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.

The Role of Beliefs in Inference for Rational Expectations Models
Bruce N. Lehmann
NBER Working Paper No. 11757
November 2005
JEL No. C1, C2, C3, C4, C5, E2, G1

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 of the model.

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

The following inference setting pervades modern empirical work in economics involving rational expectations and market efficiency. An economic model of the form $E^*[g_t^\theta \mid \mathscr{F}_{t-1}] = 0$ where $g_t^\theta \equiv g(x_t, \theta)$ is assumed to describe relations among a set of observables x_t and where * and $\mathscr{F}_{t-1}$ reflect the beliefs of and information available to relevant economic agents in the model, respectively. The model is closed by rational expectations – that is, by equating $E^*[\bullet \mid \mathscr{F}_{t-1}] = 0$ with the objective conditional expectation $E[\bullet \mid \mathscr{F}_{t-1}] = 0$ – and then estimated by the generalized method of moments (GMM) of Hansen (1982). The GMM machinery provides for large sample inference, including distribution theory for the estimates along with goodness-of-fit measures such as the overidentifying restrictions test for the proximity of the sample mean of g_t^θ to zero.

This paper provides a coherent economic interpretation of this inference setting under both the null model and, more importantly, the alternative hypothesis that the model is false. The internally consistent interpretation treats the model g_t^θ as a maintained hypothesis, making the null hypothesis rational expectations and the alternative non-rational expectations. The estimation framework is a family of minimum distance analogues of GMM that share its asymptotic properties. Such methods involve the minimization of some measure of discrepancy between the usual empirical distribution function and an estimate of it that satisfies the *a priori* moment restrictions. The resulting probabilities estimate the beliefs compatible with the moment conditions that are closest to rational in the metric defined by the distance or discrepancy function. This approach makes it possible to measure the economic significance of any violations of the null hypothesis via the degree to which these implied beliefs are distorted.

This paper traces out the implications of this distorted beliefs alternative for inference in such models. The next section is really a character sketch: it describes the structure of the beliefs of a hypothetical semiparametric Bayesian. The penultimate section discusses the way in which the beliefs of this Bayesian archetype 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

Three attributes characterize a Bayesian decision maker: a preference ordering over the possible outcomes associated with given actions, a set of constraints governing these actions, and probability beliefs regarding the likelihood of each outcome. The econometrician presumes that this archetypical Bayesian follows Bayes' rule and believes that g_t^{θ} is a martingale difference sequence but otherwise has little *a priori* knowledge of either the probability beliefs or the underlying decision problem of this agent. This hypothetical semiparametric Bayesian would formulate prior beliefs over both the distributions compatible with and the parameters underlying the moment conditions. These prior beliefs would be combined with these distributions, viewed as likelihood functions for a sample of data relevant to the decision problem, to arrive at a posterior distribution for use in drawing inferences about the data generating process. The other attribute ascribed to this archetypical Bayesian is the use to which this family of models will be put: this Bayesian's sole interest is in the use of these models for forecasting.

What follows is really a character sketch: a description of the essential attributes of this archetypical semiparametric Bayesian. The next three subsections lay out the preferences, probability models, and priors that seem to me to provide the least restrictive portrait that still has nontrivial empirical implications. The final subsection proves the only theorem of the section, one which describes the (remarkably weak) restrictions placed on prior beliefs to deliver convergence of the predictive distribution to the true distribution in this semiparametric setting.

A. Preferences

The first observation is an old one: a semiparametric Bayesian interested in using the moment conditions to provide the best *a posteriori* model for forecasting will use the so-called predictive distribution. Let $\{y_t,a_t\}$ where y_t is a vector of variables that define the relevant states of nature or that are useful for forecasting and where a_t is a vector of actions that can be taken by this Bayesian in light of the realizations of y_t. The underlying conditioning information $\mathscr{F}_{t-1}$ includes lagged values of the state variables $Y^{t-1} = \{y_s, s \le t-1\}$ but not lagged values of the actions $\{a_s, s \le t-1\}$. That is, actions may affect the state variables in a deterministic way through budget constraints and the like but do not affect the random evolution of the state vector, an assumption that is appropriate when the semiparametric Bayesian is a small player in this economic environment. Finally, suppose this archetypical Bayesian is a subjective expected utility maximizer for whom the expected present value of net benefits may be represented as:

$$V(y_t,a_t,t) = \max_{a_t \in A(y_t)} U(y_t,a_t,t) + \beta E[V(y_{t+1},a_{t+1},t+1) \,|\, \mathscr{F}_t] \qquad (1)$$

where $V(\bullet,\bullet,\bullet)$ is the value function associated with the Bayesian's dynamic program, $U(\bullet,\bullet,\bullet)$ is the per period reward function, and $E[\bullet|\mathscr{F}_t]$ is the conditional expectation, the characterization of which is the matter at hand.

There are two uncertainties underlying this conditional expectation: the state of the world y_{t+1} that will be realized next period and the probability model $P^\theta \in \mathcal{P}^\theta$ that is the data generating process. Hence, the conditional expectation of the value function next period takes the form:

$$\begin{aligned} E[V(y_{t+1},a_{t+1},t+1) \,|\, \mathscr{F}_t] &= \int_{\mathcal{Y}} \int_{\mathcal{P}^\theta} V(y_{t+1},a_{t+1},t+1) P^\theta(y_{t+1} \,|\, \mathscr{F}_t) d\Pi(P^\theta) dy_{t+1} \\ &= \int_{\mathcal{Y}} V(y_{t+1},a_{t+1},t+1) \left\{ \int_{\mathcal{P}_\theta} P^\theta(y_{t+1} \,|\, \mathscr{F}_t) d\Pi(P^\theta \,|\, \mathscr{F}_t) \right\} dy_{t+1} \qquad (2) \\ &= \int_{\mathcal{Y}} V(y_{t+1},a_{t+1},t+1) \overline{P}^\theta(y_{t+1} \,|\, \mathscr{F}_t) dy_{t+1} \end{aligned}$$

where $\overline{P}^\theta(y_{t+1} \,|\, \mathscr{F}_t)$ is the predictive distribution, the posterior weighted average of the conditional distributions of $P^\theta \in \mathcal{P}^\theta$, and $y_{t+1} \in \mathcal{Y}$. The relevance of this observation for the present analysis is that the optimal estimate of the 'parameter' $\overline{P}^\theta$ is independent of the loss function V, a consequence of the presumption that the semiparametric Bayesian does not affect the evolution of the state variables and is interested in the model only for its impact on forecasting.

This is a very large scale forecasting problem if the dimensions of the state and model spaces are large. Small world assumptions can shrink the scale of the problem. One such assumption involves partitioning the state vector into two components $y_t = \{x_t, v_t\}$. In one sense this is just a matter of conditioning since $P^\theta(y_{t+1}|\mathscr{F}_t)$ can always be written as $P^\theta(v_{t+1}|x_{t+1},\mathscr{F}_t)P^\theta(x_{t+1}|\mathscr{F}_t)$. However, analyzing $P^\theta(x_{t+1}|\mathscr{F}_t)$ and $P^\theta(v_{t+1}|x_{t+1},\mathscr{F}_t)$ independently ignores any information in the latter that is relevant for the former. Hence, the first small world assumption is that $P^\theta(x_{t+1}|\mathscr{F}_t)$ and $P^\theta(X^T) = \{x_s, s \leq T\}$ can be safely analyzed in isolation or, in the language of Engle et al. (1983), x_t is weakly exogenous with respect to v_t.

The second such assumption involves the role of conditional probabilities in the structure of the models in $\mathcal{P}^\theta$. Since the archetypical Bayesian has no prior information regarding the data generating process save for that implicit in the moment conditions, it is reasonable to suppose that the semiparametric Bayesian follows a flexible modeling strategy in parameterizing the models in $\mathcal{P}^\theta$. A convenient approach in this setting involves the decomposition of the joint distribution into the product of two components:

$$\mathscr{L}(X^T \mid P^\theta) \equiv P^\theta(X^T) = \Lambda^\theta(X^T)\prod_{t=1}^{T} p^\theta(x_t);\ \Lambda^\theta(X^T) = \frac{P^\theta(X^T)}{\prod_{t=1}^{T} p^\theta(x_t)} \tag{3}$$

where $\Lambda^\theta(X^T)$ is the likelihood ratio statistic for the null hypothesis that the joint distribution $P^\theta(X^T)$ is equal to the product of the marginals $p_\theta(x_t)$ that respect the unconditional moment conditions. The specific series approximation of $\Lambda^\theta(\mathrm{X}^\mathrm{T})$ is not important; what matters is that it is orthogonal to the product of the marginals.[1] If the archetype does not have informative priors over $\Lambda^\theta(\mathrm{X}^\mathrm{T})$, it is feasible to choose a flexible expansion that has the required orthogonality. This condition, coupled with the semiparametric Bayesian's postulated ignorance about the data generating process, suggests that the prior is separable in Λ^θ and p^θ as well.

Letting $\Pi(\Lambda^\theta,\mathrm{p}^\theta) = \Pi(\Lambda^\theta)\ \Pi(\mathrm{p}^\theta)$ denote the separable and proper[2] prior distributions over these possibly infinite-dimensional parameters, Λ^θ can be integrated out of the joint posterior via:

$$\begin{aligned}
\Pi(\mathrm{P}^\theta \mid \mathrm{X}^\mathrm{T}) &= \frac{\Pi(\mathrm{P}^\theta)\mathscr{L}(\mathrm{X}^\mathrm{T}\mid \mathrm{P}^\theta)}{\int_{\mathcal{P}^\theta}\Pi(\mathrm{P}^\theta)\mathscr{L}(\mathrm{X}^\mathrm{T}\mid \mathrm{P}^\theta)d\mathrm{P}^\theta} = \frac{\Pi(\Lambda^\theta)\Lambda^\theta(\mathrm{X}^\mathrm{T})\Pi(\mathrm{p}^\theta)\prod_{\mathrm{t}=1}^{\mathrm{T}}\mathrm{p}^\theta(\mathrm{x}_\mathrm{t})}{\int_{p^\theta}\int_{\Lambda^\theta}\Pi(\Lambda^\theta)\Lambda^\theta(\mathrm{X}^\mathrm{T})\Pi(\mathrm{p}^\theta)\prod_{\mathrm{t}=1}^{\mathrm{T}}\mathrm{p}^\theta(\mathrm{x}_\mathrm{t})d\Lambda^\theta d\mathrm{p}^\theta} \\
\Rightarrow \Pi(\mathrm{p}^\theta \mid \mathrm{X}^\mathrm{T}) &= \frac{\Pi(\mathrm{p}^\theta)\prod_{\mathrm{t}=1}^{\mathrm{T}}\mathrm{p}^\theta(\mathrm{x}_\mathrm{t})\int_{\Lambda^\theta}\Pi(\Lambda^\theta)\Lambda^\theta(\mathrm{X}^\mathrm{T})}{\int_{p^\theta}\int_{\Lambda^\theta}\Pi(\Lambda^\theta)\Lambda^\theta(\mathrm{X}^\mathrm{T})\Pi(\mathrm{p}^\theta)\prod_{\mathrm{t}=1}^{\mathrm{T}}\mathrm{p}^\theta(\mathrm{x}_\mathrm{t})d\Lambda^\theta d\mathrm{p}^\theta} = \frac{\Pi(\mathrm{p}^\theta)\prod_{\mathrm{t}=1}^{\mathrm{T}}\mathrm{p}^\theta(\mathrm{x}_\mathrm{t})}{\int_{p^\theta}\Pi(\mathrm{p}^\theta)\prod_{\mathrm{t}=1}^{\mathrm{T}}\mathrm{p}^\theta(\mathrm{x}_\mathrm{t})d\mathrm{p}^\theta}
\end{aligned} \tag{4}$$

where $\mathcal{P}^\theta = p^\theta \times \Lambda^\theta$. The same reasoning also implies that the predictive distribution for $\mathrm{x}_{\mathrm{t}+1}$ given X^T can be decomposed similarly since:

$$\begin{aligned}
\overline{\mathrm{P}}^\theta(\mathrm{x}_{\mathrm{T}+1}\mid \mathrm{X}^\mathrm{T}) &= \int_{\mathcal{P}^\theta}\mathrm{P}^\theta(\mathrm{x}_{\mathrm{T}+1}\mid \mathrm{X}^\mathrm{T})\Pi(\mathrm{P}^\theta\mid \mathrm{X}^\mathrm{T})d\mathrm{P}^\theta = \frac{\int_{\mathcal{P}^\theta}\mathrm{P}^\theta(\mathrm{x}_{\mathrm{T}+1}\mid \mathrm{X}^\mathrm{T})\mathscr{L}(\mathrm{X}^\mathrm{T}\mid \mathrm{P}^\theta)\Pi(\mathrm{P}^\theta)d\mathrm{P}^\theta}{\int_{\mathcal{P}^\theta}\mathscr{L}(\mathrm{X}^\mathrm{T}\mid \mathrm{P}^\theta)\Pi(\mathrm{P}^\theta)d\mathrm{P}^\theta} \\
&= \frac{\int_{p^\theta}\int_{\Lambda^\theta}\mathrm{p}^\theta(\mathrm{x}_{\mathrm{T}+1})\Pi(\mathrm{p}^\theta)\prod_{\mathrm{t}=1}^{\mathrm{T}}\mathrm{p}^\theta(\mathrm{x}_\mathrm{t})\Pi(\Lambda^\theta)\Lambda^\theta(\mathrm{X}^{\mathrm{T}+1})d\Lambda^\theta d\mathrm{p}^\theta}{\int_{p^\theta}\int_{\Lambda^\theta}\Pi(\Lambda^\theta)\Lambda^\theta(\mathrm{X}^\mathrm{T})\Pi(\mathrm{p}^\theta)\prod_{\mathrm{t}=1}^{\mathrm{T}}\mathrm{p}^\theta(\mathrm{x}_\mathrm{t})d\Lambda^\theta d\mathrm{p}^\theta} \\
&= \frac{\int_{p^\theta}\mathrm{p}^\theta(\mathrm{x}_{\mathrm{T}+1})\Pi(\mathrm{p}^\theta)\prod_{\mathrm{t}=1}^{\mathrm{T}}\mathrm{p}^\theta(\mathrm{x}_\mathrm{t})d\mathrm{p}^\theta\int_{\Lambda^\theta}\Pi(\Lambda^\theta)\Lambda^\theta(\mathrm{X}^{\mathrm{T}+1})d\Lambda^\theta}{\int_{p^\theta}\Pi(\mathrm{p}^\theta)\prod_{\mathrm{t}=1}^{\mathrm{T}}\mathrm{p}^\theta(\mathrm{x}_\mathrm{t})d\mathrm{p}^\theta\int_{\Lambda^\theta}\Pi(\Lambda^\theta)\Lambda^\theta(\mathrm{X}^{\mathrm{T}+1})d\Lambda^\theta} \\
&\equiv \overline{\mathrm{p}}^\theta(\mathrm{x}_{\mathrm{T}+1}\mid \mathrm{X}^\mathrm{T})\overline{\lambda}^\theta(\mathrm{x}_{\mathrm{T}+1}\mid \mathrm{X}^\mathrm{T})
\end{aligned} \tag{5}$$

[1] Note that, from a frequentist perspective, efficient semiparametric models orthogonalize the parametric and nonparametric components.

[2] The requirement that the prior be proper – that is, that it integrates to one – serves to insure that integrals over the space of probability measures are finite.

so that these assumptions imply that $\overline{p}^{\theta}(x_{T+1} \mid X^{T})$ can be safely analyzed in isolation as well.

There are two kinds of small worlds assumptions at work here that place restrictions on the structure of the beliefs of this subjective expected utility maximizer. The first is made implicitly in all applied research: that little harm is done by ignoring any information loss associated with confining the analysis to a subset the economically relevant variables x_t. The second follows from the presumption that the hypothetical semiparametric Bayesian has weak prior beliefs about the data generating process save for the information in moment conditions, which makes it reasonable to treat the likelihood ratio Λ^{θ} as a (possibly infinite-dimensional) nuisance parameter that is independent of the marginal distributions p^{θ} both *a priori* and *a posteriori*. As is readily apparent, these assumptions can be replaced by two simpler ones that are reasonable in many circumstances: that x_t represents all of the information relevant to the decision at hand and is independently and identically distributed or that x_t has two components – i.e., $x_t = \{x_{1t}, x_{2t}\}$ – which are jointly iid and the moment conditions involve the conditional mean $E[x_{1t}|x_{2t}]$.

Collectively, these assumptions simplify the representation of the semiparametric Bayesian's preferences over future outcomes (2), which is now given by:

$$\begin{aligned} E[V(y_{T+1}, a_{T+1}, T+1) \mid \mathscr{F}_T] &= \int_{\mathcal{V}}\int_{\mathcal{X}} V(v_{T+1}, x_{T+1}, a_{T+1}, T+1)\overline{P}^{\theta}(x_{T+1}, v_{T+1} \mid \mathscr{F}_T)dx_{T+1}dv_{T+1} \\ &= \int_{\mathcal{V}}\int_{\mathcal{X}} V(\bullet)\overline{P}^{\theta}(v_{T+1} \mid x_{T+1}, \mathscr{F}_T)\overline{P}^{\theta}(x_{T+1}, \mid \mathscr{F}_T)dx_{T+1}dv_{T+1} \\ &= \int_{\mathcal{X}} E_{\mathcal{V}}[V(\bullet)]\overline{p}^{\theta}(x_{T+1} \mid X^{T})\overline{\lambda}^{\theta}(x_{T+1} \mid X^{T})dx_{T+1} \end{aligned} \tag{6}$$

where $\mathcal{X}$ and $\mathcal{V}$ are the sample spaces of x_{T+1} and v_{T+1}, respectively. There is no need for a loss function because the archetype is an expected utility maximizer. There is no interaction between the learning and decision problems because the archetype's actions do not affect the state variables x_{T+1} and v_{T+1}. The problem of predicting v_{T+1} can be separated from the problem of predicting x_{T+1} because x_{T+1} is weakly exogenous. The presumption that the semiparametric Bayesian has prior information in the form of moment conditions that constrain $\overline{p}^{\theta}(x_{T+1} | X^{T})$ and none that constrains $\overline{\lambda}^{\theta}(x_{T+1} | X^{T})$ permits the archetype to integrate out the latter. In these circumstances, the Bayesian archetype can safely learn about the constrained stationary distribution $\overline{p}^{\theta}(x_{T+1} | X^{T})$ from a sample X^{T} without considering the larger decision problem.

B. Probabilities

Accordingly, let p^{θ} denote the set of all probability measures that satisfy the moment conditions for each θ where x_t takes values on a sample space $\mathcal{X} \subseteq \mathbb{R}^{d}$ and $(\mathcal{X}, \mathcal{F}, p^{\theta})$ is a probability space $\forall\ p^{\theta} \in p^{\theta}$. Note that conditioning information impinges on these measures to the extent that it is embedded in the moment conditions; for example, the moment conditions could be the scores from a parametric likelihood. As is commonplace, assume the q-dimensional parameter θ lies in the interior of the compact set $\Theta \subset \mathbb{R}^{q}$ and that $g(x,\theta)$ is continuously differentiable in an open neighborhood of the true value θ_0 with a first derivative that has full column rank.

The problem of characterizing the model space and formulating priors over it is still quite daunting. One way to simplify the problem is to approximate the stationary measures that comprise the model space with a smaller set of finite dimensional distributions. That is, one can seek a sequence of approximations $\overline{p}^{\theta}_{n}(x_{T+1} | X^{T})$ converging to $\overline{p}^{\theta}(x_{T+1} | X^{T})$ such that:

$$\max_{a_t} V_n(y_t, a_t, t) \rightarrow V(y_t, a_t, t) \tag{7}$$

which would generally obtain if the approximations converge uniformly, preferences are bounded and sufficiently smooth, and actions lie in a compact set.

Suppose the archetypical Bayesian found that the reduction of the space of possible actions to a discrete set resulted in expected utility “close enough” to that derived from the whole opportunity set. This circumstance arises when a table of actions as a function of a finite number of possible scenarios suffices. Accordingly, suppose this Bayesian discretized the choice set into N_a possible – and possibly countably infinite – actions associated with a corresponding partition of the state space into N_a nonoverlapping subsets $\mathcal{Y}_n$ such that $\mathcal{Y} = \bigcup_{n=1}^{N_a} \mathcal{Y}_n;\ \mathcal{Y}_m \cap \mathcal{Y}_n = 0\ \forall\ m \neq n$ for which:

$$\begin{aligned} V(y_t, a_t, t) &= \max_{a_t \in A(y_t)} U(y_t, a_t, t) + \beta E[V(y_{t+1}, a_{t+1}, t+1) \mid \mathscr{F}_t] \\ &\approx U[y_n, a(y_n), t] + \beta \sum\nolimits_m P[y_{t+1} \in \mathcal{Y}_m \mid \mathscr{F}_t] V(y_m, a(y_m), t+1) \mid \mathscr{F}_t] \end{aligned} \tag{8}$$

where $y_t \in \mathcal{Y}_n$ and $y_m \in \mathcal{Y}_m$ is chosen so as to best approximate expected utility (which is just a normalization). Such a Bayesian would naturally look to discrete measures to provide the required approximations.

Accordingly, consider the following family of discrete measures on the associated partition of $\mathcal{X}$ into N_a nonoverlapping subsets $\mathcal{X} = \bigcup_{n=1}^{N_a} \mathcal{X}_n;\ \mathcal{X}_m \cap \mathcal{X}_n = 0\ \forall\ m \neq n$. The discrete measures are taken to be that subset of p^θ comprised of the N_a-cell multinomial distributions $p^\theta_{M_{N_a}} = \left\{ p_n^\theta = P^\theta(\mathcal{X}_n) \geq 0 :\ \sum\nolimits_{n=1}^{N_a} p_n^\theta = 1;\ \sum\nolimits_{n=1}^{N_a} p_n^\theta g_n^\theta = 0 \right\}$ for which $g_n^\theta = E_{p^\theta}[g(x, \theta) \mid x_n \in \mathcal{X}_n]$.[3] Modulo regularity conditions, $p^\theta_{M_{N_a}}$ will approximate p^θ arbitrarily well as N_a grows without bound since, in measure-theoretic terms, these discrete measures are dense in p^θ. This is a restatement of Chamberlain's (1987) observation that any distribution can be approximated

[3] Note that there is a slight abuse of notation in this definition because $p_n^\theta = p^\theta(x_n)$ refers to each such probability distribution for a given value of θ. It might be interesting to consider more structured partitions of the sample space. For example, the cells could be Voronoi tessellations based on a given set of points $\{x_1, \ldots, x_{N_a}\}$ in which the elements of $\mathcal{X}_n$ are those values of x that are closer to x_n than to any other point x_m in this set. The discussion in Jiménez and Yukich (2002) suggests that the analysis could proceed along these lines.

arbitrarily well by a multinomial distribution.[4]

Moment conditions place considerable structure on the multinomial probabilities compatible with them. In particular, consider the projection of p_n^θ on a constant and g_n^θ:

$$p_n^\theta = a_{N_a}^\theta + g_n^{\theta\prime} b_{N_a}^\theta + \varepsilon_n^{p^\theta} \tag{9}$$

and note that the normal equations imply:

$$\begin{aligned} a_{N_a}^\theta &= \frac{1}{N_a}[1 + \bar{g}_{N_a}^\theta{}' V_{N_a}^{\theta\ -1} \bar{g}_{N_a}^\theta];\ \bar{g}_{N_a}^\theta = \frac{1}{N_a}\sum_{n=1}^{N_a} g_n^\theta \\ b_{N_a}^\theta &= -\frac{1}{N_a} V_{N_a}^{\theta\ -1} \bar{g}_{N_a}^\theta;\ V_{N_a}^\theta = \frac{1}{N_a}\sum_{n=1}^{N_a}[g_n^\theta - \bar{g}_{N_a}^\theta][g_n^\theta - \bar{g}_{N_a}^\theta]' \end{aligned} \tag{10}$$

This projection is not an estimate: the multinomial probabilities satisfy these relations arithmetically for each value of θ irrespective of the validity of the null hypothesis.[5] For each value of θ, p_n^θ has a common component $\frac{1}{N_a} - \frac{1}{N_a}\bar{g}_{N_a}^\theta{}' V_{N_a}^{\theta\ -1}[g_n^\theta - \bar{g}_{N_a}^\theta]$ and a zero mean, probability-specific residual $\varepsilon_n^{p^\theta}$ that is uncorrelated with g_n^θ by construction.

Hence, (9) can be rewritten as:

$$\begin{aligned} p_n^\theta &= \frac{1}{N_a} - \frac{1}{N_a}\bar{g}_{N_a}^\theta{}' V_{N_a}^{\theta\ -1}[g_n^\theta - \bar{g}_{N_a}^\theta] + \varepsilon_n^{p^\theta} \\ \varepsilon_n^{p^\theta} &< 1 + \frac{1}{N_a}\bar{g}_{N_a}^\theta{}' V_{N_a}^{\theta\ -1}[g_n^\theta - \bar{g}_{N_a}^\theta] - \frac{1}{N_a} \\ \varepsilon_n^{p^\theta} &> \frac{1}{N_a}\bar{g}_{N_a}^\theta{}' V_{N_a}^{\theta\ -1}[g_n^\theta - \bar{g}_{N_a}^\theta] - \frac{1}{N_a} \end{aligned} \tag{11}$$

This observation appears to be new.[6] For later reference, note that the sum of squared deviations of the probabilities from their means of $1/N_a$ is simply:

[4] See also Theorem 4.1 of Diaconis and Freedman (1986b) for an application of multinomial approximation in a Bayesian context.

[5] Of course, $\bar{g}_{N_a}^\theta \to 0$ and $V_{N_a}^\theta \to S_{N_a}^\theta = \frac{1}{N_a}\sum_n g_n^\theta g_n^{\theta\prime}$ if the null model is true. Relation (9) holds when the moment conditions are imposed but are false as well, in which case $\bar{g}_{N_a}^\theta \to \bar{g}_{N_a} \neq 0$ while $\sum_n p_n^\theta g_n^\theta = 0$ by construction.

[6] The first two terms comprise what Back and Brown (1993) refer to as implied probabilities, which need not be positive. The constraints on the residuals in (11) insure positivity.

$$\begin{aligned}\sum\nolimits_{n=1}^{N_a}\left(p_n^\theta - \frac{1}{N_a}\right)^2 &= \sum\nolimits_{n=1}^{N_a}\left(\varepsilon_n^{p^\theta} - \frac{1}{N_a}\bar{g}_{N_a}^{\theta}{}' V_{N_a}^{\theta}{}^{-1}[g_n^\theta - \bar{g}_{N_a}^\theta]\right)^2 \\ &= \frac{1}{N_a}\bar{g}_{N_a}^{\theta}{}' V_{N_a}^{\theta}{}^{-1}\bar{g}_{N_a}^\theta + N_a\sigma^2(\varepsilon_n^{p^\theta})\end{aligned} \qquad (12)$$

where $\sigma^2(\varepsilon_n^{p^\theta}) = \frac{1}{N_a}\sum\nolimits_{n=1}^{N_a}\varepsilon_n^{p^\theta\,2}$ and the leading term is the GMM overidentifying restrictions test statistic, a fact that will prove useful in the sequel.

Finally, it is worth briefly contrasting this modeling strategy with the methods of Zellner (1994,1997), Kim (2002), Lazar (2003), and Schennach (2005). Zellner, Lazar, and Schennach maximize pseudo-log-likelihoods of the form $\sum\nolimits_n q_n \ln p_n$ (with $q_n = p_n$ in Zellner and Schennach and $q_n = 1/N_a$ in Lazar) subject to different side constraints, posterior moments of the parameters of interest in Zellner and the moment conditions $\sum\nolimits_n p_n^\theta g_n^\theta = 0$ in Lazar and Schennach. Kim treats the exponential of the GMM overidentifying restrictions test statistic as a likelihood function up to normalization. None of these procedures is a truly Bayesian one in which a posterior is obtained by multiplying a prior by a likelihood function and integrating, if necessary, to make probability statements.

The present approach yields a truly Bayesian procedure in which the moment conditions are substituted into the multinomial probabilities exactly via (11). The resulting multinomial likelihood function is a true likelihood to which the full Bayesian calculus can be applied. The only approximation is that of a density by a multinomial for which there is ample justification.[7] Hence, Bayesian inference based on the N_a-cell multinomials in $p_{M_{N_a}}^\theta$ would appear to be a more internally consistent procedure than those taken in these papers.

[7] There is nothing sacred about multinomial approximation *per se*. For example, mixtures of other exponential distributions would suffice. See, for example, Barron and Sheu (1991). Multinomial approximation makes particular sense in the present context when one views the GMM econometrician as implicitly estimating the moment-constrained distribution of the data as opposed to its density. This view is compatible with Back and Brown (1993).

C. Priors

Perhaps the most nettlesome problem associated with subjective expected utility maximization is the formulation of prior beliefs for the parameters of the decision problem. This problem is further complicated in the present setting by its semiparametric nature: the space of all probability distributions compatible with the moment conditions is a "large" metric space in general and the construction of prior beliefs on such spaces is fraught with hazard.[8] Fortunately, this semiparametric Bayesian is willing to work with the smaller space of countable multinomial distributions.

Prior formulation is easier in this setting because of this Bayesian's choice of probability models. The archetype is content *ex ante* with a constrained countable cell multinomial distribution as the semiparametric model for the data for subject matter reasons relating to the adequacy of approximate decision rules and *ex ante* faith in the null model. As in Chamberlain (1987), this turns a semiparametric problem into a parametric one in this setting – albeit for somewhat different reasons – thus facilitating the analysis of prior and posterior beliefs.

The other reason it is comparatively easy to formulate priors in this case is because the priors are over probability measures, not parameter values. Prior distributions over parameter values typically change with the parameterization, the standard example being that a prior that is noninformative for a standard deviation typically is informative for the corresponding variance and vice versa. Priors over probability measures do not suffer from this problem: reparameterization of a model does not change prior beliefs in this fashion.

These considerations make it natural to approximate the priors over partitions of the space of multinomial distributions. The probabilities of the N_a-cell multinomial lie in the

[8] See Ghosh and Ramamoorthi (2003) and the references cited therein.

standard $\mathrm{N_a}$-simplex $\left\{(\mathrm{p}_1,\ldots \mathrm{p}_{\mathrm{N_a}}) \in [0,1]^{\mathrm{N_a}} : \sum_{\mathrm{n}=1}^{\mathrm{N_a}} \mathrm{p_n} = 1\right\}$. The standard $\mathrm{N_a}$-simplex is compact under the Euclidean metric as is the $\mathrm{N_a}$-simplex bounded by the hyperplanes induced by the residual constraints in (11) and so both simplices have finite subcovers. In addition, the probabilities are of order $\mathrm{N_a^{-1}}$ and this reduces the upper bound on the diameter of the relevant $\mathrm{N_a}$-simplex from √2 to $\mathrm{O_p}(\mathrm{N_a^{-1/2}})$, thus constraining the simplex to be bounded by the orthants comprised of all positive coordinates of spheres of the form $\sum_{\mathrm{n}=1}^{\mathrm{N_a}} \mathrm{p_n^2} = \mathrm{O_p}(\mathrm{N_a^{-1}})$. Put differently, the largest eigenvalue of the information matrix of the $\mathrm{N_a}$-cell multinomial is of order $\mathrm{N_a^{-2}}$.

The minimal cover of the truncated $\mathrm{N_a}$-simplex can be used to approximate prior beliefs.[9] It is given by the smallest set of points $\{p_{\mathrm{i}}^{\theta},\ \mathrm{i} = 0,\ldots,\mathrm{N_a^{\delta}} - 1 < \infty\}$, where $\mathrm{N_a^{\delta}}$ is the covering number, such that the balls $\mathrm{B}(p_{\mathrm{i}}^{\theta},\delta) = \{\mathrm{p}^{\theta} \in p^{\theta} : \left\|\mathrm{p}^{\theta} - p_{\mathrm{i}}^{\theta}\right\| \le \delta\}$ are disjoint (i.e., $\left\|p_{\mathrm{i}}^{\theta} - p_{\mathrm{j}}^{\theta}\right\| \ge 2\delta\ \forall\ \mathrm{i} \ne \mathrm{j}$) and cover $p_{\mathrm{M_{N_a}}}^{\theta}$. By convention, the minimal cover is normalized so that $\mathrm{p}^0 \in \mathrm{B}(p_0^{\theta},\delta)$, i.e., the first ball contains the true model.[10] It is natural to approximate prior beliefs by $\pi_{\mathrm{i}}^{\theta}(\delta) = \Pi[\mathrm{B}(p_{\mathrm{i}}^{\theta},\delta)]$, which, of course, need not represent the way in which the underlying prior itself was formulated.[11]

[9] See Diaconis and Freedman (1990) for a detailed discussion of Bayes estimates for the finite dimensional multinomial distribution in finite samples without the fiction of a 'true' model.

[10] There are at least two internally consistent interpretations of this true model. The first views this modeling exercise as being conditional on p^0 being the truth 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. Alternatively, the so-called true model can be replaced by the one that minimizes the Kullback-Leibler divergence between it and the truth. In these circumstances, distributions constructed to satisfy the moment conditions are perfectly well-posed but one would not expect $\overline{\mathrm{g}}_{\mathrm{T}}(\theta) \to 0$ even at the pseudo-true value θ_0 and so one would expect $\sup_{\mathrm{t}}\{\mathrm{T}\varepsilon_{\mathrm{t}}(\mathrm{p}^{\theta}) - \overline{\mathrm{g}}_{\mathrm{T}}(\theta)'\mathrm{V_T}(\theta)^{-1}[\mathrm{g_t}(\theta) - \overline{\mathrm{g}}_{\mathrm{T}}(\theta)]\} = \mathrm{O_p}(1)$, not $\mathrm{o_p}(1)$ as would be the case if the moment conditions were true. In addition, sufficiently false models would typically have probabilities that failed to be of order T^{-1}.

[11] Priors formulated in this way trouble some Bayesians when the prior depends on the sample size, as would be the case if the covering number $\mathrm{N_a^{\delta}}$ is sample size dependent. See, for example, Heath and Sudderth (1978). This basic strategy can be used to construct coherent non-informative priors by making the required accuracy of the

D. Prediction

The preceding three subsections provided a character sketch of a semiparametric Bayesian whose inferences the econometrician seeks to infer. This Bayesian is concerned with forecasting and this focus has a perhaps surprising implication in this semiparametric setting: the predictive distribution converges to the true distribution without additional regularity conditions. The twin discretizations – that is, the reduction of the space of measures that respect the moment conditions to a countable set of multinomial distributions that do so – make for predictive distributions with statistically distinguishable components.

The discrete approximation to the semiparametric Bayesian's predictive distribution based on $\pi_{\mathrm{i}}^{\theta}(\delta)$ is given by:

$$\overline{\mathrm{p}}^{\theta} = \sum\nolimits_{\mathrm{i}=0}^{\mathrm{N}_{\mathrm{a}}^{\delta}-1} \Pi(P_{\mathrm{i}}^{\theta} \,|\, \mathrm{X}^{\mathrm{N}_{\mathrm{a}}^{\delta}}) p_{\mathrm{i}}^{\theta} = \frac{\sum\nolimits_{\mathrm{i}=0}^{\mathrm{N}_{\mathrm{a}}^{\delta}-1} \pi_{\mathrm{i}}^{\theta}(\delta) P_{\mathrm{i}}^{\theta}(\mathrm{X}^{\mathrm{N}_{\mathrm{a}}^{\delta}}) p_{\mathrm{i}}^{\theta}}{\sum\nolimits_{\mathrm{i}=0}^{\mathrm{N}_{\mathrm{a}}^{\delta}-1} \pi_{\mathrm{i}}^{\theta}(\delta) P_{\mathrm{i}}^{\theta}(\mathrm{X}^{\mathrm{N}_{\mathrm{a}}^{\delta}})} \tag{13}$$

where $\mathrm{X}^{\mathrm{N}_{\mathrm{a}}^{\delta}}$ is a sample of size $\mathrm{N}_{\mathrm{a}}^{\delta}$ and $P_{\mathrm{i}}^{\theta}(\mathrm{X}^{\mathrm{N}_{\mathrm{a}}^{\delta}}) = \prod\nolimits_{\mathrm{n}=1}^{\mathrm{N}_{\mathrm{a}}^{\delta}} p_{\mathrm{in}}^{\theta}$ is the likelihood of the model around which $\mathrm{B}(p_{\mathrm{i}}^{\theta},\delta)$ is centered. Given the positivity requirement placed on the priors, the limiting properties of $\overline{\mathrm{p}}^{\theta}$ depend only on the large sample behavior of these likelihoods. Their distinguishability means that their large sample limits can be analyzed in isolation.

To recapitulate the assumptions that are scattered across the preceding three subsections, suppose x_{t} takes values on a sample space $\mathcal{X} \subseteq \mathbb{R}^{\mathrm{d}}$ and $(\mathcal{X},\mathscr{F},\mathrm{p}^{\theta})$ is a probability space $\forall\ \mathrm{p}^{\theta} \in p^{\theta}$. Each p^{θ} satisfies $\mathrm{E}_{\mathrm{p}^{\theta}}[\mathrm{g}(\mathrm{x},\theta)] = 0\ \forall\ \theta \in \Theta \subset \mathbb{R}^{\mathrm{q}}$ with $\mathrm{g}(\mathrm{x},\theta)$ continuously differentiable in an

multinomial approximation a parameter of the decision problem and then placing a noninformative prior over the number of cells needed to achieve this degree of accuracy. A variant of this approach may be found in Ghosh and Ramamoorthi (2003). The idea of defining uniform probabilities over topological objects like balls of the same size seems to have originated in Dembski (1992). In a strange evolution of ideas, he has since managed to use this idea to somehow argue for "intelligent design" in the creationism debate.

open neighborhood of θ_0 with a first derivative that has full column rank. The true measure p^0 has finite entropy (i.e., $\mathrm{E}_{\mathrm{p}^0}[\ln \mathrm{p}^0] > -\infty$). Partition $\mathcal{X}$ into $\mathcal{X} = \bigcup_{\mathrm{n}=1}^{\mathrm{N_a}} \mathcal{X}_{\mathrm{n}};\ \mathcal{X}_{\mathrm{m}} \cap \mathcal{X}_{\mathrm{n}} = 0\ \forall\ \mathrm{m} \neq \mathrm{n}$ and let $p^\theta_{\mathrm{M}_{\mathrm{N_a}}} = \left\{ \mathrm{p}^\theta_{\mathrm{n}} = \mathrm{P}^\theta(\mathcal{X}_{\mathrm{n}}) \geq 0 :\ \sum_{\mathrm{n}=1}^{\mathrm{N_a}} \mathrm{p}^\theta_{\mathrm{n}} = 1;\ \sum_{\mathrm{n}=1}^{\mathrm{N_a}} \mathrm{p}^\theta_{\mathrm{n}} \mathrm{g}^\theta_{\mathrm{n}} = 0 \right\}$ be the subset of $\mathrm{N_a}$-cell multinomial distributions on this partition for which $\mathrm{g}^\theta_{\mathrm{n}} = \mathrm{E}_{\mathrm{p}^\theta}[\mathrm{g}(\mathrm{x},\theta) \,|\, \mathrm{x}_{\mathrm{n}} \in \mathcal{X}_{\mathrm{n}}]$. Finally, let $\left\{ p^\theta_{\mathrm{i}} :\ \left\| p^\theta_{\mathrm{i}} - p^\theta_{\mathrm{j}} \right\| \geq 2\delta\ \forall\ \mathrm{i} \neq \mathrm{j},\ \mathrm{i} = 0,\ldots,\mathrm{N}^\delta_{\mathrm{a}} - 1 < \infty \right\}$ be the minimal cover of $p^\theta_{\mathrm{M}_{\mathrm{N_a}}}$ by the $\mathrm{N}^\delta_{\mathrm{a}}$ disjoint balls $\mathrm{B}(p^\theta_{\mathrm{i}},\delta) = \left\{ \mathrm{p}^\theta \in p^\theta : \left\| \mathrm{p}^\theta - p^\theta_{\mathrm{i}} \right\| \leq \delta \right\}$ with $\mathrm{p}^0 \in \mathrm{B}(p^\theta_0,\delta)$. Note that $\|\bullet\|$ denotes the Euclidean metric and $\rightarrow$ denotes almost sure convergence when applied to a random variable.

In these circumstances, we have:

Theorem 1: Let Π be a prior distribution on p^θ. If $\pi^\theta_{\mathrm{i}}(\delta) = \Pi[\mathrm{B}(p^\theta_{\mathrm{i}},\delta)] > 0\ \forall\ \delta > 0$, the predictive distribution (13) is consistent as $\mathrm{N_a} \rightarrow \infty$ and $\delta \rightarrow 0$ p^0 almost surely.

Proof: Divide the sample likelihoods in the numerator and the denominator of (13) by the true distribution so that:

$$
\begin{aligned}
\left\| \bar{\mathrm{p}}^\theta - \mathrm{p}^0 \right\| &= \left\| \sum_{\mathrm{i}=0}^{\mathrm{N}^\delta_{\mathrm{a}}-1} \pi^\theta(p^\theta_{\mathrm{i}} \,|\, \mathrm{X}^{\mathrm{N}^\delta_{\mathrm{a}}})(p^\theta_{\mathrm{i}} - \mathrm{p}^0) \right\| = \left\| \frac{\sum_{\mathrm{i}=0}^{\mathrm{N}^\delta_{\mathrm{a}}-1} \pi^\theta_{\mathrm{i}}(\delta)(p^\theta_{\mathrm{i}} - \mathrm{p}^0) \prod_{\mathrm{n}=1}^{\mathrm{N}^\delta_{\mathrm{a}}-1} p^\theta_{\mathrm{in}}}{\sum_{\mathrm{i}=0}^{\mathrm{N}^\delta_{\mathrm{a}}-1} \pi^\theta_{\mathrm{i}}(\delta) \prod_{\mathrm{n}=1}^{\mathrm{N}^\delta_{\mathrm{a}}-1} p^\theta_{\mathrm{in}}} \right\| \\
&= \left\| \frac{\sum_{\mathrm{i}=0}^{\mathrm{N}^\delta_{\mathrm{a}}-1} \pi^\theta_{\mathrm{i}}(\delta)(p^\theta_{\mathrm{i}} - \mathrm{p}^0)\ell^\theta_{\mathrm{iN}^\delta_{\mathrm{a}}}}{\sum_{\mathrm{i}=0}^{\mathrm{N}^\delta_{\mathrm{a}}-1} \pi^\theta_{\mathrm{i}}(\delta)\ell^\theta_{\mathrm{iN}^\delta_{\mathrm{a}}}} \right\| \\
&= \left\| \frac{\pi^\theta_0(\delta)(p^\theta_0 - \mathrm{p}^0)\ell^\theta_{0\mathrm{N}^\delta_{\mathrm{a}}} + \sum_{\mathrm{i}=1}^{\mathrm{N}^\delta_{\mathrm{a}}-1} \pi^\theta_{\mathrm{i}}(\delta)(p^\theta_{\mathrm{i}} - \mathrm{p}^0)\ell^\theta_{\mathrm{iN}^\delta_{\mathrm{a}}}}{\sum_{\mathrm{i}=0}^{\mathrm{N}^\delta_{\mathrm{a}}-1} \pi^\theta_{\mathrm{i}}(\delta)\ell^\theta_{\mathrm{iN}^\delta_{\mathrm{a}}}} \right\| \\
&\leq \frac{\pi^\theta_0(\delta)\ell^\theta_{0\mathrm{N}^\delta_{\mathrm{a}}} \left\| p^\theta_0 - \mathrm{p}^0 \right\|}{\sum_{\mathrm{i}=0}^{\mathrm{N}^\delta_{\mathrm{a}}-1} \pi^\theta_{\mathrm{i}}(\delta)\ell^\theta_{\mathrm{iN}^\delta_{\mathrm{a}}}} + \frac{\sum_{\mathrm{i}=1}^{\mathrm{N}^\delta_{\mathrm{a}}-1} \pi^\theta_{\mathrm{i}}(\delta)\ell^\theta_{\mathrm{iN}^\delta_{\mathrm{a}}} \left\| p^\theta_{\mathrm{i}} - \mathrm{p}^0 \right\|}{\sum_{\mathrm{i}=0}^{\mathrm{N}^\delta_{\mathrm{a}}-1} \pi^\theta_{\mathrm{i}}(\delta)\ell^\theta_{\mathrm{iN}^\delta_{\mathrm{a}}}}
\end{aligned}
\tag{14}
$$

where each $\ell^\theta_{\mathrm{iN}^\delta_{\mathrm{a}}}$ is the sample likelihood ratio statistic for the simple null hypothesis that the constrained stationary distribution is p^{θ}_{i} against the simple alternative hypothesis that it is p^0. Clearly, the denominator is bounded below by $\pi^{\theta}_{0}(\delta) > 0$. By Stein's lemma [Chernoff (1952)], the type 2 error probability for $\ell^{\theta}_{\mathrm{iN}^{\delta}_{\mathrm{a}}}$ is given by $\exp\{-\mathrm{N}_{\mathrm{a}}\mathrm{E}_{\mathrm{p}^0}[\ln(p^{\theta}_{\mathrm{i}}) - \ln(\mathrm{p}^0)] + \mathrm{o}(\mathrm{N}_{\mathrm{a}})\}$ for any Type 1 error probability $0 \le \alpha < 1$, where $\mathrm{E}_{\mathrm{p}^0}[\ln(p^{\theta}_{\mathrm{i}}) - \ln(\mathrm{p}^0)]$ is the Kullback-Leibler divergence of the i^{th} model. Hence, each $\{\psi^{\ell}_{\mathrm{iT}},\ \mathrm{i} > 0\}$ converges to zero at an exponential rate and thus the predictive distribution will lie asymptotically within $\mathrm{B}(p^{\theta}_{0}, \delta)$ at a rate given by $\min_{\mathrm{i}>0} \mathrm{E}_{\mathrm{p}^0}[\ln(p^{\theta}_{\mathrm{i}}) - \ln(\mathrm{p}^0)]$. Now let $\delta \to 0$ and convergence obtains. □

The regularity conditions underlying Theorem 1 yield consistency of the predictive distribution but are not sufficient to deliver posterior consistency: that is, $\pi^{\theta}(\mathrm{B}(p^{\theta}_{0}, \delta) \,|\, \mathrm{X}^{\mathrm{N}^{\delta}_{\mathrm{a}}}) \to 1$ and $\pi^{\theta}(\mathrm{B}(p^{\theta}_{\mathrm{i}}, \delta) \,|\, \mathrm{X}^{\mathrm{N}^{\delta}_{\mathrm{a}}}) \to 0\ \forall\ \mathrm{i} > 0$. Posterior consistency implicitly involves pairwise comparisons of the posterior probabilities of each of the distributions under consideration, not the pairwise comparison of each such distribution against p^0. Sampling variation in the *relative* likelihood ratios can impede the process of posterior convergence if the ratio of prior probabilities strays too far from unity.

Walker (2004) provides the currently weakest sufficient conditions for convergence in the more general setting in which p^{θ}_{i} is a countable set of densities: $\Pi[\mathrm{KL}_{\delta}] > 0\ \forall\ \delta > 0$ where $\mathrm{KL}_{\delta} = \{\mathrm{p} : \mathrm{E}_{\mathrm{p}^0}[\ln(\mathrm{p}) - \ln(\mathrm{p}^0)] < \delta\}$ and $\sum_i \sqrt{\Pi(p^{\theta}_{\mathrm{i}})} < \infty$.[12] The first condition, due to Schwartz (1965), is the analog of the positivity of the prior over the balls $\mathrm{B}(p^{\theta}_{\mathrm{i}}, \delta)$, the passage from Euclidean to Kullback-Leibler balls reflecting the transition from countable multinomials to

[12] See also Section 6 of the much-cited unpublished technical report of Barron (1988). Note that the summability condition has to be strengthened when p^{θ} is an uncountable space of densities.

densities. The second condition insures that distributions that (randomly) overfit the data are not given too much weight, which prevents them from interfering with posterior convergence.[13] In fact, averaging can even permit the predictive distribution to converge without the posterior doing so. In other words, the conditions on the prior that ensure posterior convergence are more delicate than those that deliver convergence of the predictive distribution.[14]

The consistency of the predictive distribution under these comparatively simple regularity conditions stands in sharp contrast to the circumstances in which posterior convergence fails documented in Freedman (1963, 1965), Freedman and Diaconis (1983, 1986a,b), and Stinchcombe (2004). Broadly speaking, two features of the priors they consider cause such failures. The first is the absence of a restriction like $\Pi[\mathrm{B}(p_{\mathrm{i}}^{\theta},\delta)] > 0$ or $\Pi[\mathrm{KL}_{\delta}] > 0\ \forall\ \delta > 0$, which results in positive probability being placed on meager sets of probability measures and, hence, in the possibility of convergence to distributions that track the data – especially large realizations – too well. The second reflects the fact that posterior convergence is more problematic than convergence of the predictive distribution, particularly when the objects of interest are metric spaces of densities. Ghosh and Ramamoorthi (2003) and Walker (2004) discuss priors that deliver posterior convergence in these circumstances.

Is there an economic motivation for confining attention to those beliefs that converge uniformly to p^0? A minimal condition for an inductive learning scheme to be deemed rational would appear to be that it produces an estimate of the marginal distribution that respects the moment conditions and that converges to its population analogue when it is feasible to do so. In the present setting, this amounts to assuming that a Bayesian learning from data would go through these "what if" calculations and avoid priors that did not satisfy $\pi_{\mathrm{i}}^{\theta}(\delta) = \Pi[\mathrm{B}(p_{\mathrm{i}}^{\theta},\delta)] > 0$

[13] See Walker et al. (2004) for a detailed investigation of this phenomenon.
[14] See Barron (1999) for more on these points.

$\forall\ \delta > 0$. On the other hand, the set of priors that satisfy this restriction is topologically small and the topologically large set of priors excluded by this criterion imply that "for essentially any pair of Bayesians, each thinks the other is crazy" (Freedman (1965), p. 455) and that they engage in "erratic, wildly inconsistent, fickle, or faddish" behavior (Stinchcombe (2004), p. 17). The question that remains is whether one should think of such behavior as economically relevant.

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

The preceding section progressively constructed the beliefs of an archetypical semiparametric Bayesian. It is time to relate these beliefs to the inferences of the econometrician studying the class of models defined by the unconditional moment conditions in large samples. The next subsection relates the analytical framework of the previous section to the estimation setting in which the econometrician resides. The penultimate subsection provides the distorted beliefs interpretation of inference in this setting. The final subsection discusses ways in which the decomposition (11) can be used to identify potentially plausible distorted beliefs.

A. Estimation

An econometrician interested in testing such a model would collect a sample of T observations $X^T = \{x_1\ x_2\ \ldots\ x_{T-1}\ x_T\}$. Since the Bayesian did not know *a priori* how the sample space would be carved up by nature, care must be taken to make sure that the econometrician looks at the sample space in a way that is compatible with the perspective of the semiparametric Bayesian. From the perspective of both, nature partitioned the sample space $\mathcal{X}$ into T nonoverlapping subsets $\mathcal{X}_t$ as in $\mathcal{X} = \bigcup_{t=1}^{T} \mathcal{X}_t;\ \mathcal{X}_s \cap \mathcal{X}_t = 0\ \forall\ s \neq t$ with $x_t \in \mathcal{X}_t$ for each t. Hence, the Bayesian's partition $\mathcal{X} = \bigcup_{n=1}^{N_a} \mathcal{X}_n;\ \mathcal{X}_m \cap \mathcal{X}_n = 0\ \forall\ m \neq n$ maps directly into that of the econometrician when $N_a \approx T$ and T is large, which insures the mapping is one-to-one.[15]

The approximate likelihood functions used by the semiparametric Bayesian must also be related to the estimation framework employed by the econometrician, which involves the replacement of conventional GMM estimation with that based on the countable cell multinomial. In particular, the log likelihood function for the T-cell multinomial is proportional to:

$$\begin{aligned}\mathcal{L}(p,\theta\,|\,X^T) \propto \frac{1}{T}\sum\nolimits_t \ln p_t^\theta;\ 1 > p_t^\theta = \frac{1}{T} - \frac{1}{T}\bar{g}_T^{\theta\prime} V_T^{\theta-1}[g_t^\theta - \bar{g}_T^\theta] + \varepsilon_t^{p^\theta} > 0 \\ \bar{g}_T^\theta = \frac{1}{T}\sum_{t=1}^{T} g_t^\theta;\ V_T^\theta = \frac{1}{T}\sum_{T=1}^{T}[g_t^\theta - \bar{g}_T^\theta][g_t^\theta - \bar{g}_T^\theta]'\end{aligned} \tag{15}$$

where the variables are as before save for being defined over the sample partition.

This is the countable cell multinomial studied by Rao (1958) that was reborn as empirical likelihood in Owen (1988, 2001) and extended to the GMM setting in Qin and Lawless (1994), although Rao allowed for sample proportions different from 1/T and for arbitrary smooth functions p_t^θ as opposed to the moment conditions. His proof of the consistency of the maximum likelihood estimator is instructive in its simplicity:

Theorem 2 [Rao (1958)]: If $E_{p^0}[\ln p^0] > -\infty$, the maximum likelihood estimators $\hat{\theta}$ and $\{\hat{p}_t^{\hat{\theta}},\ t = 1,\dots T\}$ converge to θ_0 and $\{p_t^0, t = 1,\dots,T\}$ almost surely p^0.

Proof: The sample entropies are ordered so that:

$$\frac{1}{T}\sum\nolimits_t \ln p_t^0 \le \frac{1}{T}\sum\nolimits_t \ln \hat{p}_t^{\hat{\theta}} \le \frac{1}{T}\sum\nolimits_t \ln\frac{1}{T} = -\ln T \tag{16}$$

and:

$$\frac{1}{T}\sum\nolimits_t \ln p_t^0 \to \sum\nolimits_t p_t^0 \ln p_t^0 \tag{17}$$

[15] For small values of T, there is a small problem when $N_a > T$: sampling theorist would aggregate empty cells while a Bayesian would weight them by their prior probabilities.

by the strong law of large numbers. The twin limiting conditions:

$$\begin{aligned}&\lim_{T\to\infty}\sup\frac{1}{T}\sum_t \ln\frac{1}{T}\le\sum_t p_t^0\ln p_t^0\\&\lim_{T\to\infty}\inf\frac{1}{T}\sum_t \ln\frac{1}{T}\ge\frac{1}{T}\sum_t \ln p_t^0=\sum_t p_t^0\ln p_t^0\end{aligned}\tag{18}$$

imply that:

$$\frac{1}{T}\sum_t \ln\frac{1}{T}\to\sum_t p_t^0\ln p_t^0\tag{19}$$

which, in conjunction with (17), implies that the three sums in (16) converge to the same limit. Since $\sum_t p_t^0\ln p_t^0$ is finite by assumption:

$$\begin{aligned}&\frac{1}{T}\sum_t \ln\frac{\hat{p}_t^{\hat{\theta}}}{1/T}\to 0\\&\frac{1}{T}\sum_t [\hat{p}_t^{\hat{\theta}}-\frac{1}{T}]^2\to 0\end{aligned}\tag{20}$$

as well, this last due to the fact that $\ln\frac{a}{b}\ge\frac{1}{2}(a-b)^2$. Each term in the sum is positive and so:

$$\frac{1}{T}[\hat{p}_t^{\hat{\theta}}-\frac{1}{T}]^2\to 0\Rightarrow\hat{p}_t^{\hat{\theta}}-p_t^0\to 0\tag{21}$$

which, in turn, implies convergence of the whole distribution:

$$\sum_t|\hat{p}_t^{\hat{\theta}}-p_t^0|\to 0\tag{22}$$

via Scheffés theorem. □

An immediate corollary of Theorems 1 and 2 is:

Corollary 1: Under the conditions of Theorem 1, $\{\hat{p}_t^{\hat{\theta}},\ t=1,\ldots T\}$ converges to the predictive distribution (13).

Finally, it is worth considering related estimators that are first order efficient. A convenient class in this setting arises from the family of *φ*- or f-divergences introduced by

Csiszár (1967). 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 with $\varphi(1) = \varphi'(1) = 0$. The term discrepancy serves as a reminder that $\varphi(\bullet)$ need not possess either the symmetry or triangle inequality properties of a metric. The smoothness assumption rules out weak metrics[16] such as the Kolmogorov and Prohorov but contains all of the Cressie-Read (1984,1988) power divergence family for which $\varphi(z)$ is linear in $z^{\alpha}/\alpha(\alpha-1)$ including the likelihood divergence, entropy or Kullback-Leibler information, the Hellinger metric, and Pearson's and Neyman's modified χ^2.

The divergence between p and q is measured by $D_{\varphi}(z) = E_q[\varphi(z)]$.[17] A fact that will be useful in the sequel concerns the behavior of $D_{\varphi}(z)$ when p and q are "close." A Taylor series expansion of $D_{\varphi}(z)$ for two discrete measures with probabilities p_t and q_t for t= 1,…,T yields:

$$\begin{aligned} D_{\varphi}(z) &= \sum\nolimits_t q_t \varphi(z_t) = \sum\nolimits_n q_t[\varphi(1) + \varphi'(1)(z_t - 1) + \frac{1}{2}\varphi''(\xi_t)(z_t - 1)^2];\ \xi_t \le z_t \\ &= \frac{1}{2}\sum\nolimits_t q_t\{\varphi''(1) + [\varphi''(\xi_t) - \varphi''(1)]\}(z_t - 1)^2 \\ &= \frac{1}{2}\sum\nolimits_t q_t \varphi''(1)(z_t - 1)^2 + \frac{1}{2}\sum\nolimits_t q_t[\varphi''(\xi_t) - \varphi''(1)](z_t - 1)^2 \\ &\le \frac{\varphi''(1)}{2}\sum\nolimits_t q_t(z_t - 1)^2 + \frac{1}{2}\sup\nolimits_t\{[\varphi''(\xi_t) - \varphi''(1)](z_t - 1)^2\} \end{aligned} \qquad (23)$$

where the leading term is proportional to Neyman's modified χ^2 divergence. Hence:

$$D_{\varphi}(z) \to \frac{\varphi''(1)}{2}\sum\nolimits_t q_t(z_t - 1)^2 \qquad (24)$$

uniformly if $\sup_t|z_t - 1| = o(1)$ and if $\varphi(z)$ has bounded second derivatives in the neighborhood of one.[18]

[16] See Donoho and Liu (1988) for a discussion of how such metrics can produce poorly behaved minimum distance estimates.

[17] The Csiszár divergence is sometimes defined to be the p expectation $D_{\varphi}(z) = E_p[\varphi(z)]$.

[18] $D_{\varphi}(z)$ also converges to $E_p[\varphi(z)]$. Let $z_{[t]}$ be the order statistics of z_t – that is, $z_{[1]} \le z_{[2]} \le \ldots \le z_{[T-1]} \le z_{[T]}$ – and let $\{[z_{[t-1]}, z_{[t]}], \xi_{[t]}\}$ be the associated tagged partitions with the tags $\xi_{[t]}$ given by

Now consider the divergence between the multinomial probabilities p_t^θ and the associated empirical probabilities $P^\theta(\mathcal{X}_t) = \frac{1}{T}$. Since it ignores the information contained in the moment conditions, $P^\theta(\mathcal{X}_t) = \frac{1}{T}$ is consistent and inefficient under the null but, unlike those that impose the restrictions implied by the moment conditions, is consistent under the alternative as well.[19] Setting $p_t = p_t^\theta$ and $q_t = \frac{1}{T}$ in (24), $D_\varphi(\frac{p_t^\theta}{1/T})$ converges to a variant of (12):

$$\begin{aligned} D_\varphi(\tfrac{p_t^\theta}{1/T}) &\to \frac{\varphi''(1)}{2}\sum\nolimits_t \frac{1}{T}\{T\varepsilon_t^{p^\theta} - \bar{g}_T^{\theta\prime} V_T^{\theta-1}[g_t^\theta - \bar{g}_T^\theta]\}^2 \\ &= \frac{\varphi''(1)}{2}[\bar{g}_T^{\theta\prime} V_T^{\theta-1}\bar{g}_T^\theta + T^2\sigma^2(\varepsilon_t^{p^\theta})] \end{aligned} \tag{25}$$

where $\sigma^2(\varepsilon_t^{p^\theta}) = \frac{1}{T}\sum_t \varepsilon_t^{p^\theta 2}$ and $\mathrm{Cov}[\varepsilon_t^{p^\theta}, g_t^\theta] = \frac{1}{T}\sum_t g_t^\theta \varepsilon_t^{p^\theta} \equiv 0$ by construction.

This representation is useful for two reasons. First, minimization of $D_\varphi(\frac{p_t^\theta}{1/T})$ subject to the positivity and sum constraints on the probabilities provides alternative estimators to the empirical likelihood/infinite cell multinomial estimator of Theorem 2. Second, the quadratic structure of (24) makes it easier to understand the role of the residuals $\varepsilon_t^{p^\theta}$ from (11). These considerations suggest the following theorem:

Theorem 3: Under the conditions of Theorems 1 and 2 and if $\varphi(z)$ has bounded second derivatives in the neighborhood of unity, the estimators $\hat{\theta}_\varphi$ and $\{p_t^{\hat{\theta}_\varphi}, t = 1,\ldots,T\}$ that minimize $D_\varphi(\frac{p_t^\theta}{1/T})$ converge to θ_0 and $\{p_t^0, t = 1,\ldots,T\}$ almost surely p^0 and to the predictive distribution

$\{\xi_{[t]} : \varphi(z_{[t]}) = \varphi(\xi_{[t]})[z_{[t]} - z_{[t-1]}], t = 2,\ldots,T\}$ where the initial tag satisfies $\xi_{[1]} = z_{[1]} - \zeta$ and $z_{[0]} = z_{[1]} - \frac{\varphi(z_{[1]})}{\varphi(\xi_{[1]})}$ with $\xi_{[1]} \downarrow 0$. The discrete sum (24) converges to:

$$D_\varphi(z) = \sum\nolimits_{[t]} q_{[t]}\varphi(z_{[t]}) = \sum\nolimits_{[t]} q_{[t]}\varphi(\xi_{[t]})[z_{[t]} - z_{[t-1]}] \to \int_\Omega q[x(z)]\varphi(z)dz$$

where $x(z)$ is the realization of x_t associated with z_t.

[19] Of course, 1/T would be replaced by k_t/T for any region $\mathcal{X}_t$ that contains more than one realization x_t – a case that naturally arises when its distribution contains atoms – where k_t is the cell count. I will ignore such atoms in what follows.

(13) as well. If, in addition, $\sup_t \bar{g}_T^{\theta_0\prime} V_T^{\theta_0 -1}[g_t^{\theta_0} - \bar{g}_T^{\theta_0}] = o_p(1)$,[20] $\lim_{T\to\infty} \varepsilon_t^{p^{\theta}} = 0 \ \forall \ t$ and $\frac{2T}{\varphi''(1)} D_{\varphi}(\frac{p_t^{\hat{\theta}_\varphi}}{{}^1\!/_T}) \to \chi^2_{p-q}$.

Proof: Consistency follows directly from (20) in the proof of Theorem 2 via the implied limiting equality of the maximum likelihood and minimum divergence estimators. Corollary 1 then applies to these estimators as well. Finally, the upper and lower bound constraints in (11) do not bind asymptotically if $\sup_t \bar{g}_T^{\theta_0\prime} V_T^{\theta_0 -1}[g_t^{\theta_0} - \bar{g}_T^{\theta_0}] = o_p(1)$. Hence, minimization of (25) will be such that $\varepsilon_t^{p^{\theta}} = 0 \ \forall \ t$ in large samples and so $\frac{2T}{\varphi''(1)} D_{\varphi}(\frac{p_t^{\hat{\theta}_\varphi}}{{}^1\!/_T}) = T\bar{g}_T^{\hat{\theta}_\varphi\prime} V_T^{\hat{\theta}_\varphi -1} \bar{g}_T^{\hat{\theta}_\varphi} + o_p(1)$. □

B. Inference

The large sample χ^2 test statistic $\frac{2T}{\varphi''(1)} D_{\varphi}(\frac{p_t^{\hat{\theta}_\varphi}}{{}^1\!/_T})$ obtained by minimizing (25) can be used to test the null hypothesis. Conventional practice is to select a significance level α and an associated critical value c^{α}_{p-q} that solves $\Pr(\chi^2_{p-q} \geq c^{\alpha}_{p-q}) = \alpha$. The null hypothesis is rejected if $D_{\log}(\frac{p_t^{\hat{\theta}}}{{}^1\!/_T}) > c^{\alpha}_{p-q}$ while the statistic fails to reject the null if $\frac{2T}{\varphi''(1)} D_{\varphi}(\frac{p_t^{\hat{\theta}_\varphi}}{{}^1\!/_T}) \leq c^{\alpha}_{p-q}$. As is typically the case in likelihood-based inference, the rejection region can be viewed as the complement of the 1–α per cent confidence region given by $\left\{ p_t^{\theta} : \frac{2T}{\varphi''(1)} D_{\varphi}(\frac{p_t^{\theta}}{{}^1\!/_T}) \leq c^{\alpha}_{p-q} \right\}$.

Theorem 1 and Corollary 1 provide for an economic interpretation of rejections in this

[20] This restriction will be satisfied in most circumstances since $g_t^{\theta_0}$ is naturally $O_p(1)$ and $\bar{g}_T^{\theta_0}$ converges to zero at rate $\sqrt{T}$. Consistency only requires $\sup_t \bar{g}_T^{\theta_0\prime} V_T^{\theta_0 -1}[g_t^{\theta_0} - \bar{g}_T^{\theta_0}] + \varepsilon_t^{p^{\theta_0}} = o_p(1)$ but it is not obvious to me what manner of stochastic process would violate $\sup_t \bar{g}_T^{\theta_0\prime} V_T^{\theta_0 -1}[g_t^{\theta_0} - \bar{g}_T^{\theta_0}] = o_p(1)$ without interfering with consistency or the requirement that $\mathrm{Cov}[\varepsilon_t^{p^{\theta_0}}, g_t^{\theta_0}] = 0$.

inference framework. The rejection region $\left\{ \mathrm{p}^{\theta} \in p^{\theta}_{\mathrm{M_T}} : \frac{2\mathrm{T}}{\varphi''(1)} \mathrm{D}_{\varphi}\left(\frac{\mathrm{p}_{\mathrm{t}}^{\hat{\theta}_{\varphi}}}{1/\mathrm{T}}\right) > \mathrm{c}^{\alpha}_{\mathrm{p-q}} \right\}$ is a subset of the T-cell multinomials in $p^{\theta}_{\mathrm{M_T}}$. The question at hand is simple: are there beliefs implicit in the rejection region that the econometrician would think that the archetypical semiparametric Bayesian 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–α 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 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 $\frac{1}{\mathrm{T}}\sum_{\mathrm{t}} \ln \hat{\mathrm{p}}_{\mathrm{t}}^{\hat{\theta}_{\varphi}} - \frac{1}{\mathrm{T}} \ln \frac{1}{\mathrm{T}}$ based on the estimate $\hat{\theta}_{\varphi}$ is "unreasonably low" 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 ½ and the sample success probability is q with q selected so that:

$$\frac{1}{\mathrm{T}}\sum_{\mathrm{t}} \ln \hat{\mathrm{p}}_{\mathrm{t}}^{\hat{\theta}_{\varphi}} = \tfrac{1}{2}[\ln \tfrac{1}{2} - \ln(1-\mathrm{q})] + \tfrac{1}{2}[\ln \tfrac{1}{2} - \ln \mathrm{q}] = \tfrac{1}{2}\ln \tfrac{1}{2} - \tfrac{1}{2}\ln[\mathrm{q}(1-\mathrm{q})] \qquad (26)$$

The presumption is that values of q close to ½ 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 $\frac{\mathrm{d}}{2}\ln 2\pi \mathrm{e} + \ln|\Sigma|$ where Σ is the covariance matrix. Hence:

$$\frac{1}{\mathrm{T}}\sum\nolimits_{\mathrm{t}} \ln \hat{\mathrm{p}}_{\mathrm{t}}^{\hat{\theta}_{\varphi}} = \frac{\mathrm{d}}{2}\ln 2\pi \mathrm{e} + \ln|\Sigma| \tag{27}$$

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

$$\hat{\Sigma} = \sum\nolimits_{\mathrm{t}} \hat{\mathrm{p}}_{\mathrm{t}}^{\hat{\theta}_{\varphi}}(\mathrm{x}_{\mathrm{t}} - \hat{\mu})(\mathrm{x}_{\mathrm{t}} - \hat{\mu})' \tag{28}$$

where $\hat{\mu} = \sum\nolimits_{\mathrm{t}} \hat{\mathrm{p}}_{\mathrm{t}}^{\hat{\theta}_{\varphi}} \mathrm{x}_{\mathrm{t}}$ 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.

C. Residual Analysis

Reasonable *a posteriori* probability beliefs can also be identified via the decomposition (11). This task is made easier because there are no unknown parameters in the regression of $\mathrm{p}_{\mathrm{t}}^{\theta}$ on a constant and $\mathrm{g}_{\mathrm{t}}^{\theta}$: the intercept is given by $\frac{1}{\mathrm{T}} + \frac{1}{\mathrm{T}}\bar{\mathrm{g}}_{\mathrm{T}}^{\theta\prime}\mathrm{V}_{\mathrm{T}}^{\theta-1}\bar{\mathrm{g}}_{\mathrm{T}}^{\theta}$ and the slope coefficient vector is given by $-\frac{1}{\mathrm{T}}\bar{\mathrm{g}}_{\mathrm{T}}^{\theta\prime}\mathrm{V}_{\mathrm{T}}^{\theta-1}$. This means that fitted values and residuals can be examined for given values of θ without concern for the effect of outliers and inliers on slope and intercept estimates.

One can begin by applying conventional regression diagnostics to the decomposition of $\hat{\mathrm{p}}_{\mathrm{t}}^{\hat{\theta}_{\varphi}}$ with the idea of identifying the relative contributions of the fitted values $\frac{1}{\mathrm{T}} - \frac{1}{\mathrm{T}}\bar{\mathrm{g}}_{\mathrm{T}}^{\hat{\theta}_{\varphi}\prime}\mathrm{V}_{\mathrm{T}}^{\hat{\theta}_{\varphi}-1}[\mathrm{g}_{\mathrm{t}}^{\hat{\theta}_{\varphi}} - \bar{\mathrm{g}}_{\mathrm{T}}^{\hat{\theta}_{\varphi}}]$ and the residuals $\varepsilon_{\mathrm{t}}^{\mathrm{p}^{\hat{\theta}_{\varphi}}}$. Values of either that are large in absolute value are disproportionately influential in determining the $\hat{\mathrm{p}}_{\mathrm{t}}^{\hat{\theta}_{\varphi}}$ estimates and their associated sample entropy. Large values of the residuals $\varepsilon_{\mathrm{t}}^{\mathrm{p}^{\hat{\theta}_{\varphi}}}$ may be especially informative since the residuals are identically zero if $\frac{1}{T}-\frac{1}{T}\bar{g}_T^{\hat{\theta}_\varphi}{}'V_T^{\hat{\theta}_\varphi-1}[g_t^{\hat{\theta}_\varphi}-\bar{g}_T^{\hat{\theta}_\varphi}]>0\ \forall\ t,$ a condition that will obtain in large samples according to Theorem 3 if $\sup_t \bar{g}_T^{\theta}{}'V_T^{\theta-1}[g_t^{\theta}-\bar{g}_T^{\theta}]=o_p(1)$. Hence, the fitted values and residuals are natural targets for additional scrutiny.

In fact, one can examine local perturbations of the whole probability simplex for plausible values of *θ*. Plausible values of *θ* might be obtained by minimizing (25) for different discrepancy functions *φ*(•) or by bootstrapping the model. For each such *θ*, the fitted values $\frac{1}{T}-\frac{1}{T}\bar{g}_T^{\theta}{}'V_T^{\theta-1}[g_t^{\theta}-\bar{g}_T^{\theta}]$ are fixed and so one can enumerate sets of residuals $\varepsilon_t^{p^\theta}$ that sum to zero, are orthogonal to g_t^θ, and satisfy the lower and upper bound constraints. The relative contributions of these fitted values and residuals in the resulting multinomial probabilities can also be examined for *a priori* plausibility.

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 $p_t^\theta-\frac{1}{T}$ is large – are often thought to represent *peso problems*, events that were expected to happen but that did not eventuate or that did not 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. In these circumstances, we might reasonably expect the prior predictive probability $\bar{P}^\theta(\mathcal{X}_\tau)=\sum_{i=0}^{N_a^\delta-1}\Pi(P_i^\theta)P_i^\theta(\mathcal{X}_\tau)$ of some such subset of the sample space $\mathcal{X}_\tau\subset\mathcal{X}$ to be much larger than the observed frequency 1/T, resulting in a seemingly large value of p_t^θ. Moreover, $\bar{P}^\theta(\mathcal{X}_\tau)$ *is* the *posterior* predictive probability outside the convex hull of the data.

This consideration suggests a third diagnostic to apply to candidate distributions: the calculation of asymptotic highest posterior predictive regions for the bulk of the data. For any model p_t^θ and a given confidence level 1–α, these regions are given by the largest connected subset $\bigcup_{s\in S}\mathcal{X}_s = \mathcal{X}_S \subseteq \mathcal{X}$ for which $\sum_{s\in S} p_s^\theta \le 1-\alpha$, which are just upper and lower quantiles for univariate x_t. In fact, $D_\varphi(\frac{p_t^\theta}{1/T})$ can be modified so that the objective function is the minimization of the distance between the model and given sample upper and lower quantiles for univariate data, calculations which involve straightforward modifications of the assumptions used above. Presumably the modification of $D_\varphi(\frac{p_t^\theta}{1/T})$ for the multivariate case can be handled with multivariate quantile functions of the sort discussed in Serfling (2002), particularly the ones used to estimate the volume of central regions. In any event, calculations along these lines provide explicit identification of potential outliers against which to measure the plausibility of candidate distributions.

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 simple: modulo sampling error, the 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 modest detour down the path of Bayesian semiparametrics. Semiparametric models based on moment conditions do not deliver likelihood functions and the strict application of the Bayesian calculus requires the specification of likelihoods. Moreover, the formation of prior beliefs is more challenging in semiparametric and nonparametric settings because the priors are over the space of likelihood functions that are so hard to specify because there is no guarantee that the data will swamp the prior in such settings. Finally, the literature on priors for semiparametric models is thin and it would appear to be desirable to have a broad class of priors 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 Bayesian constructed in section 2 eliminated these problems. The first was the shift from the model class comprised of densities that respect the moment conditions to that comprised of discrete measures that did so. The second was the presumption that the hypothetical semiparametric Bayesian was a consumer of economic theory who used the model solely for forecasting. The resulting predictive distribution based on a countable set of multinomial likelihood functions proved to be consistent under the weak restriction of positivity of the prior over sufficiently dense sets of multinomial distributions. 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 the probability estimates provided by empirical likelihood and related minimum divergence estimation procedures. On this interpretation, a rejection region and its complement, a confidence region, are not comprised of parameter values but rather of probability beliefs, beliefs that the econometrician can examine for their plausibility. The notion that plausible beliefs can be associated with the parameter values in a rejection region provides a framework for assessing the economic significance of distorted beliefs.

Let me conclude by suggesting four ways in which research along these lines can proceed. First, there is the extension of the tests considered in these pages beyond omnibus goodness-of-fit tests. After all, the difference between the Bayesian and frequentist treatment of nuisance parameters might make it more difficult to equate the beliefs of a semiparametric Bayesian with those of a GMM econometrician. Second, it would be nice to have a semiparametric Bayesian interpretation of higher order asymptotics such as Bartlett corrections. Third, it is natural to extend the results to conditional moment models. While this extension need not be challenging theoretically since Markov chain approximation can replace multinomial approximation, finite sample issues will be more severe since there will be so many empty cells. Finally, it might be interesting to consider a more interesting semiparametric Bayesian, one who has the same objectives but whose decisions affect the sample outcomes as is the case in rational expectations models with learning. Here, too, it might well be substantially more challenging to equate the beliefs of the Bayesian and GMM econometrician.

References

Back, K. and Brown, D. P., 1993. “Implied probabilities in GMM estimators,” *Econometrica* 61, 971-975.

Barron, A. R., 1988. “The exponential convergence of posterior probabilities with implications for Bayes estimators of density functions,” Technical Report 7, Department of Statistics, University of Illinois.

Barron, A. R., 1999. “Information-theoretic characterization of Bayes performance and the choice of priors in parametric and nonparametric problems,” in J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F.M. Smith, (eds.). *Bayesian Statistics 6*. Oxford University Press.

Barron, A. R., and Sheu, C.-H., 1991. “Approximation of density functions by sequences of exponential families**,”** *Annals* of Statistics 19**,** 1347-1369.

Chamberlain, G., 1987. “Asymptotic efficiency in estimation with conditional moment restrictions,” *Journal of Econometrics* 34, 305-334.

Chernoff, H., 1952. “Measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations,” *Annals of Mathematical Statistics* 23, 493–507.

Cressie, N. A. C., and T. R. C. Read, 1984. “Multinomial Goodness-of-Fit Tests,” *Journal of the Royal Statistical Society, Series B* 46, 440–464.

Csiszár, I., 1967. “Information-type measures of difference of probability distributions and indirect observations,” *Studia Sci. Math. Hungar.* 2, 299-318.

Dembski, W. A., 1990. “Uniform probability,” *Journal of Theoretical Probability* 3, 611-626.

Diaconis, P. and D. Freedman, 1986a. “On the Consistency of Bayes Estimates," *Annals of Statistics* 14, 1-26.

Diaconis, P. and D. Freedman, 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.

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.

Freedman, D., 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.

Heath, D., and Sudderth, W. D., 1978. “On finitely additive priors, coherence, and extended admissibility,” *Annals of Statistics* 6, 333-345.

Jiménez, R., and J. E. Yukich, 2002. “Asymptotics for Statistical Distances Based on Voronoi Tessellations,” *Journal of Theoretical Probability* 15, 503-541.

Kim, J.-Y., 2002. “Limited information likelihood and Bayesian Analysis,” *Journal of Econometrics* 107, 175-193.

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., 1988. “Empirical Likelihood Ratio Confidence Intervals for a Single Functional,” *Biometrika* 75, 237 –249.

Owen, A., 2001. *Empirical Likelihood* (New York: Chapman and Hall).

Qin, J., and J. Lawless, 1994. “Empirical likelihood and general estimating equations,” *Annals of Statistics* 22, 300–25.

Rao, C. R., 1958. “Maximum likelihood estimation for the multinomial distribution with an infinite number of cells,” Sankhya 20, 211-218.

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.

Serfling, R., 2002. “Quantile functions for multivariate analysis: approaches and applications,” *Statistica Neerlandica* 56, 214-232.

Stinchcombe, M., 2004. “The unbearable flightiness of Bayesians: generically erratic updating,” working paper, Department of Economics, University of Texas at Austin.

Walker, S. G., 2004. “New approaches to Bayesian consistency,” *Annals of Statistics* 32, 2028-

2043.

Walker, S. G., A. Lijoi, and I. Prünster, 2004. "Contributions to the understanding of Bayesian consistency," ICER Working Paper no. 13/2004.

Zellner, A., 1994. "Model, prior information and Bayesian analysis," *Journal of Econometrics* 75, 51–68.

Zellner, A., 1997. "The Bayesian method of moments (BMOM): theory and application," in Fomby, T., and Hill, R. C. (eds.). *Advances in Econometrics*. Cambridge University Press.