Journal of Econometrics 243 (2024) 105653

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Robust inference for moment condition models without rational expectations☆

Xiaohong Chen [a], Lars Peter Hansen [b], Peter G. Hansen [c,*]

[a] Cowles Foundation for Research in Economics, Yale University, United States of America
[b] University of Chicago, United States of America
[c] Purdue University, United States of America

ARTICLE INFO

JEL classification:
C14
C15
C31
C33
G40

Keywords:
Subjective beliefs
Bounded rationality
Misspecification sets
Nonlinear expectation
Divergence
Lagrange multipliers
Stochastic dual programming
Confidence sets

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.

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 assumes agents inside the economic model form Rational Expectations (RE). This postulate can be enforced as an equilibrium construct for a fully specified dynamic stochastic equilibrium model. Under this imposition of RE, the beliefs will be consistent with the Data Generating Process (DGP) only if the model happens to be correctly specified from a statistical perspective. Alternatively, as is done in the Generalized Method of Moments (GMM) analysis, RE is imposed by assuming that the beliefs of economic agents coincide with the probabilities implied by the DGP.[1] In effect, economic agents' beliefs are restricted to be consistent with the outcomes of the Law of Large Numbers, justified (approximately) by looking at long past histories of data. For this latter approach, the corresponding optimization and equilibrium conditions lead to moment restrictions that can be estimated and tested via GMM,

☆ We are grateful to James Heckman for many inspiring research discussions. We thank Orazio Attanasio, Stephane Bonhomme, Timothy Christensen, Bo Honoré, Andrey Malenko, Per Mykland, Eric Renault, Tom Sargent, Azeem Shaikh, Ken Singleton, Grace Tsiang, Edward Vytlacil, Dacheng Xiu and two anonymous referees for helpful comments.

* Corresponding author.
E-mail addresses: xiaohong.chen@yale.edu (X. Chen), lhansen@uchicago.edu (L.P. Hansen), pghansen@purdue.edu (P.G. Hansen).

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

https://doi.org/10.1016/j.jeconom.2023.105653
Received 7 October 2021; Received in revised form 7 October 2021; Accepted 6 December 2023
Available online 25 January 2024
0304-4076/© 2024 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

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. In this paper, we use RE to mean that beliefs of economic agents inside the economic model are consistent with the DGP implied by Law of Large Numbers.

1.1. What we do

When formulating dynamic stochastic equilibrium models, it is common to impose RE as a simplifying assumption. 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. For both substantive and empirical reasons we relax RE assumption to correct the misspecification. By altering the potential subjective beliefs, we allow the population moment conditions to be satisfied, building on a suggestion in Hansen (2014). Specifically, we allow for forward-looking economic agents to have subjective beliefs that depart from the Law of Large Limits implied by the DGP. In so doing, we entertain an econometric model that is only partially identified because there is a potentially large class of subjective beliefs for which the moment conditions are satisfied. This lack of full identification prevails even though the moment conditions may be identified under RE. To characterize this family, we impose statistical bounds on how far the subjective belief probabilities are from probabilities implied by the DGP. Formally, we use statistical divergence measures to limit the set of subjective probabilities that are consistent with the population moment conditions. Since our interest is in dynamic models with forward-looking economic agents, our methodology (i) extracts information on investor beliefs from equilibrium prices and from survey data, and (ii) provides revealing diagnostics for model builders that embrace specific formulations of belief distortions.

In this paper we start by considering a generic dynamic model that is posed as a vector of conditional moment restrictions. We suppose that the conditional moment model is globally misspecified under RE, and thus empirically flawed. We then follow Hansen and Singleton (1982) and Hansen and Richard (1987) by using the conditional moment restrictions to generate unconditional counterparts. We replace the RE restriction by a statistical divergence bound on the potentially distorted beliefs relative to the DGP. In effect, we use statistical divergence as a formal way to "bound irrationality". 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. While we allow for considerable flexibility in the choice of the statistical divergences, we show that some popular statistical divergence measures are problematic: they are poor at revealing some forms of distorted beliefs that interest us.

Specifying a probability distribution is equivalent to specifying an associated expectation operator applied to a rich class of measurable functions of the 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. Thus we introduce a nonlinear expectation operator as a device to represent the restrictions on the empirically plausible family of subject probabilities.

Our paper proposes inference procedures for expectation bounds as well as the corresponding bounds on model parameters. Using duality arguments, we show how the Lagrange multipliers help to characterize extremal probabilities, which is an important intermediate step in our analysis. As a result, our econometric analysis also provides way to make inferences about these multipliers. These multipliers have independent interest because they suggest ways that we could reshape the DGP probabilities to match the moment implications. In our econometric formulation, the parameter values are only set identified. To make inferences about the sets of implied expectations, parameter values, and the corresponding multipliers, we rely on econometric theory as in Chernozhukov et al. (2007) and Chen et al. (2018).

1.2. Related literature

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.

The idea of using a nonlinear expectation operator induced by minimization has important antecedents in the applied probability literature. For instance, 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.

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 allow for the incorporation of even limited survey data by adding them into moment restrictions in our framework.

Prior contributions have used stochastic dual programs and ϕ-divergence balls. For instance, there is a mathematical overlap between our work and recent contributions such as Shapiro (2017), Duchi and Namkoong (2021) and Christensen and Connault (2019) and prior literature referenced by those papers.[2] 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.

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." We deliberately shun approaches that assume this misspecification is local. In such approaches, the misspecification vanishes at a rate that is exogenously linked to sample size. While a local approach is common in much of the related theoretical literature on estimation and inference, we find it to be unappealing for the problem that we investigate. An important part of our econometric ambition is to make inferences about the subjective beliefs of economic agents without imposing RE, and thus the potential misspecifications we explore are global in nature.

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). As in the operations research literature, the GEL literature also explores local sensitivity by assuming that the 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). The global nature of our perspective differentiates our work from this approach since our analysis features subjective beliefs of economic agents that diverge from the DGP.

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.

Our approach has several connections to work by James Heckman. Unlike many behavioral finance papers which pose specific parametric models of investor beliefs, our approach is "nonparametric" in allowing for general subjective belief distributions. This approach is analogous to that taken by Heckman and Singer (1984) for avoiding misspecification coming from distributional assumptions in duration models. Moreover, while the type misspecification we consider is motivated by subjective beliefs of economic agents which differ from RE, a similar methodology could be applied to address misspecification coming from sample selection as considered by Heckman (1979).

1.3. 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 make the nonlinear expectation computationally tractable. As illustrated in Section 5.1, 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. Specifically, 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. It proposes a flexible procedure to construct confidence sets both for the nonlinear expectations as well as bounds on individual model parameters of interest. Section 8 briefly concludes.

[2] While Shapiro (2017) and Duchi and Namkoong (2021) constructed confidence intervals assuming a 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 RE 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 RE 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 $\theta \in \Theta$,

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

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 RE. 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.

Next, we give examples of how to construct moment restriction of the form (2).

Example 2.1. Consider a modification of the standard consumption-based asset pricing model with constant relative risk aversion (CRRA) utility considered in Hansen and Singleton (1982) where the investor is allowed to have subjective beliefs that differ from RE. We can construct a moment condition as in Eq. (2) by letting

$$f(X,\theta) = \delta G^{-\gamma} R - \mathbf{1}_n$$

where $\theta = (\delta, \gamma)$ consists of the subjective discount factor δ and constant relative risk aversion γ, and $X' = (G, R')$ where G represents the consumption growth of the investor and R is an n-dimensional vector of contemporaneous returns available to the investor and $\mathbf{1}_n$ is an n-dimensional vector of ones.

Example 2.2. Consider the same consumption-based asset pricing model from Example 2.1. However, suppose that the econometrician has access to a matrix of lagged conditioning variables, Z, with n rows of variables that are available to the investor when deciding on a portfolio allocation. We construct a moment condition as in Eq. (2) using

$$f(X,\theta) = \left(\delta G^{-\gamma} R - \mathbf{1}_n\right) Z. \tag{3}$$

The vector X now also includes the entries of Z. This approach to conditioning information is analogous to that used in Hansen and Singleton (1982).

Example 2.3. Consider an investor with a unitary risk aversion but an inter-temporal elasticity of substitution parameter, parameterized as $\frac{1}{\rho}$, that is unknown to the econometrician. Preferences are formulated using recursive utility as in Kreps and Porteus (1978) and Epstein and Zin (1991). Suppose data is available on the implied return to the wealth portfolio, denoted R^w. Again the investor has subjective beliefs. We now replace the f in (3) with

$$f_1(X) = \left[\left(R^w\right)^{-1} R - \mathbf{1}_n\right] Z,$$

as a first component of the function f. There are no unknown parameters in this specification. In addition we use a so-called consumption Euler equation to form the second component of the function f:[3]

$$f_2(X,\theta) = \left[\log R^w + \log\delta + (\rho - 1)\log G\right] \widetilde{Z}$$

[3] This is a substantive change from the so-called IS equation typically used in new Keynesian models. Those models use the logarithm of the riskless return instead of R^w.

where $\widetilde{Z}$ is a vector of variables available to the investors. The f_2 component of f depends on two unknown parameters: $\theta = (\delta, \rho)$. Thus

$$f(X,\theta) = \begin{bmatrix} f_1(X) \\ f_2(X,\theta) \end{bmatrix}$$

where X includes the entries R^w, R, Z and $\widetilde{Z}$. There will typically be overlap in the conditioning variables used to construct Z and $\widetilde{Z}$.

Example 2.4. Suppose the econometrician observes survey forecasts $\widehat{Y}$ of a variable of interest Y. Given a vector $\widetilde{Z}$ of conditioning variables available at the time the forecast is made, one can construct a moment condition as in Eq. (2) using

$$f(X,\theta) = \left(Y - \widehat{Y}\right)\widetilde{Z}$$

where $X' = (Y, \widetilde{Z}')$ and no inclusion of parameters is necessary. This restriction corresponds to the survey forecasts being interpreted as subjective beliefs of survey participants, which could distinct from the forecasts implied by the DGP.

Remark 2.5. The approaches described in the previous examples are not mutually exclusive. Rather, the econometrician can combine survey evidence captured by Example 2.4 with moment restrictions given in Examples 2.2 and 2.3.

For any parameter vector θ in Eq. (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 consider a family of divergences commonly used in the statistics and machine learning literature.

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.[4] For the moment, fix θ in Eq. (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).

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

$$\phi(m) = \begin{cases} \frac{1}{\eta(1+\eta)}\left[(m)^{1+\eta} - 1\right] & \eta < 0 \\ \frac{1}{\eta(1+\eta)}\left[(m)^{1+\eta} - m\right] & \eta \geqslant 0 \end{cases} \tag{4}$$

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.[5] 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.[6]

[4] Proofs and supporting analyses for this section are given in Appendix A.

[5] 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.

[6] 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.

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 (4) and used in the GEL literature are decreasing whenever $\eta < 0$. Our finding that the empirical likelihood ($\eta = -1$) and Hellinger ($\eta = -\frac{1}{2}$) 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[7]:

Problem 3.1.

$$\underline{\kappa} \stackrel{\text{def}}{=} \inf_{M>0} \mathbb{E}[\phi(M)]$$

subject to

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

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 Eq. (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) $\mathbb{E}[M f(X)] = 0$; (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 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:

$$\sup_{\lambda,\nu} \inf_{M>0} \mathbb{E}\left[\phi(M) + M\lambda \cdot f(X) + \nu(M-1)\right] \tag{5}$$

where λ and ν are Lagrange multipliers, and $\lambda \cdot f(X)$ denote the inner product between the vectors λ and $f(X)$. 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:

[7] In this paper we use $\stackrel{\text{def}}{=}$ as a definition.

[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 $\mathbb{E}[M f(X)] = 0$ is satisfied. These perturbed distributions will converge weakly to the original distribution, and the divergence will approach zero.

Problem 3.4.

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

In this problem we have already 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

$$M^* = \frac{\exp\left(-\lambda^* \cdot f(X)\right)}{\mathbb{E}\left[\exp\left(-\lambda^* \cdot f(X)\right)\right]} \tag{6}$$

where λ^* is the maximizing choice of λ.

For this candidate M^* to be a valid solution, we must restrict the probability distribution of $f(X)$. Denote $\psi(\lambda) \stackrel{\text{def}}{=} \mathbb{E}\left[\exp\left(-\lambda \cdot f(X)\right)\right]$, 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 in λ, 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 $\psi(\cdot)$ satisfies:*

(i) *$\psi(\cdot)$ 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 (6), 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:

Problem 4.1.

$$\mathbb{K}(g) \stackrel{\text{def}}{=} \min_{M \geqslant 0} \mathbb{E}\left[M g(X)\right]$$

subject to the three constraints:

$$\mathbb{E}\left[\phi(M)\right] \leqslant \kappa$$
$$\mathbb{E}\left[M f(X)\right] = 0,$$
$$\mathbb{E}\left[M\right] = 1.$$

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.

[9] This condition rules out redundant moment conditions as well as choices of f which only take on nonnegative or nonpositive values with probability one.

[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.

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 $\mathbb{K} : \mathcal{B} \to \mathbb{R}$ 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) $\mathbb{K}(g_1) + \mathbb{K}(g_2) \leqslant \mathbb{K}(g_1 + g_2)$

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.[12]

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.

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

subject to:

$$\mathbb{E}\left[M f(X)\right] = 0,$$
$$\mathbb{E}\left[M g(X)\right] \leqslant \vartheta$$
$$\mathbb{E}\left[M\right] = 1.$$

Notice that, as stated, $\mathbb{L}(\vartheta; g)$ increases as we decrease ϑ because smaller values of ϑ make the constraint set more limiting. Thus we may decrease ϑ till $\mathbb{L}(\vartheta; g)$ attains a prespecified value κ used in constructing the nonlinear expectation $\mathbb{K}(g)$ in Problem 4.1. Additionally, it is straightforward to verify that $\mathbb{L}(\vartheta; g)$ is convex in ϑ.

We can slightly modify this approach to obtain the lower and upper bounds simultaneously. Impose the equality constraint

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

instead of the current inequality constraint in Problem 4.4. Let $\widehat{\mathbb{L}}(\vartheta; g)$ denote the resulting objective function. The minimizer over ϑ recovers the expectation of g under the minimal divergence M. For $\kappa > \underline{\kappa}$, two values of ϑ attain the relaxed divergence constraint. These values of ϑ give us the lower and upper bounds on $\mathbb{E}\left[M g(X)\right]$ of interest to us as described in Remark 4.3.

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 $\mathbb{E}\left[M g(X)\right] \leqslant \vartheta$ as an additional moment condition in Problem 4.4, we include

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

[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".

[12] All values in the interval $[\mathbb{K}(g), -\mathbb{K}(-g)]$ are attained by some belief distortion M consistent with divergence less than or equal to κ. To see this note that Problem 4.1 is always attained and that ϕ-divergences are convex.

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

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

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:[13]

$$\mathbb{K}(g) = \sup_{\xi>0}\sup_{\lambda} -\xi \log \mathbb{E}\left[\exp\left(-\frac{1}{\xi}g(X) - \lambda\cdot f(X)\right)\right] - \xi\kappa. \tag{7}$$

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:

$$\widehat{\mathbb{K}}(\xi; g) \stackrel{\text{def}}{=} \sup_{\lambda} -\xi \log \mathbb{E}\left[\exp\left(-\frac{1}{\xi}g(X) - \lambda\cdot f(X)\right)\right]. \tag{8}$$

We deduce ξ and the resulting moment bound by solving:

$$\mathbb{K}(g) = \sup_{\xi>0}\left[\widehat{\mathbb{K}}(\xi; g) - \xi\kappa\right]. \tag{9}$$

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.

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 (6) 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 to the 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: $\left[g(X) \quad f(X)'\right]'$.

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*[14]

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

[13] See Appendix A for a justification.

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

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 κ of Problem 4.1. 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.[15] 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.

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

(ii) 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 4.9.

$$\mathbb{Q}(g) \stackrel{\text{def}}{=} \inf_{M \geqslant 0} \mathbb{E}[Mg(X)]$$

subject to:

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

We allow M to be zero with positive probability for mathematical convenience. Since there exists an $M > 0$ for which $\mathbb{E}[Mf(X)] = 0$, 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.

As before, we give a dual characterization of the nonlinear expectation operator. In the following $[a]^+ = \max\{a, 0\}$.

Problem 4.10.

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

Proposition 4.11. *Assume that Restriction 4.8 holds and that the supremum in Problem 4.9 is attained with* $\xi^* > 0$. *Then* $\mathbb{Q}(g) = \widehat{\mathbb{Q}}(g)$. *Furthermore, the solution* $(\xi^*, \nu^*, \lambda^*)$ *to Problem 4.10, 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 4.9 with equality, and attains the infimum, i.e. $\mathbb{E}[M^* g(X)] = \mathbb{Q}(g)$.

Proposition 4.11 follows from theorem 6.7 of Borwein and Lewis (1992). It characterizes the solution to Problem 4.9 when the divergence constraint binds. Otherwise, we can obtain the expectation bound by solving Problem 4.10 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. Other divergences

Our analysis thus far used ϕ-divergences as the relevant notion of statistical discrepancy between probability measures. Alternative measures, which have gained some recent interest in the operations research and machine learning literature, are the Wasserstein distance and the maximum mean discrepancy.[16] Here we briefly discuss how to extend our analysis to accommodate statistical neighborhoods constructed using these alternative measures.

[15] For instance, Restriction 3.5 may be violated for some exponentially affine models of stochastic discount factors with normally distributed shocks.

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

5.1. Wasserstein distance and regularization

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 $|X - Z|^p$ for an integer $p \geqslant 1$. 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 $\mathbb{E}$ 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 $\mathbb{E}[M \mid Z] = 1$.[17] We then solve:

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

where

$$\mathbb{E}[M f(X)] = 0$$
$$\mathbb{E}[M g(X)] \leqslant \vartheta$$
$$\mathbb{E}[M \mid Z] = 1$$

When X is a continuous random vector the above problem is difficult to solve. We can follow Cuturi (2013) to add a small relative entropy penalty to the objective function:

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

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

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

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 on Z:

$$\max_{\nu} \min_{M \geqslant 0} \mathbb{E}\left[M\left(|X - Z|^p + \epsilon \log M + \lambda \cdot f(X) + \rho g(X) + \nu\right) \mid Z\right] - \nu$$
$$= -\epsilon \log \mathbb{E}\left[\exp\left(-\frac{1}{\epsilon}\left[|X - Z|^p + \lambda \cdot f(X) + \rho g(X)\right]\right) \middle| Z\right]$$

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:

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

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 entropy 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. Moment conditions of the type (2) can be defined in terms of X, or they could be reconstructed in terms of a nonsingular transformation of X. One potentially nice property of a ϕ-divergence, in contrast to Wasserstein distance is that the divergence remains the same when applied to the probability distributions induced by the change-of-variables. Thus ϕ-divergences may be more easily interpretable in certain applications.

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

$$\mathbb{E}[M\varphi(Z)] = \mathbb{E}[\mathbb{E}[M\varphi(Z)|Z]] = \mathbb{E}[\varphi(Z)\mathbb{E}[M|Z]] = \mathbb{E}[\varphi(Z)].$$.

5.2. Maximum mean discrepancy

Let $\mathcal{Q}$ be the set of all probabilities over a space $\mathcal{X}$ of potential realizations of X. Let P denote the probability implied by the DGP, which is an element of $\mathcal{Q}$. We now consider a divergence measure over probabilities without reference to absolute continuity. We solve

$$\inf_{Q\in\mathcal{Q}} \sup_{h\in\mathcal{H}} \int h(x)dQ(x) - \int h(x)dP(x)$$

subject to:

$$\int f(x)dQ(x) = 0$$
$$\int g(x)dQ(x) \leqslant \vartheta$$

and $\mathcal{H}$ is a conveniently chosen collection of test functions that is rich enough so that expectations reveal probabilities.

This problem is most tractable to implement when $\mathcal{H}$ is the unit ball of a reproducing kernel Hilbert space (RKHS). Such a space is most conveniently constructed using a kernel $k(x, y)$, that is symmetric, $k(x, y) = k(y, x)$, positive definite, $k(x, y) > 0$, and bounded. For instance, k could be the Gaussian radial kernel:

$$k(x, y) = \exp\left(-\epsilon|y - x|^2\right)$$

By design, the implied inner product, $\langle\cdot,\cdot\rangle$ for the RKHS satisfies:

$$\langle k(\cdot, x), h(\cdot)\rangle = h(x).$$

The Moore–Aronszajn Theorem informs us that such a Hilbert space can be constructed given the kernel k. Thus for any $Q \in \mathcal{Q}$,

$$\int h(x)dQ(x) = \int \langle k(\cdot, x), h(\cdot)\rangle dQ(x) = \langle \mu_Q, h\rangle$$

where

$$\mu_Q(y) = \int k(y, x)dQ(x).$$

It then follows immediately from a standard Hilbert space argument that

$$\sup_{h\in\mathcal{H}} \int h(x)dQ(x) - \int h(x)dP(x) = \sqrt{\langle \mu_Q - \mu_P, \mu_Q - \mu_P\rangle}.$$

Since the kernel, k, is symmetric, we have $\int \mu_Q(x)dP(x) = \int \mu_P(x)dQ(x)$, and

$$\begin{aligned}\langle \mu_Q - \mu_P, \mu_Q - \mu_P\rangle &= \langle \mu_Q, \mu_Q\rangle - 2\langle \mu_Q, \mu_P\rangle + \langle \mu_P, \mu_P\rangle \\ &= \int \mu_Q(x)dQ(x) - 2\int \mu_Q(x)dP(x) + \int \mu_P(x)dP(x).\end{aligned}$$

In conclusion,

$$\sup_{h\in\mathcal{H}} \int h(x)dQ(x) - \int h(x)dP(x) = \sqrt{\int \mu_Q(x)dQ(x) - 2\int \mu_Q(x)dP(x) + \int \mu_P(x)dP(x)}$$

where k is associated with a RKHS. Thus we can leave the Hilbert space implicit and work directly with the kernel.

Remark 5.4. To obtain a non-degenerate divergence measure, we want the RKHS to be sufficiently rich so that the family of expectations of the functions in this space determine a unique probability. For this reason, the kernels of interest are typically ones that are "universal". For such kernels, functions in the Hilbert space can approximate a continuous function on a compact set uniformly arbitrarily well. It is common to use radial kernels, which are represented as:

$$k(x, y) = \bar{k}\left(|x - y|^2\right).$$

Provided that $\bar{k}$ is the Laplace transform of a Borel measure on the positive real numbers that is not concentrated at zero, the radial kernel is universal.[18]

Remark 5.5. While the coordinates of f may not be in a RKHS, when the kernel is universal, we may be able to approximate f with elements in the RKHS so that the optimization problem within the RKHS could be solved by standard methods.

[18] See Theorem 17 of Micchelli et al. (2006). Simon-Gabriel and Schölkopf (2018) and Simon-Gabriel et al. (2023) provide some refinements and corrections.

Remark 5.6. The direct interpretation of the divergence measure in terms of unit ball $\mathcal{H}$ of a RKHS could be of particular interest when the g functions of interest are known to reside in this ball, perhaps achieved by a suitable scaling. Alternatively, an applied researcher could provide a thoughtful defense of the particular kernel that is chosen, while sidestepping the characterization of the RKHS.

Remark 5.7. Similar to the Wasserstein distance, the maximum mean discrepancies depend on the particular construction of X, and are sensitive to non-singular transformations of the underlying data.

6. Minimum divergence estimation and inference with unknown parameters

So far, we have suppressed the parameter dependence and focused solely on characterizing probability measures that are consistent with moment conditions. We now include an unknown parameter vector θ, residing in a parameter space, in the specification of the moment conditions and suggest some large sample approximations to inference. As before, we use a divergence measured relative to the probability implied by the DGP to identify extremal probabilities of interest. As a preliminary step towards a more complete analysis of the problem of interest, we explore inference for the following nested procedure. As an inner problem, for each given parameter value θ, we find the probability that minimizes the divergence among those probability measures that satisfy the moment conditions (2). As an outer problem, we then minimize this parameter-dependent divergence bound over the parameter space.

We could view the outcome of this analysis as a way to identify what in the misspecification literature is called a "pseudo true" parameter vector. Our aim is different, however. We use this analysis as input into our characterization of probabilities that satisfy the moment conditions of interest. The fact that θ is unknown increases the family of probability of interest as the moment restrictions only need to be satisfied for one value of the θ in the parameter space. Our model is only "misspecified" under the probability implied by the DGP, and point identification of a "pseudo true" parameter vector is of no particular interest.

We focus our discussion on the use of ϕ-divergences in this and the next sections, although some of our results have clear extensions to other divergence measures. 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 (4) 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 identify an extremal distribution, which provides guidance on how one might reshape the distributions to satisfy the moment conditions.[19] While the econometric objective posed using duality is concave in the multipliers for each θ, this property only holds conditionally (on θ). We do restrict the multipliers be unique for each θ in the parameter space. Recall that Theorem 3.6 gives sufficient conditions for this uniqueness for the relative entropy divergence (i.e., $\eta = 0$).

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 $x \in supp(X)$, $f(x, \theta)$ is continuously differentiable in θ.

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

Problem 6.2.

$$\underline{\kappa} \stackrel{\text{def}}{=} \min_{\theta \in \Theta} \mathcal{L}(\theta)$$

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

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

subject to:

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

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

Problem 6.2 differs from Problem 3.1 in that it minimizes over the set of model parameters. In particular, it gives the minimum value of the divergence $\underline{\kappa}$ such that there exists a model parameter θ and belief distortion $M^*(\theta)$ with divergence $\underline{\kappa}$ which satisfy the given moment conditions. Estimation of $\underline{\kappa}$ is important because it gives guidance to the econometrician about the minimum divergence from rational expectations they must allow to obtain non-degenerate results.

Next we consider the dual formulation of Problem 6.2. Let $\mu \stackrel{\text{def}}{=} (\lambda', \nu)'$ denote the composite multipliers for the two sets of equality constraints. The minimized divergence for a given θ solves the dual problem

$$\mathcal{L}(\theta) = \max_{\mu} \inf_{M \geqslant 0} \mathbb{E}\left[\phi(M) - \lambda \cdot f(X,\theta)M - \nu(M-1)\right]. \tag{10}$$

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

$$M(x,\mu,\theta) = \begin{cases} \left(\left[\eta[\lambda \cdot f(x,\theta) + \nu] + \frac{1}{1+\eta}\right]^+\right)^{\frac{1}{\eta}}, & \eta > 0 \\ \exp\left[\lambda \cdot f(x,\theta) + \nu - 1\right], & \eta = 0 \end{cases} \tag{11}$$

for almost all x. Substituting this solution back in to the objective (10) gives the following dual alternative to Problem 6.2:

Problem 6.3.

$$\underline{\kappa} \stackrel{\text{def}}{=} \min_{\theta \in \Theta} \mathcal{L}(\theta)$$

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

$$\mathcal{L}(\theta) = \max_{\mu} \mathbb{E}\left[F(X,\mu,\theta)\right] \tag{12}$$

where

$$F(x,\mu,\theta) \stackrel{\text{def}}{=} \begin{cases} -\frac{1}{1+\eta}\left(\left[\eta[\lambda \cdot f(x,\theta) + \nu] + \frac{1}{1+\eta}\right]^+\right)^{\frac{1+\eta}{\eta}} + \nu, & \eta > 0 \\ -\exp\left[\lambda \cdot f(x,\theta) + \nu - 1\right] + \nu, & \eta = 0 \end{cases} \tag{13}$$

We note that

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

For each θ, the solution $\mu^*(\theta)$ to (12) of Problem 6.3 is given by the first-order conditions wrt μ:

$$\mathbb{E}\left[\frac{\partial F}{\partial \mu}(X,\mu,\theta)\right] = \mathbb{E}\begin{bmatrix} -f(X,\theta)M(X,\mu,\theta) \\ 1 - M(X,\mu,\theta) \end{bmatrix} = 0. \tag{14}$$

Since the moment constraints must be satisfied, these first-order conditions imply a function μ^* of θ so that

$$\mathbb{E}\left[\frac{\partial F}{\partial \mu}(X,\mu^*(\theta),\theta)\right] = 0 \quad \text{for all} \quad \theta \in \Theta^o. \tag{15}$$

The corresponding optimized probability distortion as a function of θ using formula (11) is given by

$$M^*(\theta) \stackrel{\text{def}}{=} M(X,\mu^*(\theta),\theta), \tag{16}$$

and the optimized dual objective function is

$$\mathcal{L}(\theta) = \mathbb{E}\left[F(X,\mu^*(\theta),\theta)\right] = \mathbb{E}\left[\phi(M^*(\theta))\right].$$

Consider the set of model parameter values

$$\Theta^* \stackrel{\text{def}}{=} \arg\min_{\theta \in \Theta} \mathcal{L}(\theta) = \{\theta \in \Theta : \mathbb{E}\left[F(X,\mu^*(\theta),\theta)\right] = \underline{\kappa}\}$$

which is the set of minimal ϕ-divergence implied 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 6.4. $\Theta^* \subset \Theta^o$.

By an application of the envelope theorem, we have

$$\frac{\partial \mathcal{L}(\theta)}{\partial \theta} = \mathbb{E}\left[\frac{\partial F}{\partial \theta}(X,\mu^*(\theta),\theta)\right] \quad \text{for all} \quad \theta \in \Theta^o. \tag{17}$$

Let

$$\underline{\Theta} \stackrel{\text{def}}{=} \left\{\theta \in \Theta^o : \mathbb{E}\left[\frac{\partial F}{\partial \theta}(X,\mu^*(\theta),\theta)\right] = 0\right\} \tag{18}$$

denote the set of θ solutions to the first-order conditions. Under Assumption 6.4 we have:

$$\Theta^* = \left\{\theta \in \underline{\Theta} : \mathbb{E}\left[F(X, \mu^*(\theta), \theta)\right] = \underline{\kappa}\right\} \subseteq \underline{\Theta} . \tag{19}$$

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

Assumption 6.5. For each θ in Θ, (i) the matrix

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

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

Remark 6.6. The formula for $\mathbf{H}$ in Assumption 6.5 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 $\theta \in \Theta^o$:

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

Remark 6.7. For the relative entropy ($\eta = 0$) case, we have

$$\frac{\partial F(x, \mu, \theta)}{\partial \theta} = -\frac{\partial M(x, \mu, \theta)}{\partial \theta} = -M(x, \mu, \theta)\lambda \cdot f(x, \theta).$$

Further, the $\nu^*(\theta)$ can be solved explicitly as a function of $\lambda^*(\theta)$, and the implied minimal divergence probability is

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

Finally,

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

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

6.2. Estimation

We first study the estimation of the Lagrange multipliers conditioned on the parameter θ, and then we explore the estimation of the lower bound on the divergence.

6.2.1. Estimation of the Lagrange multipliers

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 and β-mixing.

In what follows the notation X refers to a random vector distributed according to the stationary distribution of the process $\{X_t\}$.

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 (12) of Problem 6.3 is 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 $\mu^*(\theta)$ defined in (15).

This problem fits within the framework analyzed by Haberman (1989), and 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:

$$\sqrt{T}\left[\mathcal{L}_T(\theta) - \mathcal{L}(\theta)\right] = \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left[F(X_t, \mu_T(\theta), \theta) - F(X_t, \mu^*(\theta), \theta)\right] + F_T^*(\theta) \tag{20}$$

where

$$F_T^*(\theta) \stackrel{\text{def}}{=} \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \left[F(X_t, \mu^*(\theta), \theta) - \mathbb{E} F(X_t, \mu^*(\theta), \theta) \right]. \tag{21}$$

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

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

where

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

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

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

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

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

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 $\{X_t : t \geqslant 0\}$ and restrictions on the functions of X_t and θ. See Doukhan et al. (1995), and Dedecker and Louhichi (2002).

Applying theorem 2 of Kato (2009) we obtain the following result part 2, together with (20) implies result part 1.

Result 6.11. *Under Assumptions 6.1, 6.5 and 6.10, we obtain:*

1. *The estimated minimal divergence $\mathcal{L}_T(\theta)$ has the asymptotic representation*

$$\sqrt{T} \left[\mathcal{L}_T(\theta) - \mathcal{L}(\theta) \right] = F_T^*(\theta) + o_p(1) \quad \textit{uniformly in } \Theta,$$

and converges weakly to a tight Gaussian process $\{\mathcal{G}(\theta) : \theta \in \Theta\}$ uniformly over Θ.

2. *The estimated family of Lagrange multipliers $\mu_T(\theta)$ has the asymptotic representation*

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

and converges weakly to a tight Gaussian process $\{[\mathbf{H}^*(\theta)]^{-1} \mathcal{G}_{h^*}(\theta) : \theta \in \Theta\}$.

6.2.2. Estimation of $\underline{\kappa}$

We use a simple plug-in estimator for $\underline{\kappa}$

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

for $\mathcal{L}_T(\theta)$ given in Problem 6.9. We obtain a corresponding set estimate of Θ^* as

$$\Theta_T^* = \{\theta \in \Theta : \mathcal{L}_T(\theta) \leqslant \underline{\kappa}_T + \eta_T\}, \quad \text{for some } \eta_T \geqslant 0, \ \eta_T = o_P(T^{-1}).$$

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. $\sqrt{T}(\underline{\kappa}_T - \underline{\kappa}) = \min_{\theta \in \Theta^*} F_T^*(\theta) + o_p(1) \rightsquigarrow \min_{\theta \in \Theta^*} \mathcal{G}(\theta)$.

For any finite sample we have $\mathbb{E}[\underline{\kappa}_T] \leqslant \underline{\kappa}$ but $\mathbb{E}[\underline{\kappa}_T]$ increases as T increases. Observe that if Θ^* is a singleton $\{\theta_0\}$, 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 model parameter θ, this leads us to construct confidence sets for $\underline{\Theta}$ defined in (18). When $\underline{\Theta}$ is a singleton then $\underline{\Theta}$ coincides with Θ^*, and the FOC approach provides an asymptotic exact coverage for Θ^*. When $\underline{\Theta}$ is allowed to be a set that contains at least two solutions to the first-order conditions, then $\underline{\Theta}$ could be larger than Θ^*, which might make our inferences for Θ^* conservative.[20]

We find the estimated multiplier to be of interest because it allows to characterize the extremal probability density relative to the probability distribution implied by the DGP. This leads us to combine the first-order conditions for θ with those for μ:

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

with $F(x,\mu,\theta)$ given in (13). Then the joint set of first-order conditions with respect to $\beta = (\mu', \theta')'$ is:

$$\mathbb{E}[\nabla F(X,\beta)] = 0. \tag{22}$$

We let $\mathbf{B}$ be the space of admissible parameter values for β. To formulate a tractable inference approach, we view (22) as the moment conditions for a "just-identified" GMM estimation problem.

Remark 6.13. If $\underline{\Theta}$ is assumed to be a singleton $\{\theta_0\}$, 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. In what follows we do not restrict $\underline{\Theta}$ to be a singleton.

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 using a continuously-updated GMM criterion function of Hansen et al. (1996):

$$L_T(\beta) = -\frac{1}{2}\left[\frac{1}{T}\sum_{t=1}^{T}\nabla F(X_t,\beta)\right]' [\Sigma_T(\beta)]^{-}\left[\frac{1}{T}\sum_{t=1}^{T}\nabla F(X_t,\beta)\right] \tag{23}$$

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

$$\Sigma(\beta) = \lim_{T\to\infty} Var\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\nabla F(X_t,\beta)\right) \tag{24}$$

where $Var(\cdot)$ denotes the covariance matrix of the argument in parentheses. Given $L_T(\beta)$, the data $\mathbf{X} \overset{\text{def}}{=} \{X_t\}_{t=1}^{T}$, and a prior Π over $\mathbf{B}$, the quasi-posterior distribution Π_T for β given $\mathbf{X}$ is defined as

$$d\Pi_T(\beta \mid \mathbf{X}) = \frac{\exp[T L_T(\beta)] d\Pi(\beta)}{\int_{\mathbf{B}} \exp[T L_T(\beta)] d\Pi(\beta)}. \tag{25}$$

Next we explore how to make (i) joint inferences for the multipliers and parameters and (ii) marginal inferences for the parameter vector.

6.3.1. Confidence sets for the parameter and multiplier vectors

Let

$$\underline{\mathbf{B}} \overset{\text{def}}{=} \{\beta = (\mu^*(\theta),\theta) : \theta \in \underline{\Theta}\}$$

be the potentially enlarged composite parameter space. Draw a sample β's from the quasi-posterior Π_T using a numerically tractable Monte Carlo sampler. 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:[21]

1. Draw a sample $\{\beta^1, \ldots, \beta^N\}$ from the quasi-posterior distribution Π_T in (25).
2. Calculate the $(1-\alpha)$ quantile of $\{L_T(\beta^1), \ldots, L_T(\beta^N)\}$; call it $\zeta_{T,\alpha}^{mc}$.
3. Our 100α% confidence set for $\underline{\mathbf{B}}$ is then:

$$\mathbf{B}_T^{\alpha} = \{\beta \in \mathbf{B} : L_T(\beta) \geqslant \zeta_{T,\alpha}^{mc}\}. \tag{26}$$

[20] Such an outcome could be checked in the actual estimation.

[21] This is procedure 1 in Chen et al. (2018).

Remark 6.14. As Chen et al. (2018) note, confidence sets for individual components of the β vector can be constructed by minimizing over the remaining parameters and using a chi-square one threshold.[22] An analogous approach can be used to construct confidence sets for sub-vectors of β such as the vector of Lagrange multipliers.

6.3.2. Confidence sets for θ based on the profiled moment condition

While the multipliers provide valuable diagnostics, in this subsection we suppose the focus is on the underlying model parameters $\theta \in \underline{\Theta}$. This leads us to modify the previous approach to achieve further computational simplicity by using the quasi-analytical formula for the multipliers as a function of the underlying parameters. This opens the door to a plug-in procedure that allows us to extend the previous approach in a tractable way.

Recall that $\mu^*(\theta)$ satisfies (15). Importantly, μ^* is treated as a function of θ. Provided that the multipliers are uniquely identified as a function of the parameters, we can easily deduce a point-wise (in θ) limiting distribution for the resulting estimator $\mu_T(\theta)$ at the usual parametric rate of convergence. We may deduce either by applying a standard GMM approximation for inference. By using a functional central limit theory we can get a full characterization of the limiting distribution for μ_T as it depends on θ.

We then wish to base estimation of $\theta \in \underline{\Theta}$ on the moment condition:

$$\mathbb{E}\left[\frac{\partial F}{\partial \theta}(X, \mu^*(\theta), \theta)\right] = 0 \tag{27}$$

plugging $\mu_T(\theta)$ for $\mu^*(\theta)$. Plug-in approaches require an adjustment for the estimation of $\mu_T(\theta)$ when making inferences about the $\theta \in \underline{\Theta}$ of interest. This leads us to construct:

$$\begin{aligned}\frac{\partial \widetilde{F}}{\partial \theta}(x,\theta) &\stackrel{\text{def}}{=} \frac{\partial F}{\partial \theta}(x, \mu^*(\theta), \theta) + \left[\frac{\partial \mu^*}{\partial \theta}(\theta)\right]' \frac{\partial F}{\partial \mu}(x, \mu^*(\theta), \theta) \\ &= \frac{\partial F}{\partial \theta}(x, \mu^*(\theta), \theta) - \mathbb{E}\left[\frac{\partial^2 F}{\partial \theta \partial \mu'}(X, \mu^*(\theta), \theta)\right] \left(\mathbb{E}\left[\frac{\partial^2 F}{\partial \mu \partial \mu'}(X, \mu^*(\theta), \theta)\right]\right)^{-1} \frac{\partial F}{\partial \mu}(x, \mu^*(\theta), \theta)\end{aligned}$$

The second-term in this adjustment is needed for making inferences that adjust for the estimation of $\mu^*(\theta)$. This adjustment includes two terms that contribute multiplicatively to central limit approximation:

(i) the term:

$$\mathbb{E}\left[\frac{\partial^2 F}{\partial \theta \partial \mu'}(X, \mu^*(\theta), \theta)\right]$$

captures the local impact of estimation of μ^* in moment condition (27),

(ii) and the term:

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

adjusts for the estimation of $\mu^*(\theta)$ using $\mu_T(\theta)$ based on the moment condition (15), say $\frac{1}{T}\sum_{t=1}^{T} \frac{\partial F}{\partial \mu}(X_t, \mu_T(\theta), \theta) = o_p(T^{-1/2})$ uniformly in θ.

We make inferences about the identified set $\underline{\Theta}$ for θ of interest using a continuously-updated GMM criterion function:

$$\tilde{L}_T(\theta) = -\frac{1}{2} G_T(\theta)' [\widetilde{\Sigma}_T(\theta)]^{-} G_T(\theta) \tag{28}$$

where

$$G_T(\theta) \stackrel{\text{def}}{=} \frac{1}{T}\sum_{t=1}^{T} \frac{\partial F}{\partial \theta}(X_t, \mu_T(\theta), \theta),$$

and $\widetilde{\Sigma}_T(\theta)$ is a consistent estimator of $\Sigma^*(\theta) \stackrel{\text{def}}{=} \lim_{T\to\infty} Var\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} \frac{\partial \widetilde{F}}{\partial \theta}(X_t, \theta)\right)$. Importantly, $\Sigma^*(\theta)$ will differ from $\lim_{T\to\infty} Var\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} \frac{\partial F}{\partial \theta}(X_t, \mu^*(\theta), \theta)\right)$ since the latter ignores the impact of the plug-in estimation $\mu_T(\theta)$. While $\frac{\partial \widetilde{F}}{\partial \theta}(X_t, \theta)$ depends on μ^*, as is standard in plug-in approaches, we may replace μ^* and its partial derivatives with consistent estimators and compute an asymptotic covariance estimate $\widetilde{\Sigma}_T(\theta)$ based on:

$$\left\{\frac{\partial F}{\partial \theta}(X_t, \mu_T(\theta), \theta) + \left[\frac{\partial \mu_T}{\partial \theta}(\theta)\right]' \frac{\partial F}{\partial \mu}(X_t, \mu_T(\theta), \theta) : t = 1, 2, \ldots, T\right\}.$$

Remark 6.15. There has been substantial research on plug-in approaches to estimation. For instance, Hansen (1982) provides an GMM characterization for a finite-dimensional plug-in estimator.[23] This does not cover our application because μ^* is infinite-dimensional. On the hand, $\mu^*(\theta)$ can be estimated at a standard parametric rate, making the analysis very similar. For a treatment of plug-estimators allowing for more general nonparametric auxiliary estimator, see Ackerberg et al. (2014).

[22] See their procedure 3 and theorem 4.4 for details.

[23] See Hansen (2008) Section 4.2 for an elaboration,

With plug-in adjustment, we follow the previous approach to provide confidence sets for $\underline{\Theta}$. Given $\tilde{L}_T(\theta)$, the data $\mathbf{X} = \{X_t\}_{t=1}^T$, and a prior Π over Θ, the quasi-posterior distribution Π_T for θ is

$$d\Pi_T(\theta \mid \mathbf{X}) = \frac{\exp[T\tilde{L}_T(\theta)]d\Pi(\theta)}{\int_\Theta \exp[T\tilde{L}_T(\theta)]d\Pi(\theta)}. \tag{29}$$

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 (29).
2. Calculate the $(1-\alpha)$ quantile of $\{\tilde{L}_T(\theta^1), \ldots, \tilde{L}_T(\theta^N)\}$; call it $\zeta_{T,\alpha}^{mc}$.
3. Our $100\alpha\%$ confidence set for $\underline{\Theta}$ is then:

$$\widehat{\underline{\Theta}}_\alpha = \{\theta \in \Theta : \tilde{L}_T(\theta) \geqslant \zeta_{T,\alpha}^{mc}\}. \tag{30}$$

7. Estimation of nonlinear expectation functionals and inference

This section presents estimation and inference for the nonlinear expectation functionals we constructed in Section 4. In other words, we show how the econometrician can estimate bounds on the subjective expectations consistent with pre-specified moment conditions and a given statistical divergence neighborhood.

Many prior treatments of misspecification use a divergence objective for estimation and for identifying a unique "pseudo-true" parameter. The results in Section 6 could be applied directly to support such an analysis while allowing for a non-singleton set of "pseudo-true" parameters. Our aim in this section is different, however. We characterize extremal distributions for the expectations of functions, g, and the corresponding parameters vector when divergence bound is constrained to be less than a pre-specified $\kappa > \underline{\kappa}$. We reveal the full probability bound by repeating this exercise for alternative choices of g. In some applications, a specific g may be motivated by investor preferences. We allow g to depend on the unknown parameter vector θ. This gives us the flexibility to deduce and estimate model-based ambiguity sets of parameters implied by a given statistical divergence bound, κ. We reveal the full nonlinear expectation bound by repeating this exercise for alternative choices of g.

We consider two approaches to constructing confidence sets. First, we propose a direct approach which approximates the nonlinear expectation by a finite-sample counterpart. Next we suggest an alternative simulation-based approach which relies on augmenting the moment conditions and the parameter space by including the expectation of g as an additional parameter. This augmented approach allows us to apply the results of Section 6 for estimation of the non-linear expectation.

7.1. Bounding expectation functionals using divergence balls

We first extend the previous results by letting κ be a parameter satisfying $\kappa \geqslant \underline{\kappa}$. We start with a real-valued measurable function g of x and θ. We are interested in applications to multiple choices of g, but for notational convenience we leave the dependence of solutions on g implicit, and similarly for κ.[24] For a given g, we consider the following problem designed to confront parameter dependence:

Problem 7.1.

$$\mathbb{K} \stackrel{\text{def}}{=} \min_{\theta\in\Theta} \mathcal{K}(\theta) \quad \text{where}$$

$$\mathcal{K}(\theta) \stackrel{\text{def}}{=} \inf_{M\geqslant 0} \mathbb{E}[Mg(X,\theta)] \quad \text{subject to}$$

$$\mathbb{E}[Mf(X,\theta)] = 0,$$
$$\mathbb{E}[M] = 1,$$
$$\mathbb{E}[\phi(M)] \leqslant \kappa.$$

Problem 7.1 is analogous to Problem 4.1 but with two differences. The first is that it includes a minimization over the unknown model parameter θ to reflect the econometrician's parameter uncertainty. Second, it allows additional flexibility for the model parameter θ to enter the agent's function g to be bounded. Notice that to characterize agent's subjective probabilities it would suffice to let g be a measurable function of X_t only. We allow for the additional generality in $g(X_t,\theta)$ to give researchers more flexibility in diverse applications. This additional flexibility will be useful in making inferences about model parameters.

Since $\kappa \geqslant \underline{\kappa}$, the constraint set for M is not empty at least for some values of $\theta \in \Theta$.

Similar to the previous section, the dual problem is

$$\mathcal{K}(\theta) = \sup_{\xi>0} \max_{\mu} \inf_{M\geqslant 0} \mathbb{E}[Mg(X,\theta) + \xi(\phi(M) - \kappa) - \lambda \cdot f(X,\theta)M - \nu(M-1)] \tag{31}$$

[24] A consequence of this simplicity is that we recycle some notation in this section relative to previous sections that we use in analogous ways.

where $\mu \stackrel{\text{def}}{=} (\lambda', \nu)'$ denotes vector of the composite multipliers for the equality constraints. We solve the inner problem $\inf_{M \geqslant 0}[\cdot]$ by bringing M inside the expectation. Let $M(x, \xi, \mu, \theta)$ denote the resulting solution:

$$M(x,\xi,\mu,\theta) = \begin{cases} \left(\left[\eta(\xi)^{-1}[\lambda \cdot f(x,\theta) + \nu - g(x,\theta)] + \frac{1}{1+\eta} \right]^+ \right)^{\frac{1}{\eta}}, & \eta > 0 \\ \exp\left[(\xi)^{-1}[\lambda \cdot f(x,\theta) + \nu - g(x,\theta)] - 1 \right], & \eta = 0 \end{cases} \tag{32}$$

for $\xi > 0$ and almost all x under the stationary distribution for $\{X_t\}$. By substituting this solution into the objective (31), we restate the dual problem as:

Problem 7.2. For $\theta \in \Theta$,

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

where

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

We note that

$$F(x,\mu,\theta) = \begin{cases} -\frac{\xi}{1+\eta} [M(x,\mu,\theta)]^{1+\eta} + \nu - \xi\kappa, & \eta > 0 \\ -\xi M(x,\mu,\theta) + \nu - \xi\kappa, & \eta = 0 \end{cases}$$

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

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

where:

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

As to be expected from the dual problem, these first order-conditions capture the constraints:

$$\begin{aligned} -\mathbb{E}\left[f(X,\theta) M(X,\xi,\mu,\theta) \right] &= 0 \\ 1 - \mathbb{E}\left[M(X,\xi,\mu,\theta) \right] &= 0 \\ \mathbb{E}[\phi(M(X,\xi,\mu,\theta))] - \kappa &= 0 \end{aligned} \tag{33}$$

Remark 7.3. More generally we should allow for $\xi \geqslant 0$ in (33) 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. This rules out $\xi = 0$ as an optimal solution.

Let $\mu_o(\theta), \xi_o(\theta)$ denote the solution to (33), and

$$M_o(\theta) \stackrel{\text{def}}{=} M\left(X, \xi_o(\theta), \mu_o(\theta), \theta \right)$$

given in (32) denote the implied extremal probability.

Remark 7.4. In this section we focus on moment functions f and g such that, for any $\theta \in \Theta$, there is are unique multipliers $(\mu_o(\theta), \xi_o(\theta))$ and a unique belief $M_o(\theta) = M(X, \xi_o(\theta), \mu_o(\theta), \theta)$ given in (32) solves the dual Problem 7.2.

The optimized dual objective function in Problem 7.2 becomes

$$\mathcal{K}(\theta) = \mathbb{E}\left[F(X, \xi_o(\theta), \mu_o(\theta), \theta) \right] = \mathbb{E}\left[M_o(\theta) g(X, \theta) \right].$$

Let

$$\Theta_\ell \stackrel{\text{def}}{=} \arg\min_{\theta} \mathcal{K}(\theta) = \{ \theta \in \Theta : \mathbb{E}\left[M_o(\theta) g(X,\theta) \right] = \mathbb{K} \}, \tag{34}$$

which informs us of the parameter configurations that are needed to attain the lower bound on the expectation of g. Using a parallel construction and substituting $-g$ for g, a set Θ_u informs us of the parameter configurations that are needed to attain the upper bound on the expectation of g. Suppose that $\Theta_\ell \subset \Theta^o$, then we have

$$\Theta_\ell \subseteq \underline{\Theta}_\ell \stackrel{\text{def}}{=} \left\{ \theta \in \Theta^o : \frac{\partial \mathcal{K}(\theta)}{\partial \theta} = 0 \right\} = \left\{ \theta \in \Theta^o : \mathbb{E}\left[\frac{\partial F}{\partial \theta}(X, \xi_o(\theta), \mu_o(\theta), \theta) \right] = 0 \right\}.$$

Remark 7.5. For relative entropy ($\eta = 0$) divergence, we have

$$\frac{\partial F}{\partial \theta}(x,\xi,\mu,\theta) = -\xi \frac{\partial M}{\partial \theta}(x,\xi,\mu,\theta) = -M(x,\xi,\mu,\theta)\left[\lambda \cdot \frac{\partial f(x,\theta)}{\partial \theta} - \frac{\partial g(x,\theta)}{\partial \theta}\right].$$

And

$$M_o(\theta) = \frac{\exp\left[\xi_o(\theta)^{-1}(\lambda_o(\theta)\cdot f(X,\theta) - g(X,\theta))\right]}{\mathbb{E}\left(\exp\left[\xi_o(\theta)^{-1}(\lambda_o(\theta)\cdot f(X,\theta) - g(X,\theta))\right]\right)}.$$

7.1.1. Sample analog estimation of the nonlinear expectation functional

We conclude this subsection by exploring estimation and inference using the sample counterpart to Problem 7.1 for $\kappa \geqslant \underline{\kappa}_T$:

Problem 7.6.

$$\mathbb{K}_T \stackrel{\text{def}}{=} \min_{\theta\in\Theta} \mathcal{K}_T(\theta) \quad \text{where}$$

$$\mathcal{K}_T(\theta) \stackrel{\text{def}}{=} \max_{\xi>0}\max_{\mu} \frac{1}{T}\sum_{t=1}^{T} F(X_t,\xi,\mu,\theta) = \frac{1}{T}\sum_{t=1}^{T} F(X_t,\hat{\xi}(\theta),\hat{\mu}(\theta),\theta)\,, \tag{35}$$

where $\hat{\xi}(\theta), \hat{\mu}(\theta)$ are the estimates of $\xi_o(\theta), \mu_o(\theta)$

We proceed in an analogous way as in Section 6.2 with the following modifications:

(i) we use the augmented multiplier vector (ξ, μ) instead of μ;
(ii) our minimized objective from Problem 7.6 recovers our lower bound on the expectation of g instead of $\underline{\kappa}$.

Denote

$$F_T^o(\theta) \stackrel{\text{def}}{=} \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left[F(X_t,\xi_o(\theta),\mu_o(\theta),\theta) - \mathcal{K}(\theta)\right].$$

Under these analogous assumptions, by a slight extension of theorem 3.6 of Shapiro (1991) from iid data to stationary $\beta-mixing$ data, we obtain the following asymptotic result for the estimated nonlinear expectation $\mathbb{K}_T$:

Result 7.7. $\sqrt{T}(\mathbb{K}_T - \mathbb{K}) = \min_{\theta\in\Theta_\ell} F_T^o(\theta) + o_p(1) \rightsquigarrow \min_{\theta\in\Theta_\ell} \mathcal{G}_o(\theta)$, *where* $\mathcal{G}_o$ *is a tight Gaussian process on* Θ.

Precise assumptions and expressions for the asymptotic covariance function are given in Appendix B. For any finite sample size T we have $\mathbb{E}[\mathbb{K}_T] \leqslant \mathbb{K}$, but $\mathbb{E}[\mathbb{K}_T]$ increases as T increases. If $\Theta_\ell = \{\theta_0\}$ is a singleton, then $\sqrt{T}(\mathbb{K}_T - \mathbb{K}) \rightsquigarrow \mathcal{G}_o(\theta_0)$, which is a mean zero normal random variable with variance $C_o(\theta_0,\theta_0)$ given in Appendix B.

When Θ_ℓ might not be a singleton, in the stochastic programming literature, there are currently four popular approaches for confidence sets construction using the above Result 7.7: (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_ℓ is a singleton when constructing confidence intervals (see, e.g., Shapiro, 2017; Duchi and Namkoong, 2021).

In the statistics and econometrics literature, both the subsampling approach (see, e.g., Chernozhukov et al., 2007; Romano and Shaikh, 2010) and various modified bootstrap approaches (see, e.g., Dümbgen, 1993; Fang and Santos, 2019; Hong and Li, 2020) 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_ℓ is a singleton.

In principle, any of these existing approaches can be extended to our framework with β-mixing weakly dependent time series data. For concreteness, we will describe in the next subsection a procedure to construct confidence sets for $\mathbb{K}$ via the numerical bootstrap method of Hong and Li (2020) with a modified weighted bootstrap for time series data.

7.2. Confidence sets for $\mathbb{K}$ *via numerical weighted bootstrap*

Next, we describe a numerical weighted bootstrap procedure to obtain confidence sets for $\mathbb{K}$ based on the method of Hong and Li (2020).

The first step in the bootstrap iteration is to simulate positive random weights $\{W_t : t = 1,2,\ldots,T\}$ with autocorrelation satisfying the following conditions

(i) $\{W_t : t = 1,2,\ldots,T\}$ is strictly stationary and independent of $\{X_t : t = 1,2,\ldots,T\}$
(ii) $\mathbb{E}[W_t] = 1$, $\mathbb{E}[W_t^3] < \infty$, and $Cov(W_t, W_{t+j}) = \omega(j/J)$ where $\omega(\cdot)$ is a positive symmetric kernel function with $\omega(0) = 1$ and integrates to one with lag truncation parameter J.[25]

[25] When data are weakly temporally dependent, choosing $J = O(T^{1/3})$ ensures asymptotically consistent coverage.

One example is to simulate the random weights $\{W_t\}_{t=1}^T$ as exp(1) random variables. When data is iid this is known as the Bayesian bootstrap. Another example is to let $(W_t - 1)$ have standard normal marginal distribution as in Chen and Fan (1999) for time series data. They term this a "conditional Monte-Carlo approach". We include more general random weights along with an adjustment for temporal dependence.

The second step is to draw the following bootstrap quantities. Let $F_T^b(\theta)$ denote the weighted bootstrap counterpart to $F_T^o(\theta)$:

$$F_T^b(\theta) \stackrel{\text{def}}{=} \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \left[(W_t - 1)[F(X_t, \hat{\xi}(\theta), \hat{\mu}(\theta), \theta) - \mathcal{K}_T(\theta)]\right].$$

Conditional on the data $\{X_t\}_{t=1}^T$, the process $\{F_T^b(\theta) : \theta \in \Theta\}$ converges weakly to the Gaussian process $\{\mathcal{G}_o(\theta) : \theta \in \Theta\}$ (see Appendix B). Given a positive deterministic sequence ϵ_T such that $\epsilon_T \to 0$ and $\epsilon_T\sqrt{T} \to \infty$, define

$$\mathbb{D}_T^b \stackrel{\text{def}}{=} \frac{1}{\epsilon_T}\left\{\min_{\theta\in\Theta}\left[\mathcal{K}_T(\theta) + \epsilon_T F_T^b(\theta)\right] - \min_{\theta\in\Theta}\mathcal{K}_T(\theta)\right\} \tag{36}$$

Result 7.8. *Let $\epsilon_T \to 0$ and $\epsilon_T\sqrt{T} \to \infty$. Under some mild regularity conditions, we have: conditional on the data, $\mathbb{D}_T^b \rightsquigarrow \min_{\theta\in\Theta_\ell} \mathcal{G}_o(\theta)$.*

Result 7.8 implies that the sampling distribution of the bootstrap quantity $\mathbb{D}_T^b$ converges to the asymptotic distribution of $\sqrt{T}(\mathbb{K}_T - \mathbb{K})$. See Appendix B for details. Applying this insight, one can use the following procedure to obtain a confidence set for $\mathbb{K}$ with asymptotic coverage probability α:

Confidence sets for $\mathbb{K}$:

1. For $b = 1, \ldots, N$, compute the weighted bootstrap statistic $\mathbb{D}_T^b$
2. Calculate the $\frac{1-\alpha}{2}$ and $\frac{1+\alpha}{2}$ quantiles of $\mathbb{D}_T^b$. Call them $q^b_{T,\frac{1-\alpha}{2}}$ and $q^b_{T,\frac{1+\alpha}{2}}$ respectively.
3. Form the confidence set as the interval

$$\mathbf{CS}_T^\alpha = \left[\mathbb{K}_T + T^{-\frac{1}{2}} \cdot q^b_{T,\frac{1-\alpha}{2}}, \mathbb{K}_T + T^{-\frac{1}{2}} \cdot q^b_{T,\frac{1+\alpha}{2}}\right]. \tag{37}$$

Remark 7.9. If one chooses $g(X, \theta) = \theta^i$, i.e. an individual component of the parameter vector, then the corresponding estimate $\mathbb{K}_T$ gives a lower bound on values of θ^i which rationalize the moment conditions with divergence no greater than κ.

Remark 7.10. This same weighted bootstrap approach can be used to make inferences about the parameter dependent minimum divergence $(\{\sqrt{T}\left[\mathcal{L}_T(\theta) - \mathcal{L}(\theta)\right] : \theta \in \Theta\})$ and the corresponding Lagrange multiplier process $(\{\sqrt{T}\left[\mu_T(\theta) - \mu^*(\theta)\right] : \theta \in \Theta\})$ building on Result 6.11 in Section 6.2. This same approach also would support inference for the augmented Lagrange multipliers $(\xi_o(\theta), \mu_o(\theta))$ from Problem 7.2. See Appendix B for details.

7.3. Bounding expectation functionals by augmenting the moment conditions

We next describe an alternative approach framed in terms of a vector of moment conditions augmented to include expectations of g, which is a direct target of estimation. This opens the door to using a quasi-posterior method analogous to what we described in Section 6.3.

Consider the following problem that is analogous to Problem 6.2 for any fixed ϑ:

Problem 7.11. For any fixed ϑ,

$$\mathbb{L}(\vartheta) \stackrel{\text{def}}{=} \min_{\theta\in\Theta} \mathcal{L}(\theta, \vartheta)$$

where

$$\mathcal{L}(\theta, \vartheta) \stackrel{\text{def}}{=} \inf_{M\geqslant 0} \mathbb{E}\left[\phi(M)\right]$$

subject to:

$$\mathbb{E}\left[M f(X, \theta)\right] = 0,$$
$$\mathbb{E}\left[M g(X, \theta)\right] - \vartheta = 0,$$
$$\mathbb{E}\left[M\right] = 1.$$

For a given value of ϑ, this problem has the same mathematical form as that in Section 6, except that we now have an additional moment condition:

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

Since we added a moment restriction, $\mathbb{L}(\vartheta) \geqslant \underline{\kappa}$.

Denote $\theta^a \stackrel{\text{def}}{=} (\theta', \vartheta)'$ and

$$f^a(x, \theta^a) \stackrel{\text{def}}{=} \left(f(x,\theta)', g(x,\theta) - \vartheta\right)'.$$

We slightly strengthen Assumption 6.1(ii)(iii) to hold for θ^a and $f^a(x, \theta^a)$. Let $\lambda^a \stackrel{\text{def}}{=} (\lambda', \lambda_g)'$ and $\mu^a \stackrel{\text{def}}{=} ((\lambda^a)', \nu)'$.

As input into Problem 7.11, we construct $\mathcal{L}(\theta, \vartheta)$ by solving the dual problem with optimized objective,

$$\mathcal{L}(\theta, \vartheta) = \max_{\mu^a} \mathbb{E}\left[F^a(X, \mu^a, \theta^a)\right] = \mathbb{E}\left[F^a(X, \mu^{a*}(\theta^a), \theta^a)\right] = \mathbb{E}\left[\phi(M^*(\theta^a))\right]$$

where $F^a(x, \mu^a, \theta^a)$ and $M^*(\theta^a)$ are defined in the same ways as those in (13) and (16) respectively, with μ^a, θ^a, $f^a(x, \theta^a)$ replacing μ, θ, $f(x, \theta)$ respectively.

To obtain the bounds and parameters of interest, we search over ϑ until $\mathbb{L}(\vartheta)$ equals the $\kappa > \underline{\kappa}$ of interest. As mentioned in Section 4.1, $\mathbb{L}(\vartheta)$ is convex in ϑ. Let $\vartheta_\ell < \vartheta_u$ be the two solutions to

$$\mathbb{L}(\vartheta_\ell) = \mathbb{L}(\vartheta_u) = \kappa\ .$$

Then by the convexity of $\mathbb{L}$, $\{\vartheta : \mathbb{L}(\vartheta) \leqslant \kappa\} = [\vartheta_\ell, \vartheta_u]$ and $\{\vartheta : \mathbb{L}(\vartheta) > \kappa\} = (-\infty, \vartheta_\ell) \bigcup (\vartheta_u, +\infty)$.

We consider the following sample analog of Problem 7.11

Problem 7.12. For any fixed ϑ

$$\mathbb{L}_T(\vartheta) \stackrel{\text{def}}{=} \min_{\theta \in \Theta} \mathcal{L}_T(\theta, \vartheta)$$

where

$$\mathcal{L}_T(\theta, \vartheta) \stackrel{\text{def}}{=} \max_{\mu^a} \frac{1}{T} \sum_{t=1}^{T} F^a(X_t, \mu^a, \theta^a) = \frac{1}{T} \sum_{t=1}^{T} F^a(X_t, \mu_T^a(\theta^a), \theta^a).$$

We then solve for

$$\mathbb{L}_T(\hat{\vartheta}_\ell) = \mathbb{L}_T(\hat{\vartheta}_u) = \kappa \quad \text{for} \quad \hat{\vartheta}_\ell < \hat{\vartheta}_u\ .$$

Note that $\hat{\vartheta}_\ell$ is an estimator for $\vartheta_\ell = \mathbb{K}$ (in Problem 7.1), and $-\hat{\vartheta}_u$ is an estimator for the counterpart $\mathbb{K}$ when g is replaced by $-g$.

We propose a method for inference about the solution to Problem 7.11 that is entirely analogous to what we proposed for Problem 6.2, with θ^a, μ^a, $f^a(x, \theta^a)$ replacing θ, μ^a, $f(x, \theta)$, respectively. In particular, confidence sets for θ^a can be constructed in ways analogous to those described in Section 6.

We base inferences on $\mathbb{E}\left[\Psi(X_t, \mu^a, \theta^a)\right] = 0$ where

$$\Psi(x, \mu^a, \theta^a) \stackrel{\text{def}}{=} \begin{bmatrix} F^a(x, \mu^a, \theta^a) - \kappa \\ \frac{\partial F^a}{\partial \mu^a}(x, \mu^a, \theta^a) \\ \frac{\partial F^a}{\partial \theta}(x, \mu^a, \theta^a) \end{bmatrix},$$

Importantly, we do not include the partial of F^a with respect to ϑ among the set of moment conditions used for inference, as it is not included in the construction of Ψ. It is replaced by the first entry of Ψ that we include to enforce the divergence restriction. There will be two sets of solutions to the moment conditions corresponding to the upper and lower bounds on the expectation of g.

Form the sample objective

$$L_T(\mu^a, \theta^a) = -\frac{1}{2} \left[\frac{1}{T} \sum_{t=1}^{T} \Psi(X_t, \mu^a, \theta^a)\right]' [\Sigma_T(\mu^a, \theta^a)]^- \left[\frac{1}{T} \sum_{t=1}^{T} \Psi(X_t, \mu^a, \theta^a)\right] \tag{38}$$

where $\Sigma_T(\mu^a, \theta^a)$ is a consistent estimator of $\Sigma(\mu^a, \theta^a) = \lim_{T\to\infty} Var\left(\frac{1}{\sqrt{T}} \sum_{t=1}^{T} \Psi(X_t, \mu^a, \theta^a)\right)$. With this approach, the profile quasi-likelihood ratio statistics for ϑ is defined as

$$QLR(\vartheta) \stackrel{\text{def}}{=} 2T \left[\max_{\mu^a, \theta^a} L_T(\mu^a, \theta^a) - \max_{\mu^a, \theta} L_T(\mu^a, \theta, \vartheta)\right]$$

we can compute a simple $100\alpha\%$ confidence set for ϑ as:

$$\left\{\vartheta : QLR(\vartheta) \leqslant \chi^2_{1,\alpha}\right\}, \tag{39}$$

where $\chi^2_{1,\alpha}$ denotes the α quantile of the χ^2_1 distribution.

Remark 7.13. In the relative entropy ($\eta = 0$) case, we have

$$M(x, \mu^a, \theta^a) = \exp\left[\lambda^a \cdot f^a(x, \theta^a) + \nu - 1\right]$$

used in representing beliefs as a function of the multipliers: $\lambda^a = (\lambda', \lambda_g)'$ and ν. Form:

$$F^a(x, \mu^a, \theta, \vartheta) = -\exp\left[\lambda \cdot f(x, \theta) + \lambda_g[g(x, \theta) - \vartheta] + \nu - 1\right] + \nu$$

$$= -M(x, \mu^a, \theta^a) + \nu$$

with partial derivatives given by:

$$\frac{\partial F^a}{\partial \mu^a}(x, \mu^a, \theta^a) = \begin{bmatrix} -f^a(x, \theta^a) M(x, \mu^a, \theta^a) \\ 1 - M(x, \mu^a, \theta^a) \end{bmatrix}$$

$$\frac{\partial F^a}{\partial \theta}(x, \mu^a, \theta^a) = -\lambda^a \cdot \frac{\partial f^a(x, \theta^a)}{\partial \theta} M(x, \mu^a, \theta^a).$$

We use the moment restriction: $\mathbb{E}[\Psi(X, \mu^a, \theta^a)] = 0$ for statistical inference, where

$$\Psi(x, \mu^a, \theta^a) = \begin{bmatrix} -M(x, \mu^a, \theta^a) + \nu - \kappa \\ -f^a(x, \theta^a) M(x, \mu^a, \theta^a) \\ 1 - M(x, \mu^a, \theta^a) \\ -\lambda^a \cdot \frac{\partial f^a(x, \theta^a)}{\partial \theta} M(x, \mu^a, \theta^a) \end{bmatrix}.$$

Remark 7.14. While the method in Section 7.2 is applicable more generally, we only know of convenient sufficient conditions for the method in this subsection under the point identification of the θ parameters and multipliers (and two-point identification of ϑ.) For justifications of this method in a fully point-identified GMM setting, see Gallant and Jorgenson (1979), Eichenbaum et al. (1988), and Newey and West (1987). Chen et al. (2018) justify this method (procedure 3 in their paper) allowing for partial identification (of θ) but impose an additional domination condition, which we have not verified. Nevertheless, we suspect that the approach in this subsection may be more generally applicable, but leave it for future exploration.

8. Discussion and conclusion

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. The subjective beliefs, however, are not uniquely identified. We view the subjective belief specification as a form of "bounded irrationality". This leads to use a statistical measure of divergence of the subjective probabilities relative to the probabilities implied by the data generation as a way to formally bound this irrationality. This paper devises and justifies econometric methods that support this empirical approach. 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 implied probability bounds of the empirically relevant subjective beliefs with a nonlinear expectation functional. To support this approach, 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. A future challenge is to extend the econometric methods in this paper to apply to the population characterizations given in Chen et al. (2021).

Several papers (including a suggestion in Hansen, 2014 and the analysis in Ghosh and Roussellet, 2020; Korsaye, 2022) 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 one possible measure of subjective beliefs; but we do not view it as the only plausible measure of these 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 probabilities used by 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 the data generating process, and to perform a sensitivity analysis with respect to the level of divergence.

Acknowledgments

L. P. Hansen's research was partially supported by the Macro Finance Research Program, and X. Chen's research was partially supported by the Cowles Foundation for Research in Economics.

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 $-\mathbb{E}[f(X)] \in \text{int}(C)$ 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 $v(\epsilon) = \text{vol}[B(0,\epsilon)] > 0$.[26] 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. □

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^*.[27] 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 $\mathbb{E}[M f(X)] = 0$,

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

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 $\mathsf{B} = \{M \in L^1(\Omega, \mathfrak{G}, P) : \mathbb{E}[M] = 1, \mathbb{E}[M f(X)] = 0\}$. 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. □

A.3. Derivation of Eq. (7)

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

$$\sup_{\xi>0,\lambda,\nu} \inf_{M \geqslant 0} \mathbb{E}\left[M g(X) + \xi(M \log M - \kappa) + \lambda \cdot M f(X) + \nu(M-1)\right] \tag{40}$$

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

$$M g(X) + \xi(M \log M - \kappa) + \lambda \cdot M f(X) + \nu(M-1)$$

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

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

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 Eq. (40) gives Eq. (7).

We can connect these results to our earlier analysis of dual Problem 3.4 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.[28]

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

$$M_1(g; \xi) = \frac{\exp\left[-\frac{1}{\xi}g(X)\right]}{\mathbb{E}\left(\exp\left[-\frac{1}{\xi}g(X)\right]\right)}$$

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

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. □

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

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) $ess\ inf[g(X)] = \underline{g} > -\infty$,
(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) $ess\ sup(f(X)) = \mathsf{u} < \infty$,
(iii) $\lim_{\epsilon \to 0} \mathsf{P}\{f(X) \geqslant \mathsf{u} - \epsilon\} = 0$.
(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 $\underline{g} = ess\ inf\, g(X)$ 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

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

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 $\mathsf{P}\{f(X) \leqslant \mathsf{u}\} = 1$. 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 $\mathsf{u} = \text{ess sup } f(X)$ 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. □

Appendix B. More details for Section 7

To simplify notation we let $\mu^a = (\xi, \mu')'$ and $\mu_o^a = (\xi_o, \mu_o')'$. Similar to Problem 6.9, for each fixed θ, the optimization problem (35), is a standard M-estimation problem with concave criterion in μ^a. Therefore we easily obtain that problem (35) provides consistent estimators $\hat{\mu}^a(\theta) = (\hat{\xi}(\theta), \hat{\mu}(\theta)')'$ for $\mu_o^a(\theta)$, and

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

We also obtain similar asymptotic results as follows:

$$\sqrt{T}\left[\mathcal{K}_T(\theta) - \mathcal{K}(\theta)\right] = \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left[F(X_t, \hat{\mu}^a(\theta), \theta) - F(X_t, \mu_o^a(\theta), \theta)\right] + F_T^o(\theta) \tag{41}$$

where

$$F_T^o(\theta) \stackrel{\text{def}}{=} \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left[F(X_t, \mu_o^a(\theta), \theta) - \mathcal{K}(\theta)\right].$$

We again show that only the term $F_T^o(\theta)$ contributes to the approximation (41). Again due to the concavity of F in μ^a for each θ, a gradient inequality for such functions implies that

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

where

$$h_T^o(\theta) \stackrel{\text{def}}{=} \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\frac{\partial F}{\partial \mu^a}(X_t, \mu_o^a(\theta), \theta).$$

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

Assumption B.1.

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

$$C_o(\theta_1, \theta_2) \stackrel{\text{def}}{=} \lim_{T\to\infty} Cov\left[F_T^o(\theta_1), F_T^o(\theta_2)\right] = \sum_{j=-\infty}^{\infty} Cov\left[F(X_1, \mu_o^a(\theta_1), \theta_1), F(X_{1+j}, \mu_o^a(\theta_2), \theta_2)\right]$$

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

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

$$\begin{aligned}\mathbf{V}_o(\theta_1, \theta_2) &\stackrel{\text{def}}{=} \lim_{T\to\infty} Cov\left[h_T^o(\theta_1), h_T^o(\theta_2)\right] \\ &= \sum_{j=-\infty}^{\infty} Cov\left[\frac{\partial F}{\partial \mu^a}(X_1, \mu_o^a(\theta_1), \theta_1), \frac{\partial F}{\partial \mu^a}(X_{1+j}, \mu_o^a(\theta_2), \theta_2)\right]\end{aligned}$$

Result B.2. *Under Assumption B.1, we obtain the following result:*

1. $\sqrt{T}\left[\mathcal{K}_T(\theta) - \mathcal{K}(\theta)\right] = F_T^o(\theta) + o_p(1)$ *uniformly in* $\theta \in \Theta$,
   and converges weakly to the Gaussian process $\{\mathcal{G}_o(\theta) : \theta \in \Theta\}$.
2. $\sqrt{T}\left[\hat{\mu}^a(\theta) - \mu_o^a(\theta)\right] = -\left[\mathbf{H}_o^a(\theta)\right]^{-1} h_T^o(\theta) + o_p(1)$ *uniformly in* $\theta \in \Theta$,
   and converges weakly to a tight Gaussian process $\{\left[\mathbf{H}_o^a(\theta)\right]^{-1} \mathcal{G}_{h^o}(\theta) : \theta \in \Theta\}$.

Here for any $\theta \in \Theta$, $\mathbf{H}^a(\mu^a, \theta) \stackrel{\text{def}}{=} \mathbb{E}\left[\frac{\partial^2 F}{\partial \mu^a \partial \mu^{a\prime}}(X, \mu^a, \theta)\right]$, and $\mathbf{H}_o^a(\theta) \stackrel{\text{def}}{=} \mathbf{H}^a(\mu_o^a, \theta)$ is negative definite.

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

We note that

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

and

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

$$C_{T,j}(\theta_1, \theta_2) = \frac{1}{T}\sum_{t=1}^{T-j}[F(X_t, \hat{\xi}(\theta_1), \hat{\mu}(\theta_1), \theta_1) - \mathcal{K}_T(\theta_1)][F(X_{t+j}, \hat{\xi}(\theta_2), \hat{\mu}(\theta_2), \theta_2) - \mathcal{K}_T(\theta_2)].$$

Following the proof in Chen and Fan (1999) for their proposition 5.1, we can show that conditional on the data $\{X_t\}_{t=1}^T$, the process $\{F_T^b(\theta) : \theta \in \Theta\}$ converges weakly to the Gaussian process $\{\mathcal{G}_o(\theta) : \theta \in \Theta\}$. Result 7.8 follows from Theorem 2.2 of Lachout (2005) and Theorem 3.1 of Hong and Li (2018).

Let the bootstrap random weights $\{W_t\}$ satisfy Assumption B.3. Given Result 6.11, we have the following results.

Remark B.4.

1. Let $F_T^{*b}(\theta)$ denote the weighted bootstrap counterpart to $F_T^*(\theta)$:

   $$F_T^{*b}(\theta) \stackrel{\text{def}}{=} \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left[(W_t - 1)[F(X_t, \mu_T(\theta), \theta) - \mathcal{L}_T(\theta)]\right].$$

   Conditional on the data, the bootstrap process $\{F_T^{*b}(\theta) : \theta \in \Theta\}$ converges weakly to that of the process $\{F_T^*(\theta) : \theta \in \Theta\}$ and $\{\sqrt{T}\left[\mathcal{L}_T(\theta) - \mathcal{L}(\theta)\right] : \theta \in \Theta\}$.
2. Let $h_T^{*b}(\theta)$ denote the weighted bootstrap counterpart to $h_T^*(\theta)$:

   $$h_T^{*b}(\theta) \stackrel{\text{def}}{=} \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left[(W_t - 1)[\frac{\partial F}{\partial \mu}(X_t, \mu_T(\theta), \theta)]\right].$$

Conditional on the data, the bootstrap process $\{h_T^{*b}(\theta) : \theta \in \Theta\}$ converges weakly to that of the process $\{h_T^*(\theta) : \theta \in \Theta\}$. Let $\mathbf{H}_T(\theta)$ be a consistent estimate of $\mathbf{H}^*(\theta)$ using the original data, then the bootstrap process $\{-\left[\mathbf{H}_T(\theta)\right]^{-1} h_T^{*b}(\theta) : \theta \in \Theta\}$ converges weakly to that of the multiplier process $\{\sqrt{T}\left[\mu_T(\theta) - \mu^*(\theta)\right] : \theta \in \Theta\}$.

References

Ackerberg, Daniel, Chen, Xiaohong, Hahn, Jinyong, Liao, Zhipeng, 2014. Asymptotic efficiency of semiparametric two-step GMM. Rev. Econom. Stud. 81, 919–943.

Ai, Chunrong, Chen, Xiaohong, 2007. Estimation of possibly misspecified semiparametric conditional moment restriction models with differing conditioning variables. J. Econometrics 141 (1), 5–43.

Almeida, Caio, Garcia, Rene, 2012. Assessing misspecified asset pricing models with empirical likelihood estimators. J. Econometrics 170 (2), 519–537.

Alvarez, Fernando, Jermann, Urban J., 2005. Using asset prices to measure the persistence of the marginal utility of wealth. Econometrica 73 (6), 1977–2016.

Andrews, Isaiah, Gentzkow, Matthew, Shapiro, Jesse M., 2020. On the informativeness of descriptive statistics for structural estimates. Econometrica 88 (6), 2231–2258.

Antoine, Bertille, Proulx, Kevin, Renault, Eric, 2018. Pseudo-true SDFs in conditional asset pricing models. J. Financ. Econom..

Arjovsky, Martin, Chintala, Soumith, Bottou, Léon, 2017. Wasserstein generative adversarial networks. In: International Conference on Machine Learning. PMLR, pp. 214–223.

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

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

Back, Kerry, Brown, David P., 1993. Implied probabilities in GMM estimators. Econometrica 61 (4), 971–975.

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

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

Bordalo, Pedro, Gennaioli, Nicola, Ma, Yueran, Shleifer, Andrei, 2020. Over-Reaction in Macroeconomic Expectations. Amer. Econ. Rev. 110 (9), 2748–2782.

Borwein, Jonathan M., Lewis, Adrian S., 1992. Partially finite convex programming, part II: Explicit lattice models. Math. Program. 57 (1–3), 49–83.

Broniatowski, Michel, Keziou, Amor, 2012. Divergences and duality for estimation and test under moment condition models. J. Statist. Plann. Inference 142 (9), 2554–2573.

Chen, Xiaohong, Christensen, Tim, Tamer, Elie, 2018. Monte Carlo confidence sets for identified sets. Econometrica 86 (6), 1965–2018.

Chen, Xiaohong, Fan, Yanqin, 1999. Consistent hypothesis testing in semiparametric and nonparametric models for econometric time series. J. Econometrics 91, 373–401.

Chen, Xiaohong, Hansen, Lars Peter, Hansen, Peter G., 2021. Robust identification of investor beliefs. Proc. Natl. Acad. Sci. 117 (52), 33130–33140.

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

Chernozhukov, V., Hong, H., Tamer, E., 2007. Estimation and confidence regions for parameter sets in econometric models. Econometrica 75 (5), 1243–1284.

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

Cressie, Noel, Read, Timothy R.C., 1984. Multinomial goodness-of-fit tests. J. R. Stat. Soc. Ser. B Stat. Methodol. 46 (3), 440–464.

Csiszar, I., Breuer, Thomas, 2018. Expected value minimization in information theoretic multiple priors models. IEEE Trans. Inform. Theory 64 (6), 3957–3974.

Csiszar, I., Matus, F., 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. pp. 2292–2300.

Dedecker, Jérôme, Louhichi, Sana, 2002. Maximal inequalities and empirical central limit theorems. In: Empirical Process Techniques for Dependent Data. Springer, pp. 137–159.

Doukhan, Paul, Massart, Pascal, Rio, Emmanuel, 1995. Invariance principles for absolutely regular empirical processes. In: Annales de l'IHP Probabilités et statistiques, Vol. 31. pp. 393–427.

Duchi, John C., Namkoong, Hongseok, 2021. Learning models with uniform performance via distributionally robust optimization. Ann. Statist. 49 (3), 1378–1406.

Dümbgen, Lutz, 1993. On nondifferentiable functions and the bootstrap. Probab. Theory Related Fields 95 (1), 125–140.

Eichenbaum, M.S., Hansen, L.P., Singleton, K.J., 1988. A time series analysis of representative agent models of consumption and leisure choice under uncertainty. Q. J. Econ. 103.

Eichhorn, Andreas, Römisch, Werner, 2007. Stochastic integer programming: Limit theorems and confidence intervals. Math. Oper. Res. 32 (1), 118–135.

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

Epstein, Larry G., Zin, Stanley E., 1991. Substitution, risk aversion, and the temporal behavior of consumption and asset returns: An empirical analysis. J. Polit. Econ. 99 (2), 263–286.

Fang, Zheng, Santos, Andres, 2019. Inference on directionally differentiable functions. Rev. Econom. Stud. 86 (1), 377–412.

Gagliardini, Patrick, Ronchetti, Diego, 2019. Comparing asset pricing models by the conditional hansen-jagannathan distance. J. Financ. Econom. online.

Gallant, A. Ronald, Jorgenson, Dale W., 1979. Statistical inference for a system of simultaneous, non-linear, implicit equations in the context of instrumental variable estimation. J. Econometrics 11.

Ghosh, Anisha, Roussellet, Guillaume, 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., Inoue, Atsushi, 2003. The large sample behaviour of the generalized method of moments estimator in misspecified models. J. Econometrics 114 (2), 361–394.

Hansen, Lars Peter, 1982. Large sample properties of generalized method of moments estimators. Econometrica 50 (4), 1029–1054.

Hansen, L.P., 2008. Generalized method of moments estimation. In: Durlauf, S., Blume, L. (Eds.), The New Palgrave Dictionary of Economics, second ed. Palgrave Macmillan.

Hansen, Lars Peter, 2014. Nobel lecture: Uncertainty outside and inside economic models. J. Polit. Econ. 122 (5), 945–987.

Hansen, Lars Peter, Heaton, John C., Yaron, Amir, 1996. Finite-sample properties of some alternative GMM estimators. J. Bus. Econom. Statist. 14 (3), 262–280.

Hansen, Bruce E., Lee, Seojeong, 2021. Inference for iterated GMM under misspecification. Econometrica 89 (3), 1419–1477.

Hansen, Lars Peter, Richard, Scott F., 1987. The role of conditioning information in deducing testable restrictions implied by dynamic asset pricing models. Econometrica 55 (3), 587–613.

Hansen, Lars Peter, Singleton, Kenneth J., 1982. Generalized instrumental variables estimation of nonlinear rational expectations models. Econometrica 50 (5), 1269–1286.

Heckman, James J., 1979. Sample selection bias as a specification error. Econometrica 153–161.

Heckman, James, Singer, Burton, 1984. A method for minimizing the impact of distributional assumptions in econometric models for duration data. Econometrica 271–320.

Hjort, Nils Lid, Pollard, David, 1993. Asymptotics for Minimisers of Convex Processes. In: Preprint Series. Statistical Research Report, Matematisk Institutt, Universitetet i Oslo, http://urn.nb.no/URN:NBN:no-23420.

Hong, Han, Li, Jessie, 2018. The numerical delta method. J. Econometrics 206 (2), 379–394.
Hong, Han, Li, Jessie, 2020. The numerical bootstrap. Ann. Statist. 48 (1), 397–412.
Imbens, Guido W., 1997. One-step estimators for over-identified generalized method of moments models. Rev. Econom. Stud. 64 (3), 359–383.
Imbens, Guido W., Spady, Richard H., Johnson, Phillip, 1998. Information theoretic approaches to inference in moment condition models. Econometrica 66 (2), 333–357.
Kato, Kengo, 2009. Asymptotics for argmin processes: Convexity arguments. J. Multivariate Anal. 100, 1816–1829.
Kitamura, Yuichi, Otsu, Taisuke, Evdokimov, Kirill, 2013. Robustness, infinitesimal neighborhoods, and moment restrictions. Econometrica 81 (3), 1185–1201.
Kitamura, Yuichi, Stutzer, Michael, 1997. An information-theoretic alternative to generalized method of moments estimation. Econometrica 65 (4), 861–874.
Korsaye, Sofonias Alemu, 2022. Investor Beliefs and Market Frictions. Johns Hopkins Carey Business School, Available at SSRN 4277979.
Kreps, David M., Porteus, Evan L., 1978. Temporal resolution of uncertainty and dynamic choice. Econometrica 46 (1), 185–200.
Lachout, Petr, 2005. Stochastic optimisation: Sensitivity and delta theorem. PAMM: Proc. Appl. Math. Mech. 5 (1), 725–726.
Lee, Seojeong, 2016. Asymptotic refinements of a misspecification-robust bootstrap for GEL estimators. J. Econometrics 192, 86–104.
Léger, Flavien, 2020. A gradient descent perspective on sinkhorn. Appl. Math. Optim. online.
Luttmer, Erzo, Hansen, Lars P., Heaton, John, 1995. Econometric evaluation of asset pricing models. Rev. Financ. Stud. 8, 237–274.
Manski, Charles F., 2018. Survey measurement of probabilistic macroeconomic expectations: Progress and promise. NBER Macroecon. Annu. 32 (1), 411–471.
Meeuwis, Maarten, Parker, Jonathan A, Schoar, Antoinette, Simester, Duncan I, 2018. Belief Disagreement and Portfolio Choice. Tech. Rep., National Bureau of Economic Research.
Micchelli, Charles A., Xu, Yuesheng, Zhang, Haizhang, 2006. Universal kernels. J. Mach. Learn. Res. 7, 2651–2667.
Newey, Whitney K., Smith, Richard J., 2004. Higher order properties of GMM and generalized empirical likelihood estimators. Econometrica 72 (1), 219–255.
Newey, Whitney K., West, Kenneth D., 1987. Hypothesis testing with efficient method of moments estimation. Internat. Econom. Rev. 28.
Peng, Shige, 2004. Nonlinear expectations, nonlinear evaluations and risk measures. In: Frittelli, M., Runggaldier, W. (Eds.), Stochastic Methods in Finance: Lecture Notes in Mathematics. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 165–253.
Qin, Jin, Lawless, Jerry, 1994. Empirical likelihood and general estimating equations. Ann. Statist. 22 (1), 300–325.
Romano, Joseph P., Shaikh, Azeem M., 2010. Inference for the identified set in partially identified econometric models. Econometrica 78 (1), 169–211.
Schennach, Susanne M., 2007. Point estimation with exponentially tilted empirical likelihood. Ann. Statist. 35 (2), 634–672.
Shapiro, Alexander, 1991. Asymptotic analysis of stochastic programs. Ann. Oper. Res. 30, 169–186.
Shapiro, Alexander, 2017. Distributionally robust stochastic programming. SIAM J. Optim. 27 (4), 2258–2275.
Shapiro, Alexander, Dentcheva, Darinka, Ruszczyński, Andrzej, 2014. Lectures on stochastic programming: modeling and theory, SIAM.
Simon-Gabriel, Carl-Johann, Barp, Alesssandro, Schölkopf, Bernhard, Mackey, Lester, 2023. Metrizing weak convergence with maximum mean discrepancies. J. Mach. Learn. Res. 24, 1–20.
Simon-Gabriel, Carl Johann, Schölkopf, Bernhard, 2018. Kernel distribution embeddings: Universal kernels, characteristic kernels and kernel metrics on distributions. J. Mach. Learn. Res. 19, 1–29.
Smith, Richard, 1997. Alternative semi-parametric likelihood approaches to generalized method of moments estimation. Econom. J. 107, 503–519.
Xie, Yujia, Wang, Xiangfeng, Wang, Ruijia, Zha, Hongyuan, 2019. A fast proximal point method for computing exact Wasserstein distance. In: 35th Conference on Uncertainty in Artificial Intelligence, UAI 2019.