

NBER WORKING PAPER SERIES

THE ROLE OF BELIEFS IN INFERENCE FOR RATIONAL EXPECTATIONS MODELS

Bruce N. Lehmann

Working Paper 11758

<http://www.nber.org/papers/w11758>

NATIONAL BUREAU OF ECONOMIC RESEARCH

1050 Massachusetts Avenue

Cambridge, MA 02138

November 2005

This paper had an unusually long gestation period. I want to thank Lars Hansen for a long conversation I had with him sometime in the last ten years (!) and to apologize to those with whom I have had helpful conversations that I have forgotten. I also want to thank seminar participants at the Federal Reserve Bank of New York, Hong Kong University of Science and Technology, the London School of Economics, Southern Methodist University, Syracuse University, the University of Alberta, the University of Arizona, the University of California at San Diego, the University of Houston, and Yale University, students at the 2002 SSRC Summer Workshop in Applied Economics: Risk and Uncertainty, and conference participants at The First Symposium on Econometric Theory and Application. The views expressed herein are those of the author(s) and do not necessarily reflect the views of the National Bureau of Economic Research.

2005 by Bruce N. Lehmann. All rights reserved. Short sections of text, not to exceed two paragraphs, may be quoted without explicit permission provided that full credit, including notice, is given to the source.

The Role of Beliefs in Inference for Rational Expectations Models

Bruce N. Lehmann

NBER Working Paper No. 11758

November 2005, October 2007

JEL No. C1,C2,C3,C4,C5

ABSTRACT

This paper discusses inference for rational expectations models estimated via minimum distance methods by characterizing the probability beliefs regarding the data generating process (DGP) that are compatible with given moment conditions. The null hypothesis is taken to be rational expectations and the alternative hypothesis to be distorted beliefs. This distorted beliefs alternative is analyzed from the perspective of a hypothetical semiparametric Bayesian who believes the model and uses it to learn about the DGP. This interpretation provides a different perspective on estimates, test statistics, and confidence regions in large samples, particularly regarding the economic significance of rejections in rational expectations models.

Bruce N. Lehmann

University of California, San Diego

IR/PS

1415 Robinson Building Complex

La Jolla, CA 92093-0519

and NBER

blehmann@ucsd.edu

1. Introduction

A somewhat longer version of the following question appeared on the finance field exam at Columbia in 1991: Consider the excess return on a market index $R_{mt} - R_{ft}$. How would you test the null hypothesis that $E[R_{mt} - R_{ft} | I_{t-1}] = 0$ by examining p sample moment conditions $\bar{g}_T = \frac{1}{T} \sum_t g_t$; $g_t = (R_{mt} - R_{ft})z_{t-1}$; $z_{t-1} \in I_{t-1}$ for some conditioning information set I_{t-1} ? Suppose it is certain that $E[R_{mt} - R_{ft} | I_{t-1}] = 0$. How would you interpret large values of the test statistic?¹

The answer is deceptively simple. The null hypothesis can be tested with the generalized method of moments (GMM) overidentifying restrictions test statistic $T\bar{g}'_T S_T^{-1} \bar{g}_T \rightarrow \chi^2(p)$ where $S_T = \frac{1}{T} \sum_t g_t g'_t$; see Hansen (1982). Now if the null model $E[g_t | I_{t-1}] = 0$ is a maintained hypothesis, rational expectations becomes the null hypothesis and the alternative is then that expectations were not rational. Even if \bar{g}_T reliably differed from zero in a statistical sense, there might be beliefs implicit in a rejection region that seem plausible given the historical record. This possibility of assessing the economic significance of statistical rejections makes the distorted beliefs alternative a natural one in rational expectations models.

The distorted beliefs alternative arises when an econometrician specifies an economic model of the relations among a set of observables x_t of the form $E_{P_0}[g(x_t, \theta_{P_0})] = 0$, where θ_{P_0} is an unknown parameter vector and P_0 is the data generating process if expectations are rational and the model is correct. The econometrician estimates P_0 using minimum distance methods via $\inf_P |P - \hat{P}|$ where $|\cdot|$ is a measure of the distance between the empirical distribution \hat{P} and P , which satisfies the *a priori* moment restrictions. What is missing is the link between the econometrician's estimate and other *ex post* beliefs that might seem plausible.

¹ Unsurprisingly, nobody ever answers the questions I put on field exams.

This gap is closed by supposing the econometrician asks the following question: what beliefs might a hypothetical expected utility maximizer have after looking at the same data? The answer is not entirely straightforward because a semiparametric Bayesian would need to specify priors over the space of probability measures that satisfy $E_p[g(x_t, \theta_p)] = 0$. The circumstances in which the archetype's beliefs would converge to those of the econometrician can be quite delicate because a prior that places too little mass in the neighborhood of P_0 or too much outside of such neighborhoods can lead to inconsistent posteriors. Fortunately, there are weak sufficient conditions under which the archetype's beliefs will converge to those of the econometrician.

This paper is related to the extensive literature on empirical likelihood and related minimum divergence estimators; see Owen (2001) and Kitamura (2006) for recent surveys. It is most closely related to Back and Brown (1992,1993), who discuss GMM estimation of probability distributions, and to Zellner (1994,1997), Kim (2002), Lazar (2003), and Schennach (2005), who discuss Bayesian inference in GMM settings. However, the paper is almost orthogonal to the latter, in which probabilities are nuisance parameters and interest centers on inference for θ . Here probability measures are not nuisance parameters to be profiled or integrated out but rather are the focus of the analysis.

The paper is laid out as follows. The next section describes the *a posteriori* beliefs of a hypothetical semiparametric Bayesian and discusses circumstances in which they will converge to the probability measure estimated by a GMM econometrician. The penultimate section suggests some of the ways in which this insight can inform the interpretation of estimates, test statistics, and confidence regions in large samples. A brief conclusion rounds out the paper.

2. A Portrait of a Semiparametric Bayesian

This section constructs a semiparametric Bayesian archetype – one who believes in a

model comprised of moment conditions with otherwise general preferences and constraints – with a view to finding weak sufficient conditions under which posterior beliefs converge to the corresponding probability model estimated by a GMM econometrician. After setting the stage in Section 2A, Section 2B shows that the archetype will want beliefs that converge weakly and why convergence can obtain with relatively unrestricted priors in an iid setting. Section 2C proves that convergence still obtains on the subspace of discrete measures and 2D discusses the corresponding subspace for the GMM econometrician, which differs only in that each discrete approximation satisfies the moment conditions. This seemingly trivial modification results in a very simple structure that provides insight into the distorted beliefs alternative.

A. Preliminaries

To fix the setting, let x be a random variable taking values on a sample space $\mathcal{X} \subseteq \mathbb{R}^d$, let $\mathcal{B}_{\mathcal{X}}$ be the Borel σ -algebra of \mathcal{X} , and let $g(x, \theta) = \{g_j \in \mathcal{F} : \mathcal{X} \times \Theta \rightarrow \mathbb{R} \forall \theta \in \Theta \subset \mathbb{R}^q, j \leq p\}$, where \mathcal{F} is the space of all bounded real-valued uniformly continuous functions. Let \mathcal{P}^θ be a nonempty set of probability measures P on $(\mathcal{X}, \mathcal{B}_{\mathcal{X}})$ that satisfy $E_P[g(x, \theta_p)] = 0$, where $\text{rank}(E_P[g_\theta(x, \theta_p)]) = p$ and $E_P[g(x_t, \theta)] = 0 \Rightarrow \theta = \theta_p$ since θ can differ across \mathcal{P}^θ . Since \mathcal{X} is a complete separable metric space, \mathcal{P}^θ is metrizable and can be equipped with its Borel σ -algebra $\mathcal{B}_{\mathcal{P}^\theta}$; see Theorem 6.2 of Parthasarathy (1967). Finally, let P^0 denote the measure governing the realizations of x and θ_{p_0} its associated parameter value.² To avoid the notational clutter associated with atoms and P -continuity sets, each $P \in \mathcal{P}^\theta$ is taken to be dominated by a

² Some Bayesians prefer to think of P^0 as being drawn randomly from \mathcal{P}^θ . Alternatively, one can view the analysis as conditional on P^0 being true under the null with the understanding that there can be a separate modeling exercise under the alternative hypothesis. On this interpretation, the semiparametric Bayesian would possess priors over this model class and assign the remaining prior probability to all remaining model classes. This Bayesian would view P^0 as the measure that minimizes the Kullback-Leibler divergence between it and the truth under the alternative.

σ -finite measure μ with density given by the Radon-Nikodym derivative $p = dP/d\mu$.

B. On the Beliefs of a Semiparametric Bayesian

The hypothetical semiparametric Bayesian is taken to maximize expected utility (or to minimize expected expenditure or cost for a given level of utility or production) taking account of the uncertainty in both the random variables that impinge on this maximum problem and the probability law generating them. Many such problems can be cast in the following form:

A1: The archetype chooses actions $a \in \mathcal{A} \subseteq \mathbb{R}^k$ to maximize the conditional expectation of a bounded utility function $V: \mathcal{X} \times \mathcal{A} \rightarrow \mathbb{R} \cup \{-\infty\}$, where V is upper semicontinuous for almost all $x \in \mathcal{X}$, based on the information in the sub- σ -algebra $\mathcal{F} \subset \mathcal{B}_{\mathcal{X}}$.

A2: The archetype formulates a prior $\Pi(P)$ which satisfies $\int_{\mathcal{P}^\theta} \Pi(dP) = 1$ where the propriety of the prior ensures that integrals over \mathcal{P}^θ converge.

In these circumstances, the Bayesian archetype will solve the maximum problem:

$$\sup_{a \in \mathcal{A}} E_{\bar{P}}[V(x, a) | \mathcal{F}] = \sup_{a \in \mathcal{A}} \int_{\mathcal{X}} V(x, a) \bar{P}(dx | \mathcal{F}) = \int_{\mathcal{X}} V(x, a) \int_{\mathcal{P}^\theta} P(dx | \mathcal{F}) \Pi(dP | \mathcal{F}) \quad (1)$$

where $\Pi(P | \mathcal{F})$ is the posterior probability that $P = P^\theta$ and $\bar{P}(dx | \mathcal{F})$ is the predictive distribution for the next realization of x .³

Something definitive can be said about predictive and posterior distribution asymptotics under two additional assumptions, one about the measures in \mathcal{P}^θ and one about priors over \mathcal{P}^θ .

A3: Each $x_t \stackrel{\text{iid}}{\sim} P^\theta$ and there is a random sample $X^T = \{x_1, x_2, \dots, x_T\}$ with $\mathcal{F}_T = X^T$.⁴

Given A3, the predictive distribution given X^T is given by:

³ As is readily apparent, additional random variables $y \in \mathcal{Y} \subseteq \mathbb{R}^m$ can impinge on the stochastic program as long as they can be integrated out. This would be the case if x is taken to be weakly exogenous with respect to y in the language of Engle et al. (1983) and if the prior distribution is constructed to insure that the conditional distributions $P(y|x, \mathcal{F})$ are conditionally independent of the distributions $P(x|\mathcal{F})$ in \mathcal{P}^θ both *a priori* and *a posteriori*.

⁴ $\mathcal{F}_T = X^T$ can be replaced with $X^T \in \mathcal{F}_T$, which would make $\bar{P}(x_{T+1} | X^T) = E[\bar{P}(x_{T+1} | \mathcal{F}_T) | X^T]$.

$$\begin{aligned}\bar{P}(x_{T+1}|X^T) &= \int_{\mathcal{P}^\theta} P(x_{T+1})\Pi(P|X^T) = \frac{\int_{\mathcal{P}^\theta} P(x_{T+1})P(X^T)\Pi(dP)}{\int_{\mathcal{P}^\theta} P(X^T)\Pi(dP)} \\ &= \frac{\int_{\mathcal{P}^\theta} P(x_{T+1})\prod_{t \leq T} p(x_t)\Pi(dP)}{\int_{\mathcal{P}^\theta} \prod_{t \leq T} p(x_t)\Pi(dP)}\end{aligned}\quad (2)$$

where independence is only used in the passage from the first line to the second.⁵

A4. (Schwartz (1965)): The prior of the archetype satisfies $\Pi[K_\varepsilon(P^0)] \forall \varepsilon > 0$ where

$$K_\varepsilon(P) = \left\{ Q : \int p \ln(p/q) d\mu < \varepsilon \right\}.$$

Note that the archetype's prior is defined over \mathcal{P}^θ , not over θ and a nuisance parameter that defines the measures compatible with the moment conditions for each θ . In contrast to most semiparametric settings where interest centers on θ and not on nuisance parameters like probabilities, the archetype is interested in the model only for forecasting and, hence, would naturally form priors over probability measures, not parameter values.⁷

A1 – A4 suffice for weak convergence of the posterior and predictive distributions for P.

Theorem 1 (Theorem 6.1 of Schwartz (1965) and Theorem 4.4.2 of Ghosh and Ramamoorthi (2003): Let U be any weak neighborhood of P^0 and let $U^c = \mathcal{P}^\theta \setminus U$. Under assumptions A1 – A4, $\Pi(U^c | X^T) \rightarrow 0$ a.s. P^0 .

The idea of the proof is as follows. The posterior probability that P is in U^c is given by:

⁵ This serves to make it clear that there is nothing in the Bayesian calculus that makes it difficult to accommodate heterogeneity and dependence. The difficulty lies in the curse of dimensionality, the need to replace $p(x_t)$ with $p(x_t|\mathcal{F}_{t-1})$ throughout. A formal way to handle heterogeneity and dependence is to rewrite $P(X^T)$ as:

$$P(X^T) = \frac{P(X^T)}{\prod_{t \leq T} p(x_t)} \prod_{t \leq T} p(x_t) \equiv \Lambda(X^T) \prod_{t \leq T} p(x_t)$$

where $\Lambda(X^T)$ is the likelihood ratio statistic for the hypothesis that $x_t \stackrel{\text{iid}}{\sim} P$. Heterogeneity and dependence can be integrated out if $\Lambda(\bullet)$ is distributed independently of $p(\bullet)$ both *a priori* and *a posteriori*. The elucidation of the circumstances in which \mathcal{P}^θ has the required structure is beyond the scope of this paper.

⁶ $K_\varepsilon(P)$ is termed a Kullback-Leibler neighborhood of P and A4 is taken to mean that P^0 is in the Kullback-Leibler support of the prior.

⁷ In addition, priors over probability measures are invariant with respect to reparameterizations of the form $\phi = f(\theta)$.

$$\Pi\left(\mathbf{P} \in U^c \mid \mathbf{X}^T\right) = \frac{\int_{U^c} \prod_{t \leq T} p(x_t) \Pi(d\mathbf{P})}{\int_{\mathcal{P}^\theta} \prod_{t \leq T} p(x_t) \Pi(d\mathbf{P})} \quad (3)$$

The denominator can be shown to go to infinity when A4 holds and the numerator can be shown to converge to zero at an exponential rate because the likelihood ratio statistic for testing the null $\mathbf{P} = \mathbf{P}^0$ against the alternative hypothesis that $\mathbf{P} \in U^c$ is uniformly consistent.

Theorem 2 (Proposition 4.2.1 of Ghosh and Ramamoorthi (2003)): Under the conditions of Theorem 1, $\bar{\mathbf{P}}(x_{T+1} \mid \mathbf{X}^T) \Rightarrow \mathbf{P}^0(x_{T+1})$ a.s. \mathbf{P}^0 , where \Rightarrow denotes weak convergence.

$$\text{The essence of the proof is that } \left\| \int_{\mathcal{P}^\theta} \mathbf{P} \Pi(\mathbf{P} \mid \mathbf{X}^T) \Pi(d\mathbf{P}) - \mathbf{P}^0 \right\| \leq \int_{\mathcal{P}^\theta} \|\mathbf{P} - \mathbf{P}^0\| \Pi(\mathbf{P} \mid \mathbf{X}^T) \Pi(d\mathbf{P})$$

by Jensen's inequality and the right hand side converges to zero.

How is weak convergence of the predictive distribution relevant to the Bayesian archetype? A partial answer is given by the following theorem.

Theorem 3 (Propositions 2.6–2.14 of Berger and Salinetti (1995); Theorem 1 of Zervos (1999)): Suppose that V satisfies A1 and let \mathbf{P}_N be any sequence for which $\mathbf{P}_N \Rightarrow \mathbf{P}^0$. Then:

$$\sup_{a \in \mathcal{A}} \int_{\mathcal{X}} V(x, a) \mathbf{P}_N(dx) \rightarrow \sup_{a \in \mathcal{A}} \int_{\mathcal{X}} V(x, a) \mathbf{P}^0(dx) \text{ a.s.} \quad (4)$$

Theorem 2 of Zervos (1999) obtains this result under less restrictive conditions than A1.

Theorem 3 suggests that the relevance of the weak convergence criterion depends on the use to which the predictive distribution is being put. If the purpose is to learn \mathbf{P}^0 with high posterior probability, weak convergence is too weak: distributions in weak neighborhoods of \mathbf{P}^0 can look quite different from it. Dramatic examples can be found in Freedman (1963, 1965), Freedman and Diaconis (1983, 1986a,b), and Stinchcombe (2004).⁸ Measures in \mathcal{P}^θ that are

⁸ The fact that posterior convergence can fail even if the prior assigns positive mass to weak neighborhoods of \mathbf{P}^0 led Freedman (1965) to conclude that “for essentially any pair of Bayesians, each thinks the other is crazy” and Stinchcombe (2004) to say that such Bayesians engage in “erratic, wildly inconsistent, fickle, or faddish” behavior.

close to P^0 in the Prohorov, dual bounded Lipshitz, or other weak metric can be far from P^0 in a Kullback-Leibler sense and relative entropy is what is relevant for likelihood ratios.

However, this semiparametric Bayesian is using the model to solve a problem like (4). By the Portmanteau Theorem (Theorem 6.1 of Parthasarathy (1967) and Theorem 11.1.1 of Dudley (1989)), weak convergence implies that $\int f dP_N \rightarrow \int f dP \forall f \in \mathcal{F}$. If $\int g dP_N \rightarrow \int g dP = 0$ and $\int V_a dP_N \rightarrow \int V_a dP = 0$, where V_a is the set of Euler equations from (4), the archetype will learn the optimal decision rule asymptotically. To be sure, the archetype's prior might reflect *a priori* beliefs about other aspects of the measures in \mathcal{P}^θ , particularly with respect to their smoothness, but the archetype would want to ensure that the resultant prior would not interfere with weak convergence to the optimal decision rule. Achieving weak convergence when possible would appear to be a minimal condition for an inductive learning scheme to be deemed rational.

Moreover, the proviso that $\int f dP_N \rightarrow \int f dP \forall f \in \mathcal{F}$ is surely relevant for the econometrician as well. The econometrician is assuming that the semiparametric Bayesian is solving an optimization problem based on beliefs codified in the moment conditions but is making no assumptions regarding the functional form of the archetype's preferences. Consistent estimation of the probability measure by the econometrician is asymptotically equivalent to learning aspects of the beliefs of this semiparametric Bayesian that are relevant for optimal decisions irrespective of the specifics of the utility function in these circumstances.

C. Multinomial Approximation of Semiparametric Bayesian Beliefs

The weak topology is appropriate when interest centers on probability *measures* or *distributions*, not on probability *densities*; that is, weak convergence is equivalent to $\lim_{N \rightarrow \infty} P_N(\mathcal{A}) = P(\mathcal{A})$ for all Borel sets \mathcal{A} that have boundaries with P-measure zero and the natural

collection of Borel sets to contemplate is the partition of \mathcal{X} induced by sampling. The fact that the predictive distribution converges weakly to the associated countable cell multinomial links the asymptotic beliefs of the semiparametric Bayesian archetype to the probability measures estimated by the GMM econometrician.⁹ The purpose of this subsection is to make some connections that will prove useful in the sequel.

As is well-known, the set of discrete measures is dense in the space of all Borel probability measures on \mathcal{X} ; see, for example, Theorem 6.3 of Parthasarathy (1967) and Lemma 11.7.3 of Dudley (1989). A sequence of multinomial approximations can be constructed in the following manner. For each N , partition \mathcal{X} into a countable collection of Borel sets of the form:

$$\mathcal{X} = \bigcup_n \mathcal{X}_n^N; \quad \sup_{y, z \in \mathcal{X}_n^N} \|y - z\| \leq \frac{1}{N} \quad \forall n; \quad \mathcal{X}_m^N \cap \mathcal{X}_n^N = \emptyset \quad \forall m \neq n \quad (5)$$

where $\|\bullet\|$ is the usual Euclidean metric. For each n , choose any $x_n^N \in \mathcal{X}_n^N$ and set $P(x_n^N) = P(\mathcal{X}_n^N)$ so that $P_N = \sum_n P(x_n^N) \delta_{x_n^N}$, where $\delta_{x_n^N}$ is the Dirac measure at x_n^N , and the corresponding probability distribution is $F_N(x) = \sum_n P(x_n^N) 1_{x_n^N < x}$. The error in approximating any $f \in \mathcal{F}$ on \mathcal{X}_n^N by $f(x_n^N)$ is at most $\zeta_n^N = \sup(f | \mathcal{X}_n^N) - \inf(f | \mathcal{X}_n^N)$. Weak convergence obtains by the Portmanteau Theorem since $\left| \int f dP_N - \int f dP \right| \leq \sup_n \zeta_n^N \rightarrow 0$ as $N \rightarrow \infty$.

The likelihood for each P_N is given by $P_N(X^T) = \prod_t \prod_n P(\mathcal{X}_n^N)^{1_{x_t \in \mathcal{X}_n^N}}$. If, in addition, $N \gg T$, each set \mathcal{X}_n^N will contain at most one observation from X^T . Letting \mathcal{X}_{nt}^N denote the cell with $x_t \in \mathcal{X}_n^N$, $P_N(X^T) = \prod_{t \leq T} P(\mathcal{X}_{nt}^N)$. Thus the posterior and predictive distributions can be approximated on (5) by:

⁹ Chamberlain (1987) used multinomial approximation to study semiparametric efficiency but not on the weak topology, forming a neighborhood base for P^0 with measurable and integrable, not bounded or continuous, functions.

$$\begin{aligned}\Pi_N(\mathbf{P} \in \mathcal{Q} | \mathbf{X}^T) &= \frac{\int_{\mathcal{Q}} \prod_{t \leq T} P(\mathcal{X}_{nt}^N) \Pi(d\mathbf{P})}{\int_{\mathcal{P}^\theta} \prod_{t \leq T} P(\mathcal{X}_{nt}^N) \Pi(d\mathbf{P})} \Rightarrow \Pi(\mathbf{P} \in \mathcal{Q} | \mathbf{X}^T) \\ \bar{\mathbf{P}}_N(\mathbf{x}_{T+1} \in \mathcal{X}_n^N | \mathbf{X}^T) &= \int_{\mathcal{P}^\theta} P(\mathbf{x}_{T+1} \in \mathcal{X}_n^N) \Pi_N(\mathbf{P} | \mathbf{X}^T) \Pi(d\mathbf{P}) \Rightarrow P^0(\mathbf{x}_{T+1} \in \mathcal{X}_n^N)\end{aligned}\quad (6)$$

with convergence following from the Portmanteau Theorem and Theorems 1 and 2, respectively.¹⁰

For comparability with the GMM econometrician, it makes sense to consolidate the countable partition (5) into a smaller set of partitions centered on the observations in the sample \mathbf{X}^T . To be concrete, aggregate (5) into an ‘asymptotic’ Voronoi tessellation:

$$\mathcal{X} = \bigcup_t \mathcal{X}_t^T; \mathbf{x}_t \in \mathcal{X}_t^T; \mathcal{X}_t^T = \bigcup_n \left\{ \mathcal{X}_n^N : \|\mathbf{x}_n^N - \mathbf{x}_t\| \leq \|\mathbf{x}_n^N - \mathbf{x}_s\| \forall n, s \neq t \right\} \quad (7)$$

for sufficiently large N .¹¹ The associated multinomial probabilities over (7) can be taken to be $P_N(\mathcal{X}_t^T) = P(\mathcal{X}_{nt}^N)$. Alternatively, the semiparametric Bayesian can group within cells and set $P_N(\mathcal{X}_t^T) = P(\bigcup_n \mathcal{X}_n^N) = \sum_n P(\mathcal{X}_n^N)$ for all \mathcal{X}_n^N allocated to \mathcal{X}_t^T .¹² Either way, $\prod_{t \leq T} P_N(\mathcal{X}_t^T)$ is an approximate likelihood function for each $\mathbf{P} \in \mathcal{P}^\theta$ and the aggregated multinomial probabilities live on the associated standard T-simplex $\{\mathcal{S}_T^N : P_N(\mathcal{X}_t^T) > 0, \sum_t P_N(\mathcal{X}_t^T) = 1\}$.

The resulting approximate posterior and predictive distributions are given by:

$$\begin{aligned}\Pi_N(\mathbf{P} \in \mathcal{Q} | \mathbf{X}^T) &= \frac{\int_{\mathcal{Q}} \prod_{t \leq T} P_N(\mathcal{X}_t^T) \Pi(d\mathbf{P})}{\int_{\mathcal{P}^\theta} \prod_{t \leq T} P_N(\mathcal{X}_t^T) \Pi(d\mathbf{P})} \Rightarrow \Pi(\mathbf{P} \in \mathcal{Q} | \mathbf{X}^T) \\ \bar{\mathbf{P}}_N(\mathbf{x}_{T+1} \in \mathcal{X}_t^T | \mathbf{X}^T) &= \int_{\mathcal{P}^\theta} P(\mathbf{x}_{T+1} \in \mathcal{X}_t^T) \Pi_N(\mathbf{P} | \mathbf{X}^T) \Pi(d\mathbf{P}) \Rightarrow P^0(\mathbf{x}_{T+1})\end{aligned}\quad (8)$$

which converge weakly as well. It could be that (6) and (8) approximate their continuous analogues. It could be that the semiparametric Bayesian approximates the ‘true’ posterior in this

¹⁰ See also Theorem 4.1 of Diaconis and Freedman (1986b) for a related multinomial approximation in a Bayesian context based on discretization of both the sample space and the space of probability measures.

¹¹ This is not a standard Voronoi tessellation. The word ‘asymptotic’ and the large N requirement arise because each \mathcal{X}_n^N is allocated to only one \mathcal{X}_t^T and there will be points in \mathcal{X}_n^N closer to some other \mathcal{X}_s^T because the diameter of \mathcal{X}_n^N is $1/N$. When N is large, any such tie-breaking rule will suffice.

¹² Grouping is a coarse way of smoothing but is consistent with multinomial approximation.

fashion. In either case, the archetype consistently estimates the probability measure of x_t .

D. Multinomial Approximation and the GMM Econometrician

Now consider a second set of T-cell multinomial distributions on (7) given by $\{\mathcal{S}_T^G : P_t^T(\theta) > 0, \sum_{t \leq T} P_t^T(\theta) = 1, \sum_{t \leq T} P_t^T(\theta) g(x_t, \theta) = 0, \theta \in \Theta\}$, where the requirement that the moment conditions hold for each T makes \mathcal{S}_T^N differ from \mathcal{S}_T^G . However, $\mathcal{S}_T^N \Rightarrow \mathcal{S}_T^G$ – that is, $P_N(\mathcal{X}_t^T) \Rightarrow P_t^T(\theta_p) \forall P \in \mathcal{P}^\theta$ – as the cell diameters shrink to zero. This is the large sample link between probability models of the archetype and the GMM econometrician.¹³

There is a very simple theorem that provides considerable insight into the structure of \mathcal{S}_T^G . Let $\bar{g}_T(\theta) = \frac{1}{T} \sum_{t \leq T} g(x_t, \theta)$ and $V_T(\theta) = \frac{1}{T} \sum_{t \leq T} [g(x_t, \theta) - \bar{g}_T(\theta)][g(x_t, \theta) - \bar{g}_T(\theta)]'$. Then:

Theorem 4: Let $\mathcal{S}_T^G(\theta)$ be the subset of \mathcal{S}_T^G for a given $\theta \in \Theta$. Then each

$\{P_1^T(\theta), P_2^T(\theta), \dots, P_T^T(\theta)\} \in \mathcal{S}_T^G(\theta)$ satisfies:

$$\begin{aligned} P_t^T(\theta) &= \frac{1}{T} - \frac{1}{T} \bar{g}_T(\theta)' V_T(\theta)^{-1} [g(x_t, \theta) - \bar{g}_T(\theta)] + \varepsilon_t(P) \\ \sum_{t \leq T} \left(P_t^T(\theta) - \frac{1}{T}\right)^2 &= \frac{1}{T} \bar{g}_T(\theta)' V_T(\theta)^{-1} \bar{g}_T(\theta) + T \sigma_{\varepsilon_T}^2(P); \quad \sigma_{\varepsilon_T}^2(P) = \frac{1}{T} \sum_{t \leq T} \varepsilon_t(P)^2 \end{aligned} \quad (9)$$

where the residuals satisfy $\frac{T-1}{T} < \bar{g}_T(\theta)' V_T(\theta)^{-1} [g_t(\theta) - \bar{g}_T(\theta)] - \varepsilon_t(P) < \frac{1}{T}$ so that $P_t^T(\theta) > 0$.¹⁴

Proof: Trivial application of the normal equations of multiple regression with an intercept.

This is an arithmetic result: all multinomial probabilities based on the same value of θ

¹³ A discrete prior can be formed over the T-simplex \mathcal{S}_T^G (or, for that matter, on the N-simplex over (5) constrained to satisfy the moment conditions). $\text{Diam}(\mathcal{S}_T^G) \leq 2$ but the probabilities are $O(T^{-1})$ and so $\text{Diam}(\mathcal{S}_T^G) = O(\sqrt{T})$, bounding \mathcal{S}_T^G by the positive orthants of spheres of the form $\sum_i p_i^2 = O(T^{-1})$. Letting $H^2(x, y) = \sum_t (\sqrt{x_t} - \sqrt{y_t})^2$ be the squared Hellinger metric, \mathcal{S}_T^G can be covered by $\{P_m^N : H^2(P_k^T, P_m^T) \geq \delta > 0 \forall k \neq m, m = 1, \dots, M\}$ and an approximate prior is given by $\Pi_T(\delta) = \sum_m \Pi_m^T P_m^T$ where $\Pi_m^T > 0$ and $\sum_m \Pi_m^T = 1$. $\Pi_T(\delta)$ satisfies A4 because the Hellinger distance bounds relative entropy. See Ghosh and Ramamoorthi (2003) for a version of $\Pi_T(\delta)$ with a prior over a random number of cells. See also Schennach (2005).

¹⁴ The awkward notation $\varepsilon_t(P)$ arises because there will generally be many $P \in \mathcal{P}^\theta$ for each value of θ .

have the same fitted value $\frac{1}{T} - \frac{1}{T} \bar{g}_T(\theta)' V_T(\theta)^{-1} [g_t(\theta) - \bar{g}_T(\theta)]$, termed implied probabilities by Back and Brown (1993), where the residuals $\varepsilon_t(P)$ are identically zero if the implied probabilities are all positive. In large samples, $\bar{g}_T(\theta)' V_T(\theta)^{-1} [g_t(\theta) - \bar{g}_T(\theta)] = o_p(1)$ for values of θ in shrinking neighborhoods of θ_{p^0} . For such values of θ and for P in shrinking neighborhoods of P^0 , nonzero values of $\varepsilon_t(P)$ are a small sample event because $E_{p^0}[g(x_t, \theta_{p^0})] = 0$. The fact that (9) holds for any numbers $\{p_t : \sum_t p_t = 1, \sum_t p_t g_t = 0\}$ – that is, p_t need not be positive – suggests that this regression structure can be useful for interpreting their sample analogues.¹⁵

The regression sum of squares can be interpreted along similar lines. The sum of squared differences between $P_t^T(\theta)$ and $1/T$ is proportional to what Owen (1991) termed Euclidean likelihood. In large samples, it is proportional to the ϕ - or f -divergences introduced by Csiszár (1967), making for a connection with a rich literature on estimation and testing based on the minimization of empirical divergences. These divergences are defined by the discrepancy functions $\varphi(\frac{p}{q}) \equiv \varphi(z) > 0$ where p and q are two densities defined on the same sample space and where $\varphi(\bullet)$ is continuous, convex, and twice differentiable and normalized so that $\varphi(1) = \varphi'(1) = 0$ and $\varphi''(1) = 1$. The term discrepancy serves as a reminder that $\varphi(\bullet)$ need not possess either the symmetry or triangle inequality properties of a metric.¹⁶

The scaled divergence between discrete measures with probabilities p_t and q_t is measured by $D_T^\varphi(z) = 2E_q[\varphi(z)] = 2\sum_t q_t \varphi(z_t)$ and a Taylor series expansion yields:

¹⁵ The regression structure of multinomial probabilities can also be used in prior construction if the zero covariance between $g_t(\theta)$ and $\varepsilon_t(P)$ is strengthened to $E[\varepsilon_t(P) | g(x_t, \theta)] = 0$.

¹⁶ The smoothness assumption rules out weak metrics such as the Kolmogorov, Levy, Prohorov, and dual bounded Lipschitz; Donoho and Liu (1988) discuss how such metrics can produce poorly behaved minimum distance estimates. It contains the convex members of the Cressie-Read (1988) power divergence family for which $\varphi(z)$ is an affine function of z^α including the likelihood divergence, entropy or Kullback-Leibler information, the Hellinger metric, and Pearson's and Neyman's modified χ^2 .

$$\begin{aligned}
D_T^\theta(z) &= 2 \sum_t q_t \varphi(z_t) = 2 \sum_n q_t [\varphi(1) + \varphi'(1)(z_t - 1) + \frac{1}{2} \varphi''(\zeta_t)(z_t - 1)^2] \\
&= \sum_t q_t (z_t - 1)^2 + \sum_t q_t [\varphi''(\zeta_t) - 1] (z_t - 1)^2
\end{aligned} \tag{10}$$

where ζ_t is between 1 and z_t . When $p_t = P_t^\top(\theta)$ and $q_t = \frac{1}{T}$, (10) takes the form:

$$D_T^\theta(z) = T \sum_t \left[P_t^\top(\theta) - \frac{1}{T} \right]^2 + T \sum_t [\varphi''(\zeta_t) - 1] \left[P_t^\top(\theta) - \frac{1}{T} \right]^2 \tag{11}$$

for all $\{\theta, \mathcal{S}_T^G(\theta)\}$. If $\sup_t \left| P_t^\top(\theta) - \frac{1}{T} \right| = o_p(1)$ and if $\varphi''(z)$ is bounded in the neighborhood of unity, the second term in (11) converges to zero uniformly. The leading term is proportional to Pearson's χ^2 divergence, which suggests that the decomposition of the regression sum of squares in the second line of (9) can provide insight into sample analogues as well.

3. The Distorted Beliefs Interpretation of Hypothesis Tests and Confidence Regions

Section 2 provided a framework for interpreting estimates of a probability measure that satisfies given moment conditions along the lines of Back and Brown (1992). Confronting uncertainty in the form of random variables $x_t \stackrel{\text{iid}}{\sim} P^0$, a generic expected utility maximizer forms a prior distribution over \mathcal{P}^θ because this semiparametric Bayesian believes $P^0 \in \mathcal{P}^\theta$. In this setting, the archetype's predictive distribution converges weakly to P^0 when $P^0 \in \mathcal{P}^\theta$ under the sole condition that the prior assigns positive probability to all Kullback-Leibler neighborhoods of P^0 . Moreover, discrete approximations to the predictive distribution converge weakly as well and, as a consequence, converge to the GMM estimate of P^0 . The restrictions on the preferences and prior beliefs of this hypothetical semiparametric Bayesian would appear to be quite weak.

However, the archetype is a construct, a hypothetical Bayesian econometrician looking at the same data as the GMM econometrician. It is the large sample connection between the two that forms the framework proposed here: the notion that $\mathcal{S}_T^N \uparrow \mathcal{P}^\theta$ and that $\mathcal{S}_T^N \Rightarrow \mathcal{S}_T^G$. One way to exploit this insight is to actually do the work of the semiparametric Bayesian and replicate the

multinomial construction in 2C on (5) or (7) or on some other appropriate partition of the sample space. Such an analysis would require much more than the characterizations in 2B and 2C; it would necessitate formulating priors that satisfied Assumption A4 (or some analogue of it) without placing additional substantive restrictions. While it is possible to do so along the lines of footnotes 13 and 15, this sort of analysis is beyond the scope of the present paper.

This section is devoted to a discussion of the alternative hypothesis that motivated the paper: that plausible differences between the *a posteriori* beliefs of a hypothetical semiparametric Bayesian and a GMM econometrician can inform estimation and inference in GMM settings. The next subsection provides a distorted beliefs interpretation of confidence regions and goodness-of-fit statistics based on the regression sum of squares in (9). The final subsection discusses some of the uses of the corresponding residuals.

A. Test Statistics and Confidence Regions

Let $\hat{z}_T^\varphi = \{\hat{\theta}_T^\varphi, \{\hat{P}_t^\top(\hat{\theta}_T^\varphi), t \leq T\}\} = \arg \min_{\{\theta, \{P(\theta), t \leq T\}\}} D(z_T^\varphi)$ and note that $TD(\hat{z}_T^\varphi)$ is given by:

$$\begin{aligned} TD(\hat{z}_T^\varphi) &= T^2 \sum_t \left(\hat{P}_t^\top(\hat{\theta}_T^\varphi) - \frac{1}{T} \right)^2 + T^2 \sum_{t \leq T} \left\{ \varphi''[\xi_t(\hat{z}_T^\varphi) - 1] \left(\hat{P}_t^\top(\hat{\theta}_T^\varphi) - \frac{1}{T} \right) \right\}^2 \\ &= T \bar{g}_T(\hat{\theta}_T^\varphi)' V_T(\hat{\theta}_T^\varphi)^{-1} \bar{g}_T(\hat{\theta}_T^\varphi) + T^3 \sigma_{\varepsilon_T}^2 (\hat{P}_T^\top) + T^2 \sum_{t \leq T} \left\{ \varphi''[\xi_t(\hat{z}_T^\varphi) - 1] \left(\hat{P}_t^\top(\hat{\theta}_T^\varphi) - \frac{1}{T} \right) \right\}^2 \quad (12) \\ &\rightarrow T \bar{g}_T(\hat{\theta}_T^\varphi)' V_T(\hat{\theta}_T^\varphi)^{-1} \bar{g}_T(\hat{\theta}_T^\varphi) \sim \chi^2(p-q) \end{aligned}$$

where the second line involves the substitution of (9) into (12) and the convergence to a $\chi^2(p-q)$ random variable obtains if the moment conditions are valid because $\xi_t = 1 + o_p(1)$ and $T^3 \sigma_{\varepsilon_T}^2 (\hat{P}_T^\top) = o_p(1)$. Hence, $TD(\hat{z}_T^\varphi)$ differs from the GMM overidentifying restrictions test statistic in the presence of these two $o_p(1)$ terms and in the choice of estimator and covariance matrix – $\arg \min_{\theta} \bar{g}_T(\theta)' S_T(\tilde{\theta})^{-1} \bar{g}_T(\theta)$ instead of $\hat{\theta}_T^\varphi$ and $S_T(\tilde{\theta}_T) = \frac{1}{T} \sum_{t \leq T} g(x_t, \tilde{\theta}_T) g(x_t, \tilde{\theta}_T)'$ in place of $V_T(\hat{\theta}_T^\varphi)$, where $\tilde{\theta}_T$ is any \sqrt{T} consistent estimator of θ .

This large sample χ^2 test statistic can be used to test the null hypothesis and to construct confidence regions for θ under the null. Conventional practice is to select a significance level α and an associated critical value c^α that solves $\Pr(\chi_{p-q}^2 \geq c^\alpha) = \alpha$. The null hypothesis is rejected if $\text{TD}(\hat{z}_T^\varphi) > c^\alpha$ while the statistic fails to reject the null if $\text{TD}(\hat{z}_T^\varphi) \leq c^\alpha$. As is typically the case in likelihood-based inference, the rejection region can be viewed as the complement of the $1-\alpha$ per cent confidence region given by $\{\theta, \{P_t^T(\theta), t \leq T\} : \text{TD}(z_T^\varphi) \leq c^\alpha\}$.¹⁷

The link between \mathcal{S}_T^N and \mathcal{S}_T^G provides for an economic interpretation of rejections in this inference framework. The rejection region $\{\{P_t^T(\theta), t \leq T\} \in \mathcal{S}_T^G : \text{TD}(\hat{z}_T^\varphi) > c^\alpha\}$ is a subset of the T-cell multinomials in \mathcal{S}_T^G and $\mathcal{S}_T^N \Rightarrow \mathcal{S}_T^G$. The question at hand is simple: are there beliefs implicit in the rejection region that the econometrician would think that the archetype might reasonably possess *a posteriori*? Put differently, might the beliefs of such a Bayesian make a seemingly sharp rejection appear instead to be compatible with the data? Might there be plausible beliefs outside the associated $1-\alpha$ per cent confidence region?

This then is the main point of the paper. If the answer to these questions is “yes,” the econometrician could reasonably declare that the test statistic provided a *statistically* significant rejection at level α that should be thought of as *economically* insignificant. A similar statement applies to economically plausible beliefs that lie *outside* the confidence region that is the complement of the rejection region. An econometrician who did not want to draw sharp conclusions about economic as opposed to statistical significance could simply report summary

¹⁷ The empirical likelihood ratio statistic – that is, $\text{TD}(\hat{z}_T^\varphi)$ with $\varphi(z) = \ln(z)$ – is Bartlett correctable; see Chen and Cui (2006) for the moment condition version of this result. Its mean is $E\{\text{TD}(\hat{z}_T^{\log})\} = q(1 + B_c T^{-1}) + O(T^{-2})$ in large samples and the Bartlett correction takes the form $\Pr[\text{TD}(\hat{z}_T^{\log}) \leq c^\alpha(1 + \hat{B}_c T^{-1})] = \alpha + O(T^{-2})$, where \hat{B}_c can be obtained from the bootstrap. There is a subtle issue here; the prior also influences second order inference in this setting. Under suitable regularity conditions, the Bartlett-corrected empirical likelihood rejection region would be the appropriate object of inference if the prior was sufficiently flat in the neighborhood of the optimum.

statistics describing the beliefs that seem to be sufficiently compatible with the data.

One such summary statistic involves the comparison of the sample relative entropy $\frac{1}{T} \sum_t \ln \hat{P}_t^T(\hat{\theta}_T^\varphi) - \frac{1}{T} \ln \frac{1}{T}$ based on the estimate $\hat{\theta}_T^\varphi$ with that of a distribution that is more easily interpreted. McCulloch (1989) suggested one such calibration: compare the sample relative entropy with that from a hypothetical binomial experiment in which the null success probability is $\frac{1}{2}$ and the sample success probability is q with q selected so that:

$$\frac{1}{T} \sum_t \ln \hat{P}_t^T(\hat{\theta}_T^\varphi) = \frac{1}{2} [\ln \frac{1}{2} - \ln(1-q)] + \frac{1}{2} [\ln \frac{1}{2} - \ln q] = \frac{1}{2} \ln \frac{1}{2} - \frac{1}{2} \ln[q(1-q)] - \frac{1}{T} \ln \frac{1}{T} \quad (13)$$

The presumption is that values of q close to $\frac{1}{2}$ suggest that a sample entropy that is statistically significant at level α is small in this alternative metric.

A similar calibration can be based on the multivariate normal distribution for which the entropy is $\frac{d}{2} \ln 2\pi e + \ln |\Sigma|$ where Σ is the covariance matrix. Hence:

$$\frac{1}{T} \sum_{t \leq T} \ln \hat{P}_t^T(\hat{\theta}_T^\varphi) = \frac{d}{2} \ln 2\pi e + \ln |\Sigma| \quad (14)$$

can be solved for $|\Sigma|$, which, in turn, can be compared with the restricted estimate $|\hat{\Sigma}|$ from:

$$\hat{\Sigma} = \sum_{t \leq T} \hat{P}_t^T(\hat{\theta}_T^\varphi) (x_t - \hat{\mu})(x_t - \hat{\mu})' \quad (15)$$

where $\hat{\mu} = \sum_{t \leq T} \hat{P}_t^T(\hat{\theta}_T^\varphi) x_t$ is the restricted estimate of the mean. Here, too, sufficiently small differences between $|\Sigma|$ and $|\hat{\Sigma}|$ suggest that the difference between the two is “reasonably small” in this alternative metric.

B. Residual Analysis

Reasonable *a posteriori* probability beliefs can be assessed via relations (9) and (12). The relative contributions of the fitted values $\frac{1}{T} - \frac{1}{T} \bar{g}_T(\hat{\theta}_T^\varphi)' V_T(\hat{\theta}_T^\varphi)^{-1} [g(x_t, \hat{\theta}_T^\varphi) - \bar{g}_T(\hat{\theta}_T^\varphi)]$ and the residuals $\hat{\varepsilon}_t^T(\hat{\theta}_T^\varphi)$ are given in (9) and values of either that are large in absolute value have a

disproportionate impact on the $\hat{P}_t^T(\hat{\theta}_t^\phi)$ estimates and their associated sample entropy. Large residuals may be especially informative since the residuals are identically zero if the implied probabilities lie between zero and one. The incremental impact of alternative divergences can be examined via the observed covariance between the excess curvature $\phi''[\xi_t(\hat{z}_t^\phi)]-1$ and $(\hat{P}_t^T(\hat{\theta}_t^\phi) - \frac{1}{T})^2$ as codified in (12).¹⁸

The whole probability simplex \mathcal{S}_T^G can be investigated in this fashion. Plausible values of θ might be suggested by theory or introspection or might be obtained by bootstrapping, which is rigorously justifiable under A3. For each θ , $\mathcal{S}_T^G(\theta)$ can be explored by enumerating sets of residuals $\varepsilon_t(P)$ that sum to zero, are orthogonal to $g(x_t, \theta)$, and satisfy the lower and upper bound constraints, which can then be examined with the regression diagnostics. Conditioning on θ is the best way to explore \mathcal{S}_T^G if $E[\varepsilon_t(P)g(x_t, \theta)] = 0$ is strengthened to $E[\varepsilon_t(P) | g(x_t, \theta)] = 0$.

Implicit in this discussion is a particular concern for the effect of outliers on probabilities, which play a special role in models that incorporate expectations. As Back and Brown (1993) emphasized, outliers in this setting represent data that are not representative of the underlying population when the moment conditions are true. In rational expectations models, data that are underrepresented – that is, those for which $P_t^T(\theta) - \frac{1}{T}$ is large – are often thought to represent *peso problems*, events that were expected to happen but that did not eventuate or that did not

¹⁸ There is a suggestive interpretation of alternative divergences that is hard to make rigorous without taking a stand on priors. Priors generated via ex ante maximization of the expected distance, such as the Kullback-Leibler or χ^2 divergence, between prior and posterior over the sample space are called reference or default priors; see Bernardo (2005) for a survey and Kuboki (1998) for an application to parametric prediction. Maximizing this distance is analogous to minimizing the distance between \bar{P}_T and the true distribution P^0 . The empirical distribution $\hat{P} \Rightarrow P^0$ and $\hat{P}(\hat{\theta}_t^\phi) \Rightarrow \bar{P}_T$ even under the alternative. In this heuristic sense, cells for which $\phi(\hat{z}_t)$ is small are ones for which the data dominate the prior in the distance as measured by $\phi(\bullet)$ while those for which $\phi(\hat{z}_t)$ is large are ones for which the apparent impact of the prior remains sizeable. It is a considerable leap to go beyond these heuristics to an actual reference prior for semiparametric Bayesian prediction and a corresponding assessment of the impact of the prior in a given sample. The ideas in footnotes 13 and 15 are one place to start.

occur as frequently as expected. For example, the Great Depression might represent a recurrent rare event or one that will succumb to the law of large numbers. Accordingly, we might reasonably expect the prior predictive probability $\bar{P}_N(x_{T+1} \in \mathcal{X}_t^T) = \int_{\mathcal{P}^\theta} P(x_{T+1} \in \mathcal{X}_t^T) \Pi(dP)$ – that is, the predictive distribution for x_{T+1} in the absence of sample information – for some such \mathcal{X}_t^T to be much larger than the empirical probability $\frac{1}{T}$, resulting in a seemingly large value of $P_t^T(\theta)$. Note also that $\bar{P}_N(\bullet)$ is the *posterior* predictive probability outside the convex hull of the data.

4. Conclusion

This paper was based on a simple intuition. What can we learn from probability statements about sample moment conditions in rational expectations models under the maintained hypothesis that the moment conditions are true? The answer is straightforward: modulo sampling error, sample moments reflect biases in the expectations of the relevant economic actors in these circumstances. This distorted beliefs alternative would appear to be an interesting one, if only because it provides one dimension in which to distinguish between economic and statistical significance. All that is needed is a way to measure the attributes of expectations compatible with the moment conditions.

The attainment of this goal required a detour down the path of Bayesian semiparametrics. Models based on moment conditions do not deliver likelihoods and the strict application of the Bayesian calculus requires their specification. Moreover, the formation of prior beliefs is more challenging in such settings because the data need not swamp the prior when priors are over spaces of probability measures. Finally, the literature on priors for semiparametric models is thin and a broad set of priors would appear to be necessary when seeking to characterize the extent to which the expectations compatible with a given set of moment conditions are “nearly rational.”

Two attributes of the archetypical semiparametric Bayesian constructed in Section 2 eliminated these problems. The first was the presumption that the archetype was a consumer of economic theory who used the model based on moment conditions solely for forecasting. The second was the shift from densities that respect the moment conditions to discrete measures that do so. The resulting predictive distribution based on a countable set of multinomial approximations proves to be consistent under the weak restriction that the prior assigns positive probability to all Kullback-Leibler neighborhoods of the true distribution. While this observation is hardly surprising in finite-dimensional parametric settings, it is somewhat more remarkable in this semiparametric setting in which the typical requirement is far more stringent.

The result is a semiparametric Bayesian interpretation of probability estimates provided by empirical likelihood and related minimum divergence methods. From this perspective, rejection and confidence regions are comprised of probability beliefs, not parameter values, beliefs an econometrician can examine for their plausibility. This association of plausible beliefs with such regions yields a framework for assessing the economic significance of distorted beliefs.

Let me conclude by suggesting three ways in which research along these lines can proceed. First, it would be useful to have additional analytical tools beyond those described in Section 3. Second, statistics other than omnibus goodness-of-fit tests can be examined in this fashion but the difference between the Bayesian and frequentist treatment of nuisance parameters might make it more difficult to equate the beliefs of the archetype and the GMM econometrician. Finally, a more interesting archetype might be one with the same objectives but whose decisions affect sample outcomes as is the case in rational expectations models with learning or in Kurz's (1997) rational beliefs equilibria. Here, too, it might well be substantially more challenging to equate the beliefs of a semiparametric Bayesian and the GMM econometrician.

References

- Back, K. and Brown, D. P., 1993. "Implied probabilities in GMM estimators," *Econometrica* 61, 971-975.
- , 1992. "GMM, maximum likelihood, and nonparametric efficiency," *Economics Letters* 39, 23-28.
- Berger, J. O., and G. Salinetti, 1995. "Approximations of Bayes decision problems: the epigraphical approach," *Annals of Operations Research* 56, 1-13.
- Bernardo, J. M., 2005. "Reference analysis," in Dey, D. K. and C. R. Rao (eds.). *Handbook of Statistics, Volume 25: Bayesian Thinking, Modeling, and Computation*. Amsterdam: Elsevier, 17-90.
- Chamberlain, G., 1987. "Asymptotic efficiency in estimation with conditional moment restrictions," *Journal of Econometrics* 34, 305-334.
- Csiszár, I., 1967. "Information-type measures of difference of probability distributions and indirect observations," *Studia Scientifica Materia Hungaria* 2, 299-318.
- Diaconis, P. and D. Freedman, 1983. "On Inconsistent Bayes Estimates in the Discrete Case," *Annals of Statistics* 11, 1109-1118.
- , 1986a. "On the Consistency of Bayes Estimates," *Annals of Statistics* 14, 1-26.
- , 1986b. "On Inconsistent Bayes Estimates of Location," *Annals of Statistics* 14, 68-87.
- Donoho, D. L., and Liu, R. C., 1988. "Pathologies of some minimum distance estimators," *Annals of Statistics* 16, 587-608.
- Dudley, R. M., 1989, *Real Analysis and Probability*, (New York: Chapman and Hall).
- Engle, R. F., D. F. Hendry, and J.-F. Richard, 1983. "Exogeneity," *Econometrica* 51, 277-304.
- Freedman, D., 1963. "On the Asymptotic Behavior of Bayes Estimates in the Discrete Case I," *Annals of Mathematical Statistics* 34, 1386-1403.
- , 1965. "On the Asymptotic Behavior of Bayes Estimates in the Discrete Case II," *Annals of Mathematical Statistics* 36, 454-456.
- Ghosh, J. K. and R. V. Ramamoorthi, 2003. *Bayesian Nonparametrics*. Springer, New York.
- Hansen, L. P., 1982. "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica* 50, 1029-1054.

- Kim, J.-Y., 2002. "Limited information likelihood and Bayesian Analysis," *Journal of Econometrics* 107, 175-193.
- Kitamura, Y., 2006. "Empirical Likelihood Methods in Econometrics: Theory and Practice," Cowles Foundation Discussion Paper No. 1569, Department of Economics, Yale University.
- Kuboki, H., 1998. "Reference priors for prediction," *Journal of Statistical Planning and Inference* 69, 295-317.
- Kurz, M., (ed.), 1997. *Endogenous Economic Fluctuations: Studies in the Theory of Rational Beliefs*. Springer, New York.
- Lazar, N., 2003. "Bayesian empirical likelihood," *Biometrika* 90, 319–26.
- McCulloch, R. E., 1989. "Local Model Influence," *Journal of the American Statistical Association*, 84, 473-478.
- Owen, A., 1991. "Empirical Likelihood for Linear Models," *Annals of Statistics* 19, 1725-1747.
- , 2001. *Empirical Likelihood* (New York: Chapman and Hall).
- Parthasarathy, K. R., 1967, *Probability Measures on Metric Spaces*, (New York: Academic Press).
- Read, T. R. C., and N. A. C. Cressie, 1988. *Goodness-of-fit statistics for discrete multivariate data* (New York: Springer-Verlag).
- Schennach, S. M., 2005. "Bayesian exponentially tilted empirical likelihood," *Biometrika* 92, 31–46.
- Schwartz, L., 1965. "On Bayes procedures," *Z. Wahrsch. Verw. Gebiete* 4, 10-26.
- Stinchcombe, M., 2004. "The unbearable flightiness of Bayesians: generically erratic updating," working paper, Department of Economics, University of Texas at Austin.
- Zellner, A., 1994. "Model, prior information and Bayesian analysis," *Journal of Econometrics* 75, 51–68.
- , 1997. "The Bayesian method of moments (BMOM): theory and application," in Fomby, T., and Hill, R. C. (eds.). *Advances in Econometrics*. Cambridge University Press.
- Zervos, M., 1999. "On the Epiconvergence of Stochastic Optimization Problems," *Mathematics of Operations Research* 24, 495-508.