Robust Estimation and Inference when Beliefs are Subjective*

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

This draft: January 4, 2021

Abstract

Applied researchers using structural models of economic dynamics 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 unconditional moment restrictions. We suppose that the model is globally misspecified, and thus empirically flawed in a way that is not econometrically subtle. In RE models the expectations used in defining the moment conditions are both consistent with the beliefs of the economic agents within the model and population limit of the empirical distribution. We relax this assumption by allowing subjective beliefs to differ from rational expectations while still maintaining that the moment conditions are satisfied under the subjective beliefs of economic agents whose behavior is captured by the dynamic economic model. This form of misspecification alters econometric identification and inferences in a substantial way. The underlying parameter vector ceases to be identified and the subjective beliefs may unduly weak. Therefore, we explore the consequences of restricting the statistical divergence between subjective belief distortions and their RE counterparts. In so doing, we are lead to address some new econometric challenges.

JEL Classification: C14, C15, C31, C33

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

†Yale University. Email: xiaohong.chen@yale.edu
‡University of Chicago. Email: lhansen@uchicago.edu
§MIT Sloan. Email: pghansen@mit.edu

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

1 Introduction

Models of economic dynamics have forward-looking decision makers who form beliefs about their uncertain environment. For instance, investment decisions depend on guesses about the future, firms speculate about the future demand for their products, and strategic players in a dynamic game setting make conjectures about the other players actions. One common approach is to assume agents inside the economic model form rational expectations. This postulate can be enforced as an equilibrium construct, or as is done in the generalized method of moments (GMM) approach presume that beliefs of economic agents coincide with those of data generating process as revealed by empirical evidence.[1] For this latter approach, the corresponding optimization and equilibrium conditions lead to moment restrictions that can be tested via GMM or generalized empirical likelihood (GEL) 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.

Rather than pursue the often illusive or unrevealing goal of finding a correctly specified models under rational expectations, we take a different approach. We suppose that the rational expectations model is misspecified. We depart from a common approach of entertaining a uniquely identified pseudo true parameter vector under misspecification. Moreover, we deliberately shun approaches that assume this misspecification is local. Instead we explore the consequences of allowing subjective beliefs not to be consistent with the underlying data generating process. By altering the subjective beliefs, we allow the population moment conditions to be satisfied, following on a suggestion in Hansen (2014). This approach to misspecification introduces an identification challenge as there is potentially a very large set of subjective beliefs for which the moment conditions will be satisfied. Moreover, this is true potentially for an entire set of the unknown parameters of the model.

We avoid excessively large identified sets by imposing a form of bounded rationality which we implement with a constraint on the statistical divergence between the subjective beliefs and the rational expectations counterpart. This constraint limits both the set of potential beliefs and the underlying parameters. Thus, rather than testing a specific (behavioral) model of distorted beliefs, we explore families of probabilities that belong to a convex divergence ball around the data generating process. Since boundedly rational economic agents will not embrace beliefs that are obviously inconsistent with statistical

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

evidence, we may justify the magnitude of the imposed constraint based on the statistical interpretation of the divergence that we use. Alternatively, we may explore how sensitive the identification is to changes in the magnitude of this constraint or in the divergence measure that used in formulating it. We show that some popular statistical divergence measures are problematic for studying subjective beliefs.[2] As we will see, some GEL methods emerge as a limiting version of parts of our analysis, but such methods are not the focal point of the econometric methods that we motivate and justify.

Specifying a probability distribution is equivalent to specifying an associated expectation operator applied to a rich class of functions of underlying random vector. Extending this insight, we represent bounds on a family of probability distributions as a nonlinear expectation operator constructed as follows. Minimize the expectation of the function of the random vector, for each possible function. Because of the minimization, the resulting operator is nonlinear. Since the class of functions is sufficiently large to include both a function and its negative, this operator gives both upper and lower bounds on the admissible set of expectations. This nonlinear operator is a mathematically is convenient way to depict the identified set of probabilities that capture subjective beliefs.[3]

Central to the econometric analysis are inferences on the Lagrange multipliers of the model-implied moment conditions evaluated at the subjective probabilities of the economic agents inside the model. These multipliers inform us how we must reshape the historical distribution to match the moment model implications. To make inferences about these multipliers, we apply existing econometric theory such as in Chernozhukov et al. (2007) and Chen et al. (2018) to construct confidence sets for the "misspecification set." Thus, our results feature the estimation of sets of models with similar magnitudes of belief distortions. When applied to a dynamic model with forward-looking economic agents, our methodology can be viewed as a way to (i) extract information on investor beliefs from equilibrium prices and from survey data and (ii) to provide revealing diagnostics for model builders that embrace specific formulations of belief distortions.

[2]Schennach (2007) also demonstrates a problematic aspect of the divergence associated with empirical likelihood (EL), but her approach is substantially different from ours in terms of both motivation and implementation.

[3]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.

1.1 Organization

The rest of the paper is organized as follows. Section 2 presents our framework. We follow the GEL literature by using the Cressie and Read (1984) family of divergences between probability measures. We illustrate how some divergences may be problematic for identifying (global) misspecification in financial applications. Using relative entropy divergence, we propose a nonlinear expectation functional for bounding agent's expectations subject to model-implied and a divergence constraint. We give dual representations that make evaluating the nonlinear expectation computationally tractable. Section 3 considers inference on minimal divergence measure or the confidence sets for the "misspecification set" of parameter values associated with the minimal divergence. Section 4 studies estimation and inference on nonlinear expectation functionals associated with the family of beliefs satisfying unconditional moment restrictions and a strictly convex divergence constraint. Section 5 briefly concludes.

1.2 Relation to the existing literature

There is a well known and long-standing literature on the important role of subjective beliefs in determining investment and other economic decisions. This literature to 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 models of subjective belief models. For instance, see Manski (2018), Meeuwis et al. (2018), Bordalo et al. (2020), Bhandari et al. (2019), and Attanasio et al. (2019). We can allow for the incorporation of even limited survey data by adding them into moment restrictions in our framework, but we do not explore consequences of prominent measurement errors in such data.

In terms of econometrics, our approach is related to the GEL literature on estimation and testing of moment restriction models. GEL estimates parameters and probabilities jointly in hopes of improving second-order statistical efficiency over GMM estimates for correctly specified moment restrictions. It presumes the expectations used in representing the moment conditions coincides with data generating process. See, for example, Qin and Lawless (1994), Imbens (1997), Kitamura and Stutzer (1997), Smith (1997), Imbens et al. (1998) and Newey and Smith (2004). Rather than improving the statistical performance of a correctly specified moment restrictions model, we feature a particular form of model misspecification for which there is considerable substantive interest: economic agents' subjective beliefs that diverge from the data generating process.

When the moment restrictions are misspecified, it is well known that the population criterion determines a so-called pseudo true parameter vector as the minimizer over the parameter space. In general, the pseudo true parameter vector depends on the choice of the criterion function 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 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 (2019) for a more generic global approach to misspecification, again featuring identified pseudo true parameters. In contrast, we add structure to the misspecification, and this puts the notion of a pseudo true parameter vector to the wayside.

A second approach is to assume misspecification is local around the unique "true" parameter value that satisfies the unconditional moment conditions. See, e.g., Kitamura et al. (2013), Bonhomme and Weidner (2018), Armstrong and Kolesár (2018), and Andrews et al. (2020). Unlike these papers, we take a global view of misspecification, which we find to be more conducive to our ambition. Given that we entertain subjective beliefs that differ from the data generating process, we are lead naturally to entertain set identification of both beliefs and parameters.

In operations research, stochastic optimal control, robust statistics, machine learning, sensitivity analysis and other diverse fields, distributionally robust optimization (DRO) approach is becoming increasingly popular. Many of the mathematical tools are summarized in Shapiro et al. (2014) on stochastic programming. In terms of the application of stochastic programming tools, our work has similarity to Duchi et al. (2016), Christensen and Connault (2019) and the references therein. While these papers study sensitivity with respect to the radius of a ball centering at a benchmark model, we focus on a possibly large set of subjective beliefs. Since under misspecification, our moment restriction models entertain a large set of subjective probabilities, parameters are typically only partially identified.

This paper is similarly motivated and complementary to our recent publication Chen et al. (2020). The Chen et al. (2020) contribution features identification only and does not explore estimation. 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. (2020) paper features conditional moment conditions whereas in this paper we take a finite number of unconditional moment conditions as the starting point.

2 Bounding Beliefs Using Divergence and Moment Restrictions

In dynamic economic applications, moment conditions are often justified via an assumption of rational expectations. 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 applied to time series data.

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

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

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

In this paper we allow for agents' beliefs that are revealed by the data to differ from the rational expectations beliefs implied by (infinite) histories of 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 rational expectations are altered by agent beliefs. By changing M, we allow for alternative densities. Notice that we are restricting the implied probability measures to be absolutely continuous with respect to the original probability measure implied by rational expectations. That is, we restrict the agent beliefs so that any event that has probability measure zero under the original distribution 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 under rational expectations.

2.1 Bounding Beliefs

For any parameter vector θ in equation (2), there are typically many specifications of beliefs M that will satisfy the model implied moment conditions. Rather than imposing ad hoc assumptions to resolve this identification failure, we will characterize the multiplicity by using bounds on statistical divergence. A statistical divergence quantifies how close two probability measures are. In our analysis, one of these probability measures governs the data evolution while the other governs the investment decisions or the equilibrium pricing relations. We define a range of allowable probability measures, and we consider a family of divergences commonly used in the statistics literature. We then study which of these divergences are most revealing for assessing misspecification in asset pricing models. Proofs and supporting analyses for this section are given in appendix A.

For the moment, fix θ in equation (2) and write $f(X)$. Initially we will also abstract from the role of conditioning information, but the expectations can be interpreted as being conditioned on sigma algebra $\mathfrak{I}$. Later we will investigate the role of conditioning information explicitly. 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 family of divergences $\mathbb{E}[\phi(\cdot)]$ are known as f-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 in the GEL literature, and are nested in the family of f-divergences introduced by Cressie and Read (1984) defined by

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

For $\eta = -1$ or 0, we can apply L'Hôpital's rule to obtain cases (i) and (iii) respectively. The divergence corresponding to $\eta = -\frac{1}{2}$ is equivalent to the Hellinger distance between probability densities. Empirical likelihood methods use the $\eta = -1$ divergence. 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. 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, which is known to have close links to GMM.

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.

2.1.1 Problematic divergences

For the purposes of misspecification analysis, we show that monotone decreasing divergence functions are problematic. For instance, the Cressie and Read divergences defined by (3) and used in the GEL literature are decreasing whenever $\eta < 0$. Our finding that the empirical likelihood ($\eta = -1$)and Hellinger (the case $\eta = -\frac{1}{2}$) divergences are problematic under model misspecification is noteworthy, as both have been widely used in statistics and econometrics.[4] 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:

[4]In particular, Hellinger distance has been used for the purpose of local robustness under misspecification.

Problem 2.1.

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

subject to

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

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

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

When f depends on an unknown parameter θ, we presume this inequality applies for all θ in a prespecified parameter space. This leads us to ask if this inequality is revealed by a strictly positive minimized objective in problem 2.1.

For a divergence to be of interest to us, the greatest lower bound on the objective should inform us as to how big of a statistical discrepancy is needed to satisfy equation (2). Therefore the infimum should be strictly positive whenever $\mathbb{E}[f(X)] \neq 0$. Conversely, notice that under correct specification, $E[f(X)] = 0$, and $M = 1$ is in constraint set of problem 2.1. By the design of a divergence measure, for $M = 1$ the minimized objective for problem 2.1 is zero.

Theorem 2.2. *Assume that $\phi(m)$ is decreasing in m, $\mathbb{E}[f(X)] \neq 0$, $f(X)$ is absolutely continuous w.r.t. the Lebesgue measure on $\mathbb{R}^d$, and there exists a convex cone $C \subset \mathbb{R}^d$ such that $f(X)$ has strictly positive density on C and $-\mathbb{E}[f(X)] \in int(C)$. Then for any $\kappa > 0$ there exists a belief distortion M such that i) $M > 0$ on $supp[f(X)]$; ii) $\mathbb{E}[M] = 1$; iii) $\mathbb{E}[Mf(X)] = 0$; iv) $\mathbb{E}[\phi(M)] < \kappa$.*

Theorem 2.2 shows dramatically that when the vector $f(X)$ has unbounded support, problem 2.1 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 $f(X)$.[5] We view the assumption of unbounded support as empirically relevant, since moment conditions coming from asset pricing typically have terms that

[5] 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

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 2.2 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 $\eta < 0$.

To further understand the degeneracy for $\eta < 0$ it is helpful to consider the associated dual problem to problem 2.1. The dual problem is typically easier to solve than the primal problem, and is often the starting point for generalized empirical likelihood estimation. Consider:

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

where λ and ν are Lagrange multipliers. Minimizing over M leads us to the dual problem:

Problem 2.3.

$$\begin{array}{ll} \sup_{\lambda,\nu} -\frac{1}{1+\eta}\mathbb{E}\left[\left(-\eta\left[\lambda \cdot f(X) + \nu\right]\right)^{\frac{1+\eta}{\eta}}\right] - \frac{1}{\eta(1+\eta)} - \nu & \textit{if } \eta \in (-1, 0) \\ \sup_{\lambda,\nu} \mathbb{E}\left(\log[\lambda \cdot f(X) + \nu]\right) + 1 - \nu & \textit{if } \eta = -1 \end{array}$$

provided that $\lambda \cdot f(X) + \nu \geqslant 0$.

The optimized objective from problem 2.3 is necessarily less than or equal to that of the original primal problem 2.1. When the solution to the dual problem:

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

is feasible for the primal problem, then the two optimized objectives will coincide. The support restriction on $\lambda \cdot f(X) + \nu$ can be problematic under misspecification, sometimes leading to a degenerate solution of the form $\lambda^* = 0$ and $\nu^* = 1$ with the implied M^* not being feasible.

Remark 2.4. *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*

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

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.

2.2 Relative Entropy Divergence

This section considers the relative entropy divergence (i.e., $\phi(m) = m \log m$). As known from a variety of sources and reproduced in the appendix, the dual to problem 2.1 with relative entropy divergence is:

Problem 2.5.

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

The first-order conditions for this problem are $\mathbb{E}[M^* f(X)] = 0$ where M^* is constructed using

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

where λ^* is the maximizing choice of λ.

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

Restriction 2.6. *The moment generating function ψ satisfies:*

(i) ψ *is continuous in* λ;

(ii) $\lim_{|\lambda|\to\infty} \psi(\lambda) = +\infty$.[6]

A moment generating function is infinitely differentiable in neighborhoods in which it is finite. To satisfy condition (i) of restriction 2.6, 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.

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

Restriction 2.6 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. More importantly for us, restriction 2.6 allows for $f(X)$ to have unbounded support.

Theorem 2.7. *Suppose that restriction 2.6 is satisfied. Then problem 2.2.2 has a unique solution λ^*. Using this λ^* to form M^* in (5), which satisfies the two constraints imposed in problem 2.1. Thus the optimized objective for both problems is*

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

2.2.1 Bounding expectations

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 κ's will lead to bigger sets of potential expectations.

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

Problem 2.8.

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

subject to the three constraints:

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

As before we can solve this problem using convex duality.[7] 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.

[7]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 rather than to verify regularity conditions in this literature.

Consider now 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 2.9. *The mapping $\mathbb{K} : \mathcal{B} \to \mathbb{R}$ has the following properties*[8]*:*

(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 by solving a minimization problem, and the M's that solve this problem can differ depending on g.

Remark 2.10. *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.

Next we give a dual representation of $\mathbb{K}(g)$ as justified in appendix A:

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

Notice that conditioned on ξ, the maximization over λ does not depend on κ because $-\xi\kappa$ is additively separable.

It is convenient to explore the supremum over λ for each $\xi > 0$. Write:

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

[8] 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".

We deduce ξ and the resulting moment bound by solving:

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

Remark 2.11. *For sufficiently large values of κ used to constrain relative entropy, it is possible that this constraint 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.*

2.2.2 Alternative formulation

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

Problem 2.12.

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

subject to:

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

Notice that $\mathbb{L}(\vartheta; g)$ increases as we decrease ϑ because values of ϑ make the constraint set more limiting. By imitating our previous logic for the minimum divergence problem subject to moment conditions, the corresponding dual problem is:

Problem 2.13.

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

The variable ρ is a Lagrange multiplier on the moment restriction involving g. We may hit a relative entropy target varying ϑ.

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

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

By setting $\vartheta = \mathsf{u}_g$, the solution to problem 2.13 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 2.12 as an extended version of problem 2.1 (for $\eta = 0$) with an additional moment restriction added. This leads us to state following analog to theorem 2.7.

Theorem 2.14. *Suppose*

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

ii) restriction 2.6 is satisfied for the random vector: $\begin{bmatrix} g(X) & f(X)' \end{bmatrix}'$.

Then problem 2.13 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 2.12 and 2.13 coincide.*[9]

The relative entropy objective for problem 2.12 increases as we decrease ϑ. For instance, by decreasing ϑ in this way we could hit the relative entropy threshold of problem 2.8. 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.

2.2.3 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 A. Given a function $g(X)$ of the data X, we can extend our previous arguments to

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

produce a bound on the conditional expectation. Instead of entering $\mathbb{E}\left[Mg(X)\right] \leqslant \vartheta$ as an additional moment condition in problem 2.12, we include

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

in the constraint set and vary ϑ to attain an entropy target. In practice, we solve the dual problem 2.2.2 as a function of ϑ tracing out the family of implied relative entropies.

2.3 Quadratic Divergence

While the $\eta = 0$ 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.[10] 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 2.8.

We impose two assumptions to ensure non-degenerate bounds.

Restriction 2.15. *$f(X)$ and $g(X)$ have finite second moments.*

Restriction 2.16. *There exists an $M > 0$ such that $\mathbb{E}[M] = 1$, $\mathbb{E}\left[Mf(X)\right] = 0$ and $\frac{1}{2}\mathbb{E}[M^2 - M] \leqslant \kappa$.*

The problem of interest is:

Problem 2.17.

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

subject to:

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

[10]For instance, if we specify S as an exponential-affine model of the form $S = \exp(\psi \cdot Z + Z'\Psi W)$ where W is a conditionally Gaussian shock, then restriction 2.6 may be violated.

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 2.9.

Problem 2.18.

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

Proposition 2.19. *Assume that restrictions 2.15 and 2.16 hold and that the supremum in problem 2.17 is attained with* $\xi^* > 0$. *Then* $\mathbb{Q}(g) = \widehat{\mathbb{Q}}(g)$. *Furthermore, the solution* $(\xi^*, \nu^*, \lambda^*)$ *to problem 2.18, 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 2.17 with equality, and attains the infimum, i.e. $\mathbb{E}[M^* g(X)] = \mathbb{Q}(g)$.

Proposition 2.19 follows from theorem 6.7 of Borwein and Lewis (1992). It characterizes the solution to problem 2.17 when the divergence constraint binds. Otherwise, we can obtain the expectation bound by solving problem 2.18 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.

3 Estimation

In this section, we present consistent estimation of the minimal belief distortion and of the nonlinear expectation function when the moment restrictions depend on unknown parameter vector θ. In view of the arguments made in the previous section, we focus only on the divergences ϕ's with $\eta \geqslant 0$. First, we are interested in implications across the entire set of possible θ's rather than necessarily targeting the ones with the smallest divergence. Second we are interested in the Lagrange multipliers computed from the dual problem because they inform us how we must reshape the historical distribution to match the moment model implications. Third, we exploit the concavity of the dual problems in the lagrange multipliers in justifying large sample approximations. This concavity, however, is only true conditioned on each θ and not necessarily across the various θ's. Nevertheless, we do restrict the space of admissible θ's:

Assumption 3.1. *(i) Θ is compact with non-empty interior; (ii) the data $\{X_t\}_{t=1}^T$ is strictly stationary, $\beta-mixing$.*

Given the sufficient conditions discussed in the previous sections, in this section we assume the existence and uniqueness of lagrange multipliers for any fixed $\theta \in \Theta$.

3.1 Minimal belief distortions

We explore minimal belief distortions as a starting point towards the study of beliefs constrained by divergence balls. Thus we start by studying the following problem.

Problem 3.2.

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

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

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

subject to:

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

Let $\mu \doteq (\lambda', \nu)'$ denote the composite multipliers for the two sets of constraints. From the previous sections, the minimized divergence solves the dual problem

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

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

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

for almost all X. Then Problem 3.2 could be re-expressed as:

Problem 3.3.

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

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

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

where

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

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

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

And $M^*(\theta) \doteq M(X, \mu^*(\theta), \theta)$ given in (9) denote the corresponding implied belief distortion. Both are functions of $\theta \in \Theta$. By definition of $\mu^*(\theta)$, we have

$$\frac{d\mathcal{L}(\theta)}{d\theta} = \mathbb{E}\left[\frac{\partial F}{\partial \theta}(X, \mu^*(\theta), \theta)\right] \quad \text{for all} \ \ \theta \in \Theta. \tag{12}$$

In the rest of this section we study divergences with $0 \leqslant \eta < 1$. For this range of η's the function F has a continuous second derivative with respect to the multiplier vector μ, which implies that $\mu^*(\theta)$ is continuously differentiable in θ under the following mild assumption:

Assumption 3.4. *For any* $\theta \in \Theta$*,(i)* $f(x, \theta)$ *is continuously differentiable in* θ *for almost every* x*; (ii) the matrix*

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

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

Indeed, by definition of $\mu^*(\theta)$, we have

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

By the implicit function theorem, we have that $\mu^*(\theta)$ is continuously differentiable in θ:

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

Remark 3.5. *For the relative entropy ($\eta = 0$) case, the above results simplify as $\nu^*(\theta)$ could be solved explicitly as a function of $\lambda^*(\theta)$, and the implied minimal belief is*

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

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

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

3.1.1 Minimal ϕ-divergence pseudo-true model parameters

Let $\mathcal{M}^* \doteq \{M^*(\theta) = M(\cdot, \mu^*(\theta), \theta) : \theta \in \Theta\}$ denote the collection of minimal ϕ-divergence implied probabilities that satisfy the model moment restrictions. Members of $\mathcal{M}^*$ could be ranked according to their degree of divergence (from the rational expectation belief) $\mathbb{E}\left[\phi(M^*(\theta))\right]$. Recall that

$$\underline{\kappa} = \min_{\theta\in\Theta} \mathcal{L}(\theta) = \min_{\theta\in\Theta} \mathbb{E}\left[\phi(M^*(\theta))\right].$$

Let $\underline{\Theta} \doteq \arg\min_{\theta\in\Theta} \mathcal{L}(\theta)$, which is the set of minimal ϕ-divergence implied pseudo-true model parameter values. $\underline{\Theta}$ is typically assumed to be a singleton in the literature on GEL. Here it is allowed to be a non-singleton. We note that correct model specification corresponds to $\underline{\kappa} = 0$, or equivalently, $M^*(\theta) = 1$ (or equivalently, $\mu^*(\theta) = 0$) for all $\theta \in \underline{\Theta}$. In our paper we consider belief distortions and hence assume $\underline{\kappa} > 0$.

Let

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

denote the set of stationary solutions.

Assumption 3.6. *Θ_S belongs to the interior of Θ.*

Under Assumptions 3.4 and 3.6, we have:

$$\underline{\Theta} = \left\{\theta \in \Theta_S : \frac{d^2\mathcal{L}(\theta)}{d\theta d\theta'} \text{ is positive definite}\right\}.$$

That is, $\underline{\Theta}$ is the set of solutions in Θ_S satisfying the second order condition.

Remark 3.7. *When $\underline{\Theta} = \Theta_S$, Problem 3.3 could be equivalently solved as the following just-identified system of equations:*

$$\begin{bmatrix} \mathbb{E}[\frac{\partial F}{\partial \mu}(X, \mu, \theta)] \\ \mathbb{E}[\frac{\partial F}{\partial \theta}(X, \mu, \theta)] \end{bmatrix} = 0. \tag{13}$$

Remark 3.8. *For the relative entropy ($\eta = 0$) case, Problem 3.3 becomes*

$$\underline{\kappa} = -\log(\underline{v}) \ , \quad \underline{v} = \max_{\theta \in \Theta} \min_{\lambda} \mathbb{E} \left(\exp \left[\lambda \cdot f(X, \theta) \right] \right). \tag{14}$$

All the solutions satisfy the following first-order conditions:

$$\mathbb{E} \begin{bmatrix} \exp \left[\lambda \cdot f(X, \theta) \right] f(X, \theta) \\ \exp \left[\lambda \cdot f(X, \theta) \right] \frac{\partial f(X,\theta)}{\partial \theta} \cdot \lambda \end{bmatrix} = 0. \tag{15}$$

3.1.2 Profile M-estimation of Lagrange multipliers

With the above formulation of Problem 3.3, the estimation of $\mu^*(\theta)$ for each θ is a special case of an M-estimation problem with a concave objective function. The sample counterpart to Problem 3.3 is

Problem 3.9.

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

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

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

We proceed by obtaining a functional central limit approximation for:

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

where

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

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

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

where

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

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

Assumption 3.10.

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

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

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

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

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

Under Assumption 3.10, we obtain the following result:

Result 3.11.

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

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

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

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

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

Define

$$\widehat{\underline{\kappa}} = \min_{\theta \in \Theta} \mathcal{L}_T(\theta) \quad \text{and} \quad \widehat{\underline{\Theta}} = \{\theta \in \Theta : \mathcal{L}_T(\theta) = \widehat{\underline{\kappa}} + o_P(T^{-1})\}.$$

Then under mild conditions (including those ensuring the uniqueness of $\mu^*(\theta)$, a slight extension of theorem 3.6 of Shapiro (1991) from iid data to stationary $\beta - \textit{mixing}$ data, we have

Result 3.12. $\sqrt{T}(\widehat{\underline{\kappa}} - \underline{\kappa}) = \min_{\theta \in \underline{\Theta}} \sqrt{T}\left[\mathcal{L}_T(\theta) - \mathcal{L}(\theta)\right] + o_p(1) \rightsquigarrow \min_{\theta \in \underline{\Theta}} \mathcal{G}(\theta)$.

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

Remark 3.13. *If* $\underline{\Theta}$ *is further assumed to be a singleton set* $\{\theta_0\}$, *then either the sample analog of Problem 3.3 or the just-identified moment problem* (13) *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 this just-identified moment conditions to establish root-T asymptotic normality of their estimators for* $(\mu^*(\theta_0), \theta_0)$ *jointly for possibly* $E[f(X, \theta)] \neq 0$ *with iid data. Almeida and Garcia (2012) establish root-T asymptotic normality for* $(\mu^*(\theta_0), \theta_0)$ *under stationary strongly mixing data. Even for iid data and assuming unique pseudo-true parameter value* θ_0, *there is no "efficiency" claim as that in Ghosh et al. (2017).*

3.2 Bounding expectation functionals using divergence balls

We extend the previous results by letting κ be a parameter satisfying $\kappa > \underline{\kappa}$. Define

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

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

Problem 3.14. *Let $\kappa > \underline{\kappa}$. Assume that for each $\theta \in \Theta$, $g(x, \theta) \notin span\{f(x, \theta)\ , 1\ \}$.*

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

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

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

Similar to the previous section, the dual problem is

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

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

$$M(X, \xi, \mu, \theta) = \begin{cases} \left(\left[\eta(\xi)^{-1}[\lambda \cdot f(X, \theta) + \nu - g(X, \theta)] + \frac{1}{1+\eta}\right]^+\right)^{\frac{1}{\eta}}, & \eta > 0 \\ \exp\left[(\xi)^{-1}[\lambda \cdot f(X, \theta) + \nu - g(X, \theta)] - 1\right], & \eta = 0 \end{cases} \tag{16}$$

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

Problem 3.15. *For any fixed $\theta \in \Theta$,*

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

where

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

We have:

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

$$\frac{\partial F}{\partial\xi}(x,\xi,\mu,\theta,\kappa) = \phi(M(x,\xi,\mu,\theta))-\kappa$$

The solutions to Problem 3.15 are given by the first-order conditions wrt (μ,ξ):

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

which is

$$\begin{aligned} -\mathbb{E}\left[f(X,\theta)M(X,\xi,\mu,\theta)\right] &=0 \\ 1-\mathbb{E}\left[M(X,\xi,\mu,\theta)\right] &=0 \\ \mathbb{E}[\phi(M(X,\xi,\mu,\theta))]-\kappa &=0 \end{aligned} \tag{17}$$

Let $\mu_0(\theta,\kappa),\xi_0(\theta,\kappa)$ denote the solution to (17), and $M_0(\theta,\kappa)=M(X,\xi_0(\theta,\kappa),\mu_0(\theta,\kappa),\theta)$ given in (16) denote the (g,κ,θ) implied belief. Then Problem 3.14 becomes

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

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

We note that

$$\frac{\partial\mathcal{K}(\theta,\kappa)}{\partial\theta}=\mathbb{E}\left[\frac{\partial F}{\partial\theta}(X,\xi_0(\theta,\kappa),\mu_0(\theta,\kappa),\theta,\kappa)\right],$$

$$\frac{\partial\mathcal{K}(\theta,\kappa)}{\partial\kappa}=-\xi_0(\theta,\kappa)<0.$$

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

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

Remark 3.16. *For relative entropy ($\eta = 0$) case, Problem 3.14 simplifies to*

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

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

with

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

3.2.1 Profile M-estimation of Lagrange multipliers

The sample counterpart to Problem 3.14 is Problem 3.17:

Problem 3.17. *For any $\kappa > \underline{\kappa}$*

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

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

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

To simplify notation, let $\theta^a = (\theta', \kappa)'$ and $\mu^a = (\xi, \mu')'$. Again similar to Problem 3.9, the estimation for $\mu_0^a(\theta^a)$ is the M-estimation with convave criterion, we will obtain similar results as follows:

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

where

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

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

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

where

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

This leads us to make the following assumption similar to assumption 3.10

Assumption 3.18.

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

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

for any θ_1^a, θ_2^a.

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

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

Under Assumption 3.18, we obtain the following result:

Result 3.19.

1. Uniformly over θ^a,

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

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

2. Uniformly in θ^a,

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

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

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

3.2.2 Estimation of the expectation functional

Let $\kappa > \underline{\kappa}$ and define

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

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

Result 3.20. $\sqrt{T}(\widehat{\mathbb{K}}(\kappa)-\mathbb{K}(\kappa)) = \min_{\theta\in\Theta_\kappa}\sqrt{T}\,[\mathcal{K}_T(\theta^a) - \mathcal{K}(\theta^a)]+o_p(1) \rightsquigarrow \min_{\theta\in\Theta_\kappa}\mathcal{G}^a(\theta,\kappa)$.

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

Given Result 3.20, one could apply weighted bootstrap to construct confidence interval for $\mathbb{K}(\kappa)$. However, we prefer the quasi Bayes procedure as presented in the next section.

3.3 Frontier belief distortion estimation under expectation bounds

Define

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

Problem 3.21. *Assume* $\vartheta < \overline{\vartheta}$.

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

where for fixed θ*,*

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

subject to:

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

Note that for any $\vartheta > \overline{\vartheta}$, we have $\mathbb{L}(\vartheta) = \underline{\kappa}$ and the inequality constraint is satisfied without binding, which is not interesting. So we assume $\vartheta < \overline{\vartheta}$, which will lead to solutions with binding constraint. The dual problem becomes

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

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

$$M(X,\zeta,\mu,\theta) = \begin{cases} \left(\left[\eta[\lambda \cdot f(X,\theta) + \nu - \zeta g(X,\theta)] + \frac{1}{1+\eta}\right]^{+}\right)^{\frac{1}{\eta}} , & \eta > 0 \\ \exp\left[\lambda \cdot f(X,\theta) + \nu - \zeta g(X,\theta) - 1\right] , & \eta = 0 \end{cases} \tag{18}$$

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

Problem 3.22. *For any fixed $\theta \in \Theta$,*

$$\mathcal{L}(\theta;\vartheta) = \max_{\zeta>0} \max_{\mu} \ \mathbb{E}\left[F(X,\zeta,\mu,\theta,\vartheta)\right],$$

where

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

We have:

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

$$\frac{\partial F}{\partial \zeta}(x,\zeta,\mu,\theta,\vartheta) = g(x,\theta)M(x,\zeta,\mu,\theta) - \vartheta$$

The solutions to Problem 3.22 are given by the first-order conditions wrt (μ,ζ):

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

which is

$$-\mathbb{E}\left[f(X,\theta)M(X,\zeta,\mu,\theta)\right] = 0$$

$$
\begin{aligned}
1-\mathbb{E}\left[M(X,\zeta,\mu,\theta)\right] =& 0\\
\mathbb{E}[g(X,\theta)M(X,\zeta,\mu,\theta)]-\vartheta =& 0
\end{aligned} \tag{19}
$$

Let $\mu^*(\theta,\vartheta),\zeta^*(\theta,\vartheta)$ denote the solution to (19), and $M^*(\theta,\vartheta) = M(X,\zeta^*(\theta,\vartheta),\mu^*(\theta,\vartheta),\theta)$ given in (18) denote the (g,ϑ,θ) implied "optimal" belief. Then Problem 3.21 becomes

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

$$
\mathcal{L}(\theta,\vartheta) = \max_{\zeta>0}\max_{\mu}\ \mathbb{E}\left[F(X,\zeta,\mu,\theta,\vartheta)\right] = \mathbb{E}\left[F(X,\zeta^*(\theta,\vartheta),\mu^*(\theta,\vartheta),\theta,\vartheta)\right] = \mathbb{E}\left[\phi(M^*(\theta,\vartheta))\right].
$$

We note that

$$
\begin{aligned}
\frac{\partial\mathcal{L}(\theta,\vartheta)}{\partial\theta} &= \mathbb{E}\left[\frac{\partial F}{\partial\theta}(X,\zeta^*(\theta),\mu^*(\theta),\theta)\right],\\
\frac{\partial\mathcal{L}(\theta,\vartheta)}{\partial\vartheta} &= -\zeta^*(\theta,\vartheta) < 0.
\end{aligned}
$$

Thus $\mathbb{L}(\vartheta)$ increases as ϑ decreases for $\vartheta < \overline{\vartheta}$; while $\mathbb{L}(\vartheta) = \underline{\kappa}$ for all $\vartheta \geqslant \overline{\vartheta}$.

3.3.1 Profile M-estimation of Lagrange Multipliers

The sample counterpart is

Problem 3.23. *For each fixed* $(\theta,\vartheta>0)$*,*

$$
\mathcal{L}_T(\theta,\vartheta) = \max_{\zeta>0}\max_{\mu}\ \frac{1}{T}\sum_{t=1}^{T}F(X_t,\zeta,\mu,\theta,\vartheta) = \frac{1}{T}\sum_{t=1}^{T}F(X_t,\zeta_T,\mu_T,\theta,\vartheta)
$$

where (ζ_T,μ_T) *is the corresponding estimate for* $(\zeta^*(\theta,\vartheta),\mu^*(\theta,\vartheta))$*.*

Problem 3.24. *Assume* $\vartheta < \overline{\vartheta}$*.*

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

4 Inferences

This section presents two approaches to construct confidence sets. The first one is a weighed bootstrap for time series data. The second one is Monte Carlo simulation based (quasi posteriors).

4.1 Inferences via weighted bootstrap

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

Assumption 4.1. *(i) $\{W_t\}_{t=1}^T$ is strictly stationary and independent of data $\{X_t\}_{t=1}^T$; (ii) $E[W_t] = 1$, $E[W_t)^3] < \infty$, $Cov(W_t, W_{t+j}) = \omega(j/J)$, where $\omega()$ is a positive symmetric kernel function with J be the lag truncation parameter.*

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

4.1.1 Bootstrap inference on minimal divergence

We define the weighed bootstrap analog of $\mathcal{L}_T(\theta)$ as follows:

Problem 4.2.

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

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

Result 4.3. *Under Assumptions 3.10 and 4.1, we obtain the following result:*

1. *Conditionalondata, $\{\sqrt{T}\left[\mathcal{L}_T^B(\theta) - \mathcal{L}_T(\theta)\right] : \theta \in \Theta\}$ converges weakly to a tight mean zero Gaussian process $\{\mathcal{G}(\theta) : \theta \in \Theta\}$, with covariance function $C^*(\theta_1, \theta_2)$.*

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

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

The following is intuitively why the bootstrap works.

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

where

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

Then, again by concavity of F we have

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

where

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

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

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

and

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

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

Similarly we have:

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

and

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

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

Similarly, we define the weighed bootstrap analog of $\widehat{\underline{\kappa}}$ as follows:

$$\widehat{\underline{\kappa}}^B = \min_{\theta \in \Theta} \mathcal{L}_T^B(\theta).$$

Result 4.4. *Conditional on data,* $\sqrt{T}(\widehat{\underline{\kappa}}^B - \widehat{\underline{\kappa}}) = \min_{\theta \in \Theta} \sqrt{T}\left[\mathcal{L}_T^B(\theta) - \mathcal{L}_T(\theta)\right] + o_p^*(1) \rightsquigarrow$

$\min_{\theta \in \underline{\Theta}} \mathcal{G}(\theta)$.

4.1.2 Bootstrap inference on nonlinear expectation functionals

The same bootstrap procedure also works for nonlinear expectation functionals.

4.2 Confidence sets via quasi posteriors

Even when $\underline{\Theta}$ is a singleton $\{\theta_0\}$, it is known that GEL estimators of θ_0 is difficult to compute. When the model is correctly specified, Chernozhukov and Hong (2003) has suggested to apply Monte Carlo simulations quasi posterior to provide confidence intervals for GEL estimators of point-identified θ_0. Precisely, they propose a quasi Bayesian procedure for estimation of θ_0 based on the profiled GEL criterion (in terms of our paper notation)

$$\theta_0 = \arg\min_{\theta} \mathcal{L}(\theta) \ , \quad \mathcal{L}(\theta) = \mathbb{E}\left[\phi(M^*(\theta))\right].$$

We find that, even for iid data, the MCMC posterior draw is no longer consistent for frequentist inference under non-RE beliefs (i.e., misspecification). Recently, Chib et al (2018) also find that the Bayesian posterior corresponding to ELET estimator does not coincide with the frequentist result in Snnach.s—although the point estimates coincide, but the asymptotic variance disagree. Instead, we find that the MCMC draw using the just identified moment (15) remains valid.

4.2.1 Confidence set for "least misspecification set" via quasi posteriors

Recently, Chen et al. (2018) has proposed a slightly different way to compute critical values based on the quasi posterior approach that are valid for set-valued $\underline{\Theta}$. Below we apply procedure 1 of Chen et al. (2018) to provide a consistent confidence set for the identified set for the whole set of parameters $\underline{\mathbf{B}} \doteq \{\beta = (\mu^*(\theta), \theta) : \theta \in \underline{\Theta}\}$. In addition, we can apply Procedure 2 of Chen et al. (2018) to obtain confidence set for $\widehat{\underline{\Theta}}$. Denote

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

Then the joint first-order conditions wrt (μ, θ) is:

$$\mathbb{E}[\rho(X, \mu, \theta)] = 0. \tag{20}$$

Following Hansen, Heaton and Yaron (1996) we define the continuously updated GMM criterion function:

$$L_T(\mu,\theta) = -0.5\left[\frac{1}{T}\sum_{t=1}^{T}\rho(X_t,\mu,\theta)\right]'[\hat{\Sigma}(\mu,\theta)]^{-}\left[\frac{1}{T}\sum_{t=1}^{T}\rho(X_t,\mu,\theta)\right]$$

where, for each (μ,θ),$[\hat{\Sigma}(\mu,\theta)]^{-}$ is the generalized inverse of $\hat{\Sigma}(\mu,\theta)$, which is a consistent estimator of $\Sigma(\mu,\theta)$:

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

Note that for each fixed θ, we have unique solution of μ. But θ could be partially identified. We could directly apply CCT to the "just-identified" moment condition (20).

Let $\beta = (\mu,\theta)$ and $\mathbf{B} = \{\beta = (\mu,\theta) \in \Lambda \times \Theta\}$. Let $\underline{\mathbf{B}} = \{\beta = (\mu^*(\theta),\theta) : \theta \in \underline{\Theta}\}$. Given $L_T(\beta)$, the data $\mathbf{X}$ and a prior Π over $\mathbf{B}$, the quasi-posterior distribution Π_T for β given $\mathbf{X}$ is defined a

$$d\Pi_T(\beta \mid \mathbf{X}) = \frac{\exp[TL_T(\beta)]d\Pi(\beta)}{\int_{\mathbf{B}}\exp[TL_T(\beta)]d\Pi(\beta)}\,. \tag{21}$$

We draw a sample $\{\beta^1,\ldots,\beta^B\}$ from the quasi-posterior Π_T. Any Monte Carlo sampler could be used. Chen et al. (2018) suggested to use an adaptive sequential Monte Carlo (SMC) algorithm that is known to perform well for drawing from irregular, multi-modal distributions. We seek a CS $\widehat{\mathbf{B}}_\alpha$ for $\underline{\mathbf{B}}$ such that $\lim_{T\to\infty} Pr(\underline{\mathbf{B}} \subseteq \widehat{\mathbf{B}}_\alpha) = \alpha$.

Confidence sets for $\underline{\mathbf{B}}$:

1. Draw a sample $\{\beta^1,\ldots,\beta^B\}$ from the quasi-posterior distribution Π_T in (21).

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

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

$$\widehat{\mathbf{B}}_\alpha = \{\beta \in \mathbf{B} : L_T(\beta) \geqslant \zeta^{mc}_{T,\alpha}\}\,. \tag{22}$$

Let $\widehat{\beta} \in \mathbf{B}$ denote an approximate maximizer of L_T, i.e.:

$$L_T(\widehat{\beta}) = \max_{\beta\in\mathbf{B}} L_T(\beta) + o_p(T^{-1})\,.$$

and define the quasi-likelihood ratio (QLR) (at a point $\beta \in \mathbf{B}$) as:

$$QLR_T(\beta) = 2T[L_T(\hat{\beta}) - L_T(\beta)]\,. \tag{23}$$

Let $\xi^{mc}_{T,\alpha}$ denote the α quantile of $\{QLR_T(\beta^1), \ldots, QLR_T(\beta^B)\}$. The confidence set

$$\widehat{\mathbf{B}}'_\alpha = \{\beta \in \mathbf{B} : QLR_T(\beta) \leqslant \xi^{mc}_{T,\alpha}\}$$

is equivalent to $\widehat{\mathbf{B}}_\alpha$ defined in (22) for $\underline{\mathbf{B}}$.

The simple projection-based confidence set (CS) for $\underline{\Theta}$ is given by:

$$\widehat{\underline{\Theta}}^{proj}_\alpha = \{\theta : (\mu, \theta) \in \widehat{\mathbf{B}}_\alpha \text{ for some } \mu\} \tag{24}$$

which is a valid $100\alpha\%$ CS for $\underline{\Theta}$ whenever $\widehat{\mathbf{B}}_\alpha$ is a valid $100\alpha\%$ CS for $\underline{\mathbf{B}}$, and is known to be conservative:

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

We follow procedure 2 of CCT for a Monte Carlo based CS for $\underline{\Theta}$. The profile criterion for a point $\theta \in \Theta$ is simply $L_T(\mu_T(\theta), \theta)$, and the profile criterion for $\underline{\Theta}$ is

$$PL_T(\underline{\Theta}) \equiv \inf_{\theta \in \underline{\Theta}} \max_{\mu \in \Lambda_\theta} L_T(\mu, \theta)\ , \quad \text{with} \;\; \Lambda_\theta \doteq \{\mu : (\mu, \theta) \in \mathbf{B}\} \tag{25}$$

Let $\Delta(\beta^b) = \{\beta \in \mathbf{B} : M^*(\theta) = M^*(\theta^b)\}$ for likelihood models, and $\Delta(\beta^b) = \{\beta \in \mathbf{B} : E[\rho(X, \beta)] = E[\rho(X, \beta^b)]\}$ for moment-based models. Let $\Theta(\beta^b) = \{\theta : (\mu, \theta) \in \Delta(\beta^b) \text{ for some } \mu\}$ and

$$PL_T(\Theta(\theta^b)) \equiv \inf_{\theta \in \Theta(\theta^b)} \max_{\mu \in \Lambda_\theta} L_T(\mu, \theta)$$

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

1. Draw a sample $\{\beta^1, \ldots, \beta^B\}$ from the quasi-posterior distribution Π_n in (21).

2. Calculate the $(1 - \alpha)$ quantile of $\left\{PL_T(\Theta(\theta^b)) : b = 1, \ldots, B\right\}$; call it $\zeta^{mc,p}_{n,\alpha}$.

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

$$\widehat{\underline{\Theta}}_\alpha = \left\{\theta \in \Theta : \sup_{\mu \in \Lambda_\theta} L_T(\mu, \theta) \geqslant \zeta^{mc,p}_{n,\alpha}\right\}. \tag{26}$$

Applying CCT we have:

$$\lim_{T\to\infty} Pr(\underline{\Theta} \subseteq \underline{\widehat{\Theta}}_\alpha) = \alpha \ .$$

Recall the definition of the QLR QLR_T in (23). We define the profile QLR for the set $\Theta(\beta^b)$ analogously as

$$PQLR_T(\Theta(\beta^b)) \equiv 2T[L_T(\widehat{\beta}) - PL_T(\Theta(\beta^b))] \ = \ \sup_{\theta\in\Theta(\beta^b)} \inf_{\mu\in\Lambda_\theta} QLR_T(\mu,\theta)\,. \tag{27}$$

Let $\xi^{mc,p}_{n,\alpha}$ denote the α quantile of the profile QLR draws $\{PQLR_T(\Theta(\beta^b)) : b = 1,\ldots,B\}$. The confidence set:

$$\underline{\widehat{\Theta}}'_\alpha = \left\{\theta \in \Theta : \inf_{\mu\in\Lambda_\theta} QLR_T(\mu,\theta) \leqslant \xi^{mc,p}_{n,\alpha}\right\}$$

is equivalent to $\underline{\widehat{\Theta}}_\alpha$.

Remark 4.5. *When θ is a scalar, we can also apply Procedure 3 of CCT for CS for θ via simple chi sqaured one upper bound as follows.*

1. *Calculate a maximizer $\widehat{\beta}$ for which $L_T(\widehat{\beta}) \geqslant \sup_{\beta\in\boldsymbol{B}} L_T(\beta) + o_p(T^{-1})$.*
2. *Our 100α% confidence set for $\underline{\Theta} \subset \mathbb{R}$ is then:*

$$\underline{\widehat{\Theta}}^\chi_\alpha = \left\{\theta \in \Theta : \inf_{\mu\in\Lambda_\theta} QLR_T(\mu,\theta) \leqslant \chi^2_{1,\alpha}\right\} \tag{28}$$

where QLR_T is the QLR statistics in (23) and $\chi^2_{1,\alpha}$ denotes the α quantile of the χ^2_1 distribution.

4.3 Inference for frontier belief distortions under expectation bounds

We can also apply Procedure 3 of CCT for confidence set for the nonlinear expectation functional.

1. Calculate a maximizer $\hat{\theta}$ for which $L_T(\hat{\theta}) \geqslant \sup_{\theta\in\Theta} L_n(\theta) + o_p(n^{-1})$.
2. Our 100α% confidence set for $M_I \subset \mathbb{R}$ is then:

$$M^\chi_\alpha = \left\{\mu \in M : \inf_{\eta\in H_\mu} Q_n(\mu,\eta) \leqslant \chi^2_{1,\alpha}\right\} \tag{29}$$

where Q_n is the QLR in (23) and $\chi^2_{1,\alpha}$ denotes the α quantile of the χ^2_1 distribution.

Based on the above various equivalence formulations of the problems, we could construct the incremental Hansen J test as a simple inference tool. Let $\mathcal{L}_T(\theta;\vartheta)$ denote the sample analog of $\mathcal{L}(\theta;\vartheta)$ as in Problem 3.9 (for $\mathcal{L}(\theta)$). Then all the previous properties $\mathcal{L}_T(\theta)$ remain valid for $\mathcal{L}_T(\theta;\vartheta)$ by augmenting θ to (θ,ϑ). In addition, we could consider the incremental Hansen J test statistic as:

$$\mathcal{J}_T \doteq \min_{\theta\in\Theta} \mathcal{L}_T(\theta;\vartheta) - \min_{\theta\in\Theta} \mathcal{L}_T(\theta),$$

which will be asymptotically chi-square distributed with one degree of freedom.

Result 4.6. $\mathcal{J}_T \rightsquigarrow \chi^2(1)$.

5 Discussion and Conclusion

GEL methods target a single parameter vector for which moment conditions hold under data-generating process. Under misspecification, such methods will sometimes (but not necessarily, see Theorem 2.2 or Schennach (2007)) consistently estimate a pseudo true parameter. The implied empirical probabilities will approximate a minimally divergent probability measure under which the moment conditions hold. In our subjective belief framework, this approach will produce model parameters and implied beliefs which are consistent with model-implied moments and have a minimal divergence relative to rational expectations. Several papers (including a suggestion in Hansen (2014) and the analysis in Ghosh and Roussellet (2020)) treat the minimal divergent beliefs and corresponding model parameter as the target of estimation. In contrast, we view the distorted probability recovered in this way merely as a plausible candidate for subjective beliefs, and we do not view it as the only plausible measure of subjective beliefs. Additionally, we do not necessarily view an identified parameter vector associated with the minimal divergent beliefs as the only parameter value of interest. After all, we see no economic principle which states that the subjective beliefs of market participants must appear to the econometrician to have minimal divergence relative to rational expectations. Instead we consider it more fruitful to characterize and bound sets of plausible beliefs and model parameters consistent with certain levels of divergence from rational expectations (i.e. “misspecification sets”) and perform sensitivity analysis with respect to the level of divergence. We therefore view the methods developed in Chen et al. (2020) and in this paper as more appropriate to analyze implied subjective beliefs of economic agents than existing GEL methods.

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

Our recently published paper Chen et al. (2020) uses a similar perspective to explore identification using conditional moment restrictions and dynamic counterparts to the divergence measures we consider in this paper. As an important future challenge, we plan to extend the econometric methods in this paper to apply to population characterizations given in Chen et al. (2020).

References

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

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

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

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

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

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

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

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

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

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

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

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

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

Chen, Xiaohong, Lars Peter Hansen, and Peter G. Hansen. 2020. Robust identification of investor beliefs. *Proceedings of the National Academy of Sciences* 117 (52):33130–33140.

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

Chernozhukov, Victor and Han Hong. 2003. An MCMC approach to classical estimation. *Journal of Econometrics* 115 (2):293–346.

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

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

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

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

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

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

Duchi, John, Peter Glynn, and Hongseok Namkoong. 2016. Statistics of robust optimization: A generalized empirical likelihood approach. *arXiv preprint arXiv:1610.03425* .

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

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

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

Ghosh, Anisha, Christian Julliard, and Alex P. Taylor. 2017. What is the Consumption-CAPM Missing? An Information-Theoretic Framework for the Analysis of Asset Pricing Models. *Review of Financial Studies* 30 (2):442–504.

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

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

Hansen, Bruce E and Seojeong Lee. 2019. Inference for iterated GMM under misspecification. *Unpublished Manuscript* .

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Appendix

A Proofs and Derivations for Section 2

A.1 Proof of Theorem 2.2

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 $v(\epsilon) = \text{vol}[B(0, \epsilon)] > 0$.[11] 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. □

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

A.2 Derivation of Problem 2.3

By standard duality arguments, the dual formulation of problem 2.1 is the saddlepoint equation

$$\sup_{\lambda,\nu} \inf_{M>0} \mathbb{E}\left[\phi(M) + M\lambda \cdot f(X) + \nu\,(M-1)\right] \tag{30}$$

where λ and ν are Lagrange multipliers.

The objective function is separable over the realized values of M, and this leads us to minimize:

$$\phi(M) + M\lambda \cdot f(X) + \nu\,(M-1)$$

The first-order condition for optimizing over M is:

$$\frac{1}{\eta}M^{\eta} + \lambda \cdot f(X) + \nu = 0.$$

Thus the minimizing M is

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

Substituting the minimizing M back into (4) leads us to

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

as in dual problem 2.3.

A.3 Proof of Theorem 2.7

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 2.1 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 2.1.

To verify that M^* solves Problem 2.1, 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^*.[12] Compute

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

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

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

We conclude that

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

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

A.4 Derivation of equation (6)

By standard duality arguments, the dual formulation of problem 2.8 is the saddlepoint equation

$$\sup_{\xi > 0, \lambda, \nu} \inf_{M \geqslant 0} \mathbb{E}\left[Mg(X) + \xi(M \log M - \kappa) + \lambda \cdot Mf(X) + \nu(M - 1)\right] \tag{31}$$

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

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

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

with respect to M. The first-order condition is

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

Thus,

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

Substituting back into equation (31) gives equation (6).

We can connect these results to our earlier analysis of dual Problem 2.2.2 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 2.7 for the existence of a solution where Restriction 2.6 is imposed under the change of measure.[13]

A.5 When will the relative entropy constraint bind?

We first give a high-level sufficient condition under which the relative entropy constraint in problem 2.8 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

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

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

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

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 2.8 binds for any value of $\kappa > \overline{\kappa}$.

Proof. An application of the Envelope Theorem gives that

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

where

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

Applying L'Hôpital's rule, we see that

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

The result follows. □

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

Example A.3 establishes that the relative entropy constraint will bind in problem 2.8 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 2.8 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) $\lim_{\epsilon \to 0} \mathsf{P}\left\{f(X) \geqslant \mathsf{u} - \epsilon\right\} = 0$.

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

Then for any $\kappa > 0$, the relative entropy constraint in Problem 2.8 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.6 Auxiliary results

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

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

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

Proof:

Write

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

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

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

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

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

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

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

which simplifies to

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

It follows that

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

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

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

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

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

$$\geqslant \frac{1}{2}\log\left[\frac{1}{2h(\epsilon)}\right] - \frac{1}{e}$$

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 $\mathsf{u} =$ 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

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

Now choosing ϵ sufficiently small so that the lower bound exceeds κ gives the desired result. □