SOME GENERICITY ANALYSES IN NONPARAMETRIC STATISTICS[‡]

MAXWELL B. STINCHCOMBE

ABSTRACT. Many nonparametric estimators and tests are naturally set in infinite dimensional contexts. Prevalence is the infinite dimensional analogue of full Lebesgue measure, shyness the analogue of being a Lebesgue null set.

A prevalent set of prior distributions lead to wildly inconsistent Bayesian updating when independent and identically distributed observations happen in class of infinite spaces that includes $\mathbb{R}^n$ and $\mathbb{N}$.

For any rate of convergence, no matter how slow, only a shy set of target functions can be approximated by consistent nonparametric regression schemes in a class that includes series approximations, kernels and other locally weighted regressions, splines, and artificial neural networks.

When the instruments allow for the existence of an instrumental regression, the regression function only exists for a shy set of dependent variables. The instruments allow for existence in a counterintuitive dense set of cases, shyness is an open question.

A prevalent set of integrated conditional moment (ICM) specification tests are consistent, a dense subset of the finitely parametrized ICM tests are consistent, prevalence is an open question.

Date: December 10, 2002.

[‡]Many thanks Xiaohong Chen, Stephen Donald, and Haskell Rosenthal, Dan Slesnick, Hal White, and Paul Wilson for helpful conversations about this paper. Tom Wiseman and Jeff Ely helped with some difficult terminological issues, but I'm not sure they want to be thanked.

1. Introduction

For some, but not all prior distributions, Bayesian updating based on i.i.d. real-valued random variables is consistent. Given a sequence $r_n \to 0$, for some, but not all functional relations $f(x) = E(Y\big|X = x)$, consistent nonparametric regression techniques for estimating f converge at rate $\mathcal{O}(r_n)$. Under the compact operator assumptions used in devising estimators for instrumental regressions, for some, but not all dependent variables, instrumental regressions exist. Some, but not all, integrated conditional moment (ICM) tests are consistent for any deviation from the null.

These observations mean that absolute answers to a number of important questions are not possible. We cannot say that Bayesian updating is consistent. We cannot say that all functions are $\mathcal{O}(r_n)$ approximable. We cannot say that instrumental regressions always exist. We cannot say that all ICM tests are consistent.

This paper answers the questions that flow from the lack of absolute answers. "How big is the set of priors for which Bayesian updating is consistent?" "How big is the set of functions approximable at rate $\mathcal{O}(r_n)$?" "How big is the set of dependent variables for which instrumental regressions exist?" Finally, "How big is the set of consistent ICM tests?"

The answers to these questions are given in terms of prevalence and shyness. The nonparametric estimators and tests considered here are naturally set in infinite dimensional contexts. Prevalence is the infinite dimensional analogue of full Lebesgue measure, or genericity, shyness the analogue of being a Lebesgue null set, of being non-generic.

In short, the answers are:

1. Given a true distribution θ with infinite support in a locally compact space, it is a prevalent property of priors that with probability 1, for every strictly positive ϵ and every non-empty open G, the Bayesian posterior distribution assign mass at least $1 - \epsilon$ to G infinitely often. Thus, Bayesian updating can only be consistent for a shy set of prior distributions.
2. Given any rate $r_n \to 0$, only a shy set of functions is approximable at rate $\mathcal{O}(r_n)$ using any of a wide class of nonparametric estimation techniques. This means that the regularity conditions invoked to guarantee faster rates of estimator convergence, usually smoothness assumptions, restrict attention

to a shy class of functions, something that is at odds with the intended interpretations of consistency results.

3. Given the compact operator assumptions invoked to devise estimation techniques, instrumental regressions exist only for a shy set of dependent random variables. The sufficient conditions for the compact operator assumptions are themselves only satisfied for a shy set of distributions of the explanatory and instrumental variables.
4. Finally, within the set of ICM tests, consistency is prevalent. Within the set of ICM tests that are finitely parametrized, consistency is dense, whether or not it is prevalent is an open question.

Genericity analyses run the risk of being wrong-headed. For example, suppose an estimation technique is consistent and/or efficient only if the mean, μ, and the standard deviation, σ, of the population are equal. When μ and σ are viewed as unrelated (beyond both being positive), the whole positive orthant must be considered. Relative to the positive orthant, the diagonal along which $\mu = \sigma$ is a non-generic set. In this case, the technique would be judged to be generically inconsistent and/or inefficient. This conclusion is not informative if (μ, σ) is known to lie on the diagonal (as it might in some counting processes or after some variance stabilizing transformations). More generally, the conclusion that a set $E \subset \mathfrak{X}$ is "large" relative to $\mathfrak{X}$ can happen because $\mathfrak{X}$ is "too small," the conclusion that a set $E \subset \mathfrak{X}$ is "small" relative to $\mathfrak{X}$ can happen because $\mathfrak{X}$ is "too large." Each of sections concludes with an examination of these interpretational issues, and the final section considers these issues in a unified fashion.

The next section contains the necessary background on genericity for the infinite dimensional contexts considered here. The subsequent sections discuss, in turn, the generic inconsistency of Bayesian updating, the shyness of the set of regression functions that can be approximated at any given rate, the generic properties of nonparametric instrumental variable regressions, and the generic consistency of integrated conditional moment tests. The material in the section on nonparametric instrumental variable regression depends slightly on the material in the section on general nonparametric regression. Aside from this, the sections are mutually independent. Proofs are gathered in the appendix.

2. Small Sets and Large Sets

Throughout, $\mathfrak{X}$ denotes an infinite dimensional, locally convex, topological vector space that is also a complete separable metric (csm) space. From Rudin (1973, Theorem 1.24, p. 18), the topology on $\mathfrak{X}$ can be metrized by a translation invariant metric $d(\cdot,\cdot)$, that is, $d(x,y) = d(x+z, y+z)$ for all $x, y, z \in \mathfrak{X}$. This class of spaces includes (but is not limited to):

1. separable Banach spaces such as the $L^p(\Omega, \mathcal{F}, P)$ spaces, $1 \leq p < \infty$, $\mathcal{F}$ countably generated;
2. $C(X)$, the continuous functions with the sup norm when X is compact;
3. $C(X)$ with the topology of uniform convergence on compact sets when X is locally compact and separable;
4. the Sobolev spaces $S_m^p(\mathbb{R}^r, \mu)$ defined as the metric completion of $C_m^p(\mathbb{R}^r, \mu)$, the space of m times continuously differentiable functions, $m \geq 0$, on $\mathbb{R}^r$ having finite norm $\|f\|_{p,m,\mu} = \sum_{|\alpha|\leq m} \left[\int |D^\alpha f(x)|^p \, d\mu(x)\right]^{\frac{1}{p}}$, $p \in [1, \infty)$, μ a Borel probability measure on $\mathbb{R}^r$;
5. $C^m(X)$, the space of m times continuously differentiable functions on a compact X with the norm $\sum_{|\alpha|\leq n} \max_{x\in X} |D^\alpha f(x)|$.

There are two notions of rarity available for $\mathfrak{X}$. The topological notion called *meagerness* is due to Baire (1899, §59-61, pp. 65-67). The measure theoretic notion called *Haar zero* sets is due to Christensen (1974, Ch. 7). Its properties and applications were more thoroughly investigated under the name of *shy* sets by Hunt, Sauer and Yorke (HSY, 1992), who especially applied these techniques to the study of the generic behavior of dynamical systems. There are subtle and difficult problems in extending shyness to a definition of non-generic for subsets of convex subsets of vector spaces that are themselves shy, e.g. spaces of probability measures. These problems were discovered and resolved by Anderson and Zame (2001).

2.a. **Meager and residual sets.** A closed set with no interior seems small.

Definition 2.1. *A set S is* **nowhere dense** *if its closure has no interior. A set S is* **meager** *if it can be expressed as a countable union of nowhere dense sets. A set E is* **residual** *or* **Baire large** *if it is the complement of a meager set.*

Baire large sets are, equivalently, the countable intersection of open dense sets. The countable union of meager sets is meager, the countable intersection of Baire large sets is a Baire large set. Because $\mathfrak{X}$ is a csm space, any Baire large subset is dense, and to some extent this justifies thinking of residual sets as being "large" or "generic". Baire large sets can have Lebesgue measure 0 and seem quite small in $\mathbb{R}^k$ ($k < \infty$ throughout).

Example 2.1. *Let q_n be an enumeration of the vectors in $\mathbb{R}^k$ with rational coordinates. For any rational $\epsilon > 0$, let E_ϵ be the union of open balls centered at q_n, $\cup_n B(q_n, \epsilon/2^n)$. E_ϵ is an open dense subset of $\mathbb{R}^k$ having Lebesgue measure less than ϵ. The set $E = \cap_\epsilon E_\epsilon$ is a residual set having Lebesgue measure 0.*

2.b. **Shy and prevalent sets.** For $\mathbb{R}^k$, we have the following.

Lemma 2.1. *For a universally measurable $S \subset \mathbb{R}^k$, the following are equivalent*

1. *$\Lambda^k(S) = 0$ where Λ^k is k-dimensional Lebesgue measure,*
2. *$G^k(S) = 0$ where G^k is a non-degenerate Gaussian distribution, and*
3. *there exists a compactly supported probability η such that $\eta(S + x) = 0$ for all $x \in \mathbb{R}^k$.*

The third condition in Lemma 2.1 generalizes to $\mathfrak{X}$. Taking the measure η to be the continuous linear image of U^k, the uniform distribution on $[0, 1]^k$, is so useful that it merits a special name.

Definition 2.2 (Christensen. Hunt, Sauer, and Yorke)**.** *A subset S of a universally measurable $S' \subset \mathfrak{X}$ is* **shy** *if there exists a compactly supported probability η such that $\eta(S' + x) = 0$ for all x. S is* **finitely shy** *if η can be taken the continuous linear image of U^k for some k. The complement of a (finitely) shy set is a* (**finitely**) **prevalent** *set.*

From HSY (1992, Facts $2'$ and $3''$), no S containing an open set can be finitely shy in $\mathfrak{X}$ so that prevalent sets are dense, and countable unions of shy sets are shy, equivalently, countable intersections of prevalent sets are prevalent

2.c. **Approximately flat sets.** For $A, B \subset \mathfrak{X}$, define $A + B = \{a + b : a \in A,\ b \in B\}$. For any sequence of A_n of sets, $[A_n$ i.o.], read as "A_n infinitely often," is defined as $\bigcap_m \bigcup_{n \geq m} A_n$. In a similar fashion, $[A_n$ a.a.], read as "A_n

almost always," is defined as $\bigcup_m \bigcap_{n \geq m} A_n$. The following definition and Lemma will be used frequently.

Definition 2.3. *A set $F \subset \mathfrak{X}$ is* **approximately flat***, if for every $\epsilon > 0$, there is a finite dimensional subspace W of $\mathfrak{X}$ such that $F \subset W + B(0, \epsilon)$ where $B(x, r)$ is the ball around x with radius r.*

Any finite union of approximately flat sets is approximately flat, and every compact set is approximately flat — let W be the span of a finite ϵ-net. The following is the basic lemma used Stinchcombe (2001).

Lemma 2.2. *For any sequence, F_n, of approximately flat sets and any $r_n \to 0$, the set $[(F_n + B(0, r_n))$ i.o.$]$ is shy.*

Taking $F_n \equiv F$ shows that the closure of any approximately flat (e.g. any compact set) set is shy.

One intuition for the shyness of $[(F_n + B(0, r_n))$ i.o.$]$ comes from how small approximately flat sets are. If W^d is a d-dimensional subspace of $\mathbb{R}^k$, then, as a proportion of the unit ball, $W^d + B(0, \epsilon)$ is on the order of ϵ^{k-d}. This leads one to suspect that approximately flat sets are "small" in infinite dimensional $\mathfrak{X}$'s.

2.d. **Shy subsets of convex sets.** Suppose that C is a convex subset of $\mathbb{R}^k$. Defining $S' \subset C$ to be shy relative to C by asking that $S' = S \cap C$ for some shy $S \subset \mathbb{R}^k$ make C a shy subset of itself if $\dim(C) < k$. However, taking $\mathbf{aff}(C)$ to be the smallest affine subspace containing C and using the lower dimensional Lebesgue measure on $\mathbf{aff}(C)$ delivers the appropriate definition.

An example directly relevant to this paper demonstrates that the affine subspace approach does not generally work in $\mathfrak{X}$. Take $\mathfrak{X}$ to be the set of countably additive, finite, signed measures on $2^{\mathbb{N}}$, $C = \Delta(\mathbb{N}) \subset \mathfrak{X}$ to be the probability measures on $2^{\mathbb{N}}$. $\Delta(\mathbb{N})$ is a finitely shy subset of $\mathfrak{X}$. (Let η be the uniform distribution on the line L joining the 0 measure and any point mass, δ_n. For any $x \in \mathfrak{X}$, $L \cap (\Delta(\mathbb{N}) + x)$ contains at most one point, so that $\eta(\Delta(\mathbb{N}) + x) = 0$.) However, $\mathbf{aff}(\Delta(\mathbb{N})) = \mathfrak{X}$.

Working from the "outside," that is, with $\mathbf{aff}(C)$, is not appropriate in $\mathfrak{X}$. The path-breaking work of Anderson and Zame (2001) give a definition of shy subsets of C that works from the "inside." For any $c \in C$, C convex, and any $\epsilon > 0$, the set $\epsilon C + (1 - \epsilon)c$ is a convex subset of C that shrinks C toward c.

For a universally measurable $C \subset \mathfrak{X}$, $\Delta^K(C)$ denotes the compactly supported probability measures on C.

Definition 2.4 (Anderson and Zame)**.** *Let C be a convex subset of $\mathfrak{X}$ that is topologically complete in the relative topology. A subset S of a universally measurable $S' \subset C$ is* **shy relative to** *C if for all $c \in C$, all neighborhoods U_c of c, and all $\epsilon > 0$, there exists a $\eta \in \Delta^K(C)$ such that $\eta(U_c \cap [\epsilon C + (1-\epsilon)c]) = 1$ and $(\forall x \in \mathfrak{X})[\eta(S' + x) = 0]$. S is* **finitely shy relative to** *C if there exists a if $\eta \in \Delta^K(C)$ that is the continuous affine image of U^k for some k such that $(\forall x \in \mathfrak{X})[\eta(S' + x) = 0]$. The complement of a (finitely) shy set is a* (**finitely**) **prevalent** *set.*

Anderson and Zame (2001) show that finite shyness is sufficient for shyness, that the countable union of shy sets is shy, and that shyness is equivalent to the "Lebesgue measure on $\mathbf{aff}(C)$" definition in $\mathbb{R}^k$. They demonstrate the utility of their definition of shyness relative to convex sets in a number of contexts in theoretical economics.

For $S \subset C$, C a convex subset of $\mathfrak{X}$, $\overline{\mathbf{co}}\, S$ is the closure of its convex hull. For $S, T \subset C$, $\operatorname{rel\,int}_S(T)$ is the interior of S relative to T. The following sufficient condition for shyness relative to a convex set will be useful below.

Lemma 2.3. *Let C be a convex subset of $\mathfrak{X}$ that is topologically complete in the relative topology. If $\overline{\mathbf{co}}\, S$ has empty interior relative to C, then S is a finitely shy subset of C.*

2.e. **Interpretational issues.** In $\mathbb{R}^k$, Lemma 2.1 ties together Lebesgue measure, a probabilistic interpretation, and a translation invariant property of the smallness of a set S. Lebesgue measure fails to extend to $\mathfrak{X}$ because there is no translation invariant measure on $\mathfrak{X}$ assigning positive mass to any open set. If there were, it would have to assign equal, and strictly positive mass to every open ball $B(x, \epsilon/4)$. Since $\mathfrak{X}$ is infinite dimensional, every $B(y, \epsilon)$ contains countably many disjoint balls with radius $\epsilon/4$, and the measure assigned to every open set would therefore be infinite.

Probability measures on csm spaces are tight, that is, for every $\epsilon > 0$, there is a compact set, F_ϵ, with $P(F_\epsilon) > 1 - \epsilon$. Probabilistic interpretations fail to extend to $\mathfrak{X}$ directly because the tightness of any probability P on $\mathfrak{X}$ implies

that $P(S) = 1$ for S being the countable union of compact, hence shy, sets. Probabilistic interpretations of shyness also fail to be approximately true.

Let Y_i be an independent and identically distributed (iid) sequence of random variables distributed P. Suppose that that $r_n \to 0$, and that $N_n \to \infty$. A point $x \in \mathfrak{X}$ is (r_n, N_n)-**lonely** if $P^\infty(A(x)) = 0$ where $A(x) = [A_n(x) \text{ i.o.}]$, $A_n(x) = \{d(Y_i, x) < r_n \text{ for some } i \leq N_n\}$. In other words, the x is lonely if, with probability 1, $B(x, r_n)$ eventually receives no more visits from $Y_1, \ldots, Y_{N_n}$. Stinchcombe (2001) shows that, no matter how slowly r_n goes to 0 or how quickly N_n goes to ∞, a prevalent set of points are (r_n, N_n)-lonely.

3. The Generic Inconsistency of Bayesian Updating

Bayesian updating approaches to and understandings of statistical problems and results are widespread. It is therefore striking that Bayesian updating is generically inconsistent. The classic result is that for a Baire large set of priors on the distributions on the integers, Bayesian updating is wildly inconsistent (Freedman 1965). Since Baire large sets can be shy, and several of the constructions in Freedman's proof use sets that turn out to be shy, one might have hoped (as the author did) that this wild inconsistency was non-generic.

Bayesian updating and optimization in the face of uncertainty are intimately tied, nowhere more so than in the theory of learning. Nachbar's (1997) crucial result for infinitely repeated games is that, when combined with optimization, Bayesian updating of priors about other players' repeated game strategies is leads the players to play strategies that others were certain were not going to be played. Generic inconsistency implies that for interesting single agent games, Bayesian updating is "objectively" sub-optimal. It also provides a optimizing rationalization of fads, bubbles, and other seemingly irrational behavior.

3.a. **Overview for real-valued random variables.** Suppose that we observe a sequence of iid $\mathbb{R}$-valued random variables $(Y_n)_{n\in\mathbb{N}}$. Let $\theta \in \Delta(\mathbb{R})$ denote the true, full support distribution of each Y_n. Let θ^∞ denote the distribution of the sequence of Y_n's. A prior distribution, μ, is a distribution on $\Delta(\mathbb{R})$. Notationally, $\mu \in \mathbb{M} = \Delta(\Delta(\mathbb{R}))$.

After observing a partial history $h^t = (x_1, \ldots, x_t) \in \mathbb{R}^t$, the posterior beliefs $\mu(\cdot|h^t)$ are formed using Bayes' Law. Bayesian updating is **consistent** if for θ^∞

almost all histories, posterior beliefs converge, in the weak* topology, to putting mass 1 on θ, $\mu(\cdot|h^t) \to_{w^*} \delta_\theta$.

Null sets can cause serious problems in updating — conditional probabilities are arbitrary on null sets, and the set of observed h^t typically belong to null sets. The best solution is to use densities. To avoid these issues, one must work in a countably infinite space such as $\mathbb{N}$.

Fix a full support, σ-finite reference measure, λ, on $\mathbb{R}$, e.g. Lebesgue measure. Let C_+^λ denote the set of continuous, non-negative functions f such that $\int_X f\, d\lambda = 1$, $C_{++}^\lambda \subset C_+^\lambda$ is the set of strictly positive f. Each $f \in C_+^\lambda$ is uniquely associated with the probability $\theta_f \in \Delta(X)$ defined by $\theta_f(A) = \int_A f\, d\lambda$.

Updating after partial history $h^t = (x_1, \ldots, x_t) \in X^t$ is done using the values of the densities at h^t and the prior, μ, M_B^λ,

$$\mu(B|h^t) := \frac{\int_B \Pi_{i=1}^t f(x_t)\, d\mu(f)}{\int_{C_+^\lambda} \Pi_{i=1}^t f(x_t)\, d\mu(f)},\ B \subset C_+^\lambda. \tag{1}$$

$\mathbb{M}_B^\lambda \subset \mathbb{M}^\lambda$ are those beliefs which will never involve division by 0 in (1),

$$\mathbb{M}_B^\lambda = \{\mu \in \mathbb{M} : \mu(C_+^\lambda) = 1,\ \forall (x_1, \ldots, x_t)\ \mu(\{f : \Pi_{i=1}^t f(x_i) > 0\}) > 0\,\}. \tag{2}$$

Theorem 3.1 shows that for any full support θ, Bayesian updating is consistent only for a shy set of $\mu \in \mathbb{M}_B^\lambda$. This is true even when $\theta = \theta_f$ for some $f \in C_+^\lambda$. One step in the proof contains some insight into how inconsistency can arise.

Example 3.1. *For arbitrary full support θ and θ°, there is a dense set of priors with the property that θ^∞-a.e., posterior beliefs converge to θ°.*

To see why, let $M \subset \mathbb{M}$ denote the set of beliefs, ν, that rule out some non-empty, open $V \subset \mathbb{R}$, that is, for which V, $\nu(\{\theta : \theta(V) > 0\}) = 0$. Let $\mathbb{D}^\circ$ be the set of μ's of the form $\alpha_0 \delta_{\theta^\circ} + \sum_{i \le I} \alpha_i \nu_i$ where $\alpha_0 > 0$ and $\nu_i \in M$. $\mathbb{D}^\circ$ is easily seen to be dense in $\mathbb{M}_B^\lambda$. Since the true θ is full support, every non-empty open V will be visited infinitely often θ^∞-a.e. Therefore, for every $\mu \in \mathbb{D}^\circ$, only the full support θ° can have positive posterior weight in the limit, that is, $\mu(\cdot|h^t) \to_{w^*} \delta_{\theta^\circ}$.

It is important to note that the prior beliefs used in Example 3.1 are not an indication of the kind of priors that one must have for Bayesian updating to be inconsistent, they are but a device used in the proof. Indeed, a direct implication of Theorem 3.1 is that Bayesian updating is only consistent for a shy subset of the full support beliefs in $\Delta(C_{++}^\lambda)$. By contrast, the priors in Example 3.1 may put

mass less than 1 on $\Delta(C^\lambda_{++})$. The inconsistency of Example 3.1 only scratches the surface of how badly Bayesian updating of generic priors behaves.

A pair (μ, θ) is **erratic** written $\mu \in \text{err}(\theta)$, if for all non-empty open subsets G of $\Delta(\mathbb{R})$, $\limsup_t \mu(G|h^t) = 1$ for a set of histories having θ^∞ probability 1. Theorem 3.1 shows that for any full support $\theta \in \Delta(\mathbb{R})$, $\text{err}(\theta)$ is prevalent in $\mathbb{M}^\lambda_B$.

Because the set of full support μ is a prevalent subset of $\mathbb{M}^\lambda_B$, the result does not arise from some kind of failure of support conditions. For the same reason, the result continues to hold if one restricts attention to beliefs that are full support on the set of probabilities having only strictly positive densities. Further, since the result holds for *any* full support θ, it also does not arise by picking the true θ to be outside the set supporting the prior beliefs.

3.b. **Generic inconsistency.** Fix an infinite, locally compact, complete, separable metric (csm) space (X, d) with $\mathcal{X}$ denoting Borel sigma-field. (A space is locally compact if every point has a neighborhood with compact closure. The spaces $\mathbb{R}^\ell$ and $\mathbb{N}$ are locally compact, infinite dimensional topological vector spaces are not.) $\Delta(X)$ denotes the set of (countably additive, Borel) probabilities on $\mathcal{X}$. An iid sequence of draws, $(Y_n)_{n\in\mathbb{N}}$, is made according to a distribution $\theta \in \Delta(X)$, and θ^∞ denotes the corresponding product distribution on $X^{\mathbb{N}}$. Prior beliefs, μ, are points in $\Delta(\Delta(X))$, the set of distributions on the set of distributions on X. Both $\Delta(X)$ and $\Delta(\Delta(X))$ are csm's in the weak* topology.

Define $\theta_\mu \in \Delta(X)$ by $\theta_\mu(E) = \int_\Theta \theta(E)\, d\mu(\theta)$ for $\mu \in \Delta(\Delta(X))$. If $\theta_\mu \not\gg \theta$, the choice of version of the conditional probabilities matters quite sharply. As noted above, to relate μ to the updated, versions of μ conditional on finite histories of draws, one must make a whole system of coordinated choices of versions. This will be done by assuming that the μ's under study put mass 1 on the dense set of θ's having continuous densities with respect to some full support reference measure. Throughout, a simple reference case has X countable and discrete, in which case the σ-finite reference measure, λ, can be taken to be counting measure, and all probabilities have densities with respect to λ.

1. λ is a full support, σ-finite reference measure on $\mathcal{X}$. $C(X)$ is the set of continuous functions on X. $C^\lambda_+ \subset C(X)$ is the set of non-negative f such that $\int_X f\, d\lambda = 1$, $C^\lambda_{++} \subset C^\lambda_+$ is the set of strictly positive f. Each $f \in C^\lambda_+$ is

associated with a probability $\theta_f \in \Delta(X)$ defined by $\theta_f(A) = \int_A f\, d\lambda$. When X is countable and discrete, $C^\lambda_+ = \Delta(X)$.

2. Both C^λ_{++} and C^λ_+ are G_δ's in the csm $\Delta(X)$, implying that there are complete separable metrics, d_{++} and d_+ inducing the weak* topology. (A G_δ is a countable intersection of open sets. The relative topology on any G_δ in a csm can be metrized with a complete separable metric. It is easy to give explicit metrics making C^λ_+ and C^λ_{++} into csm's.)
3. $\mathbb{M}^\lambda \subset \Delta(\Delta(X))$ denotes the set of probabilities on probabilities $\Delta(C^\lambda_+)$, while $\mathbb{M}^\lambda_{++}$ denotes $\Delta(C^\lambda_{++})$. $\mathbb{M}^\lambda_B$ are those for which Bayes updating using densities will never involve division by 0, formally,

$$\mathbb{M}^\lambda_B = \{\mu \in \mathbb{M}^\lambda : \forall (x_1, \ldots, x_t)\ \mu(\{f : \Pi^t_{i=1} f(x_i) > 0\}) > 0\,\}.$$

From the definitions, $\mathbb{M}^\lambda_{++} \subset \mathbb{M}^\lambda_B \subset \mathbb{M}^\lambda$. It can be shown that $\mathbb{M}^\lambda_{++}$ is a convex, topologically complete, prevalent subset of the convex csm $\mathbb{M}^\lambda$, and $\mathbb{M}^\lambda_B$ is a G_δ, hence topologically complete.

Assuming that $\mu \in \mathbb{M}^\lambda_B$, updating after partial history $h^t = (x_1, \ldots, x_t) \in X^t$ is done using the values of the densities at h^t and the prior, μ,

$$\mu(B\big|h^t) := \frac{\int_B \Pi^t_{i=1} f(x_t)\, d\mu(f)}{\int_{C^\lambda_+} \Pi^t_{i=1} f(x_t)\, d\mu(f)},\ B \subset C^\lambda_+. \tag{3}$$

Definition 3.1. *For any $\theta \in \Delta(X)$, Cons$(\theta) \subset \mathbb{M}^\lambda_B$ denotes the set of μ in $\mathbb{M}^\lambda_B$ that are* **consistent for** *θ, that is, the set of beliefs that satisfy $\mu(\cdot\big|h^t) \to_{w^*} \delta_\theta$ θ^∞-a.e. A pair (μ, θ) is* **erratic**, **wildly inconsistent**, **fickle**, *or* **faddish**, *written $\mu \in \mathrm{err}(\theta)$, if for all non-empty open subsets G of $\Delta(X)$, $\limsup_t \mu(G\big|h^t) = 1$ θ^∞-a.e.*

Being erratic is a very strong form of failing to be consistent.

Theorem 3.1. *For any full support $\theta \in \Delta(X)$, $\mathrm{err}(\theta)$ is prevalent in $\mathbb{M}^\lambda_B$.*

3.c. **Comments.** Because the set of full support μ's is prevalent in $\mathbb{M}^\lambda$, the set of full support elements of $\mathrm{err}(\theta)$ is prevalent. This means that Theorem 3.1 does not arise because of some generalized failure of support conditions. In a similar vein, since $\mathbb{M}^\lambda_{++} \subset \mathbb{M}^\lambda_B \subset \mathbb{M}^\lambda$ and $\mathbb{M}^\lambda_{++}$ is a prevalent subset of $\mathbb{M}^\lambda$, $\mathrm{err}(\theta) \cap \mathbb{M}^\lambda_{++}$ is a prevalent subset of $\mathbb{M}^\lambda_{++}$. Also, if $f \in C^\lambda_{++}$, θ_f is full support and $\mathrm{err}(\theta_f)$ is

prevalent. Theorem 3.1 does not arise because the full support θ is necessarily outside the set of probabilities supporting μ.

The continuity of the densities can be weakened — the result holds if the set of densities being considered are continuous with respect to a metric for which: (a) λ is still full support, and (b) the Borel σ-field is still σ-field $\mathcal{X}$.

When $X \subset \mathbb{R}$, the Glivenko-Cantelli theorem tells us that θ^∞-a.e., the empirical cdf converges uniformly to the cdf of θ. Generically, Bayes estimators behave much differently, not converging to the true θ nor to anything else. When $X = \mathbb{N}$, Freedman (1965) shows that a Baire large set of (μ, θ) pairs in $\Delta(\Delta(\mathbb{N})) \times \Delta(\mathbb{N})$ are erratic. This uses a "Fubini" theorem for Baire sets. Anderson and Zame (2001, Example 4, p. 57) show that no such Fubini result is available for prevalent sets.

3.d. **Interpretational issues.** First, it is possible that the set of consistent priors being shy might is a result of working in the space of all priors, and this is simply "too large" a space. Second, if a generic Bayesian optimizer updates an erratic prior, their strategy, while seeming to ignore the accumulating evidence, is still optimal given their preferences. Third, a population of identical agents with an erratic prior demonstrate episodic near certainty about the truth of very different propositions, much as intellectual (or other) fads do. In other words, fads may arise from optimal updating.

3.d.1. *Too large a set of priors?* Suppose that $X = \mathbb{N}$, and suppose that the prior is essentially finite dimensional in the inference patterns relating different observations to each other. In this case, we have a small enough class of priors that the consistent ones form a prevalent subset. Freedman (1963) and Diaconis and Freedman (1986a) discuss the following intuition. Priors encode detailed prejudices about the likely shape of tail behavior of the distribution. Further, using Bayes' Law implies that priors can react very strongly to rare events. Among the huge number of partial histories, there will, for typical priors, be at least one that interacts arbitrarily strongly with the prior's prejudices about the relations between rare events.

When $X = \mathbb{N}$, each $\theta \in \Delta(X)$ is specified by the countably many numbers $\theta(n)$. Let $S_k(\theta) = \sum_{n \leq k} \theta(n)$. Picking a θ according to μ gives rise to countably many random variables $Y_k = (1 - S_{k-1}(\theta))^{-1}\theta(k)$. Following Freedman (1963),

a prior μ is **tail-free** if $\mu(S_k < 1) = 1$ for all k, and there exists a K such that the random vector $(\theta_k)_{k=1}^K$ and the random variables $Y_{K+1}, Y_{K+2}, \ldots$ are mutually independent (see Ferguson (1973) for a wide set of applications of these ideas). With tail-free priors, large observations in $\mathbb{N}$ have no information about the smaller observations, and there is essentially only a finite dimensional set of relations between the different observations.

The extent to which these intuitions generalize to more general X is somewhat unclear, see Arnold *et al* (1984) and Diaconis and Freedman (1986b) for relatively natural settings (competing risks and location estimators respectively) in which Bayes estimators are not consistent.

3.d.2. *Optimization and consistency.* The paradox of wildly inconsistent updating is that the posterior beliefs, by not settling down, contrast sharply with a major implication of the maintained iid assumption, that the empirical cdf does settle down. It might seem that such behavior has negative implications for the optimality of dynamic expected utility maximization in iid contexts.

Suppose that at each t, a decision maker chooses $a_t \in \{a, b\}$ after having observed the realizations of iid θ random integers $X_1, \ldots, X_{t-1}$ (with X_0 an arbitrary constant). After a_t is chosen, X_t is realized and observed, and the decision maker receives reward $R_t = R(a_t, X_t) \in [0, 1]$. Corresponding to any infinite sequence $((a_t), (x_t))_{t=1}^\infty$ of actions and realizations is the sequence of rewards $\tilde{R} = (R_t)_{t=1}^\infty = (R(a_t, x_t))_{t=1}^\infty$. Sequences of rewards are evaluated using the Bernoulli utility function $V_\beta(\tilde{R}) := (1-\beta)\sum_t \beta^{t-1} R_t$, $0 < \beta < 1$. The decision maker's preferences over distributions on possible reward sequences are represented by the expected utility function $E^\mu V_\beta(\tilde{R})$.

Let $\tilde{R}^e$ (respectively $\tilde{R}^\mu$) be the achievable sequence of rewards associated with choosing a_t to best respond to the empirical distribution of the history $h^{t-1} = X_0, \ldots, X_{t-1}$ (respectively to the posterior $\mu(\cdot|h^{t-1})$). Suppose that a is not a dominant action, i.e. there exists a $x^\circ \in \mathbb{N}$ such that $R(a, x^\circ) < R(b, x^\circ)$, but that a is the strict best response to θ, i.e. $\overline{r} = \int R(a, x)\, d\theta(x) > \underline{r} = \int R(b, x)\, d\theta(x)$.

Since the true distribution is θ^∞, the "objective" maximizing strategy is $a_t \equiv a$, delivering an "objective" expected utility $\overline{r}$. By the Glivenko-Cantelli theorem, with θ^∞ probability 1, $\tilde{R}^e$ is $\overline{r}$ in expectation with at most finitely many exceptions. By contrast, the wild inconsistency of Bayesian updating implies that $\tilde{R}^\mu$

is $\underline{r}$ infinitely often. This seems to indicate that patient optimizers will prefer $\tilde{R}^e$ to $\tilde{R}^\mu$, that the consistency embodied in the empirical cdf approach leads to higher payoffs, at least for patient decision makers.

Formalizing such an argument would require the inequality

$$E^{\delta_\theta} V_\beta(\tilde{R}^e) > E^{\delta_\theta} V_\beta(\tilde{R}^\mu), \tag{4}$$

and it is here that the fallacy appears clearly. It is quite easy to give generic μ's and θ in the support of μ satisfying (4). However, an expected utility chooses actions bearing in mind a wide range of possibilities, and knowing that this caution may involve the "wrong" action being chosen from time to time. Evaluating a course of action under δ_θ, that is, under certainty about the true distribution is, from the decision maker's point of view, an entirely irrelevant exercise.

3.d.3. *Fads.* Suppose that the action set $\{a, b\}$ above is replaced by an infinite set with the property each action is the unique maximizer of $R(a, \theta)$ for some $\theta \in \Delta(\mathbb{N})$. Another way to understand the oddity of wild inconsistency is that new observations on the same process arrive, and beliefs and actions will wander arbitrarily far away from the historical record infinitely often. If a large population of identical agents behaves in so erratic a fashion, one is tempted to look for a model of irrationality, perhaps a model of informational cascades, or a model of bubbles, or to look for exogenous sources of variability, in short, to look for a model of fads. One point to be taken from the present result is that there is a huge variety of rational behavior, even in the quite limited case of iid observations.

4. Rates of Convergence for Nonparametric Regression

Suppose that interest centers on finding $f(x) = E(Y|X = x)$, and the target function f belongs to separable Banach space $\mathfrak{X}$, typically an $L^2(\mathbb{R}^\ell, \mu)$ space or a Sobolev space, $S_m^p(\mathbb{R}^\ell, \mu)$, of smooth functions. In this section, $\mathfrak{X}$ is assumed to be a Banach space. For any rate of approximation $r_n \to 0$, no matter how slow, only a shy set of functions f are $\mathcal{O}(r_n)$-approximable by a wide class of the available nonparametric regression techniques. The present analysis concerns only the deterministic part of the regression error, not the part due to noise in the observations.

Consistency results for nonparametric regression techniques show that the approximation error converges to 0. Rate of convergence results show that the

approximation error is $\mathcal{O}(r_n)$ if the target function satisfies regularity conditions, usually smoothness conditions of some sort. The result here shows that the regularity conditions can only be satisfied by shy sets of functions, even when approximation is being done within infinite dimensional classes of smooth functions.

The argument for the shyness of the set of $\mathcal{O}(r_n)$-approximable functions has two steps:

1. Available nonparametric regression estimators are functions in a sequence $C_n \subset \mathfrak{X}$ of compactly generated two-way cones.
2. Given a sequence C_n of estimators, $[C_n + B(0, M \cdot r_n)$ a.a.] is the set of f at distance less than or equal to $M \cdot r_n$ almost always. If the C_n are compactly generated two-way cones, this set is shy.

The arguments are simplest in the context of Fourier series estimators.

4.a. **Fourier series estimators.** Interest centers on estimating $f(x) = E\left(Y \middle| X = x\right)$ from repeated observations of the random variables X and Y. We assume that $f \in L^2(X)$, the set of square integrable functions of X. Fix an orthonormal basis $\{e_k : k \in \mathbb{N}\}$ for $L^2(X)$. The unknown function f has an infinite series representation $f = \sum_k \beta_k e_k$ where $\beta_k = \langle f, e_k \rangle$, $\langle g, h \rangle := \int g \cdot h \, dP$ for $g, h \in L^2(X)$, and $\|f\| = \sum_k \beta_k^2 < \infty$.

When f is estimated using κ terms, the deterministic part of the error is $\text{err}_\kappa = \|f - \sum_{k \leq \kappa} \beta_k e_k\| = \left(\sum_{k > \kappa} \beta_k^2\right)^{\frac{1}{2}}$. Because $\sum_k \beta_k^2 < \infty$, $\text{err}_\kappa \downarrow 0$ as $\kappa \uparrow \infty$. If the β_k are consistently estimated by $\widehat{\beta}_k$, say by OLS with $\kappa_n \to \infty$ sufficiently slowly as the number of data points, n, increases, then the sequence of estimators of the form $\sum_{k \leq \kappa_n} \widehat{\beta}_k e_k$ are consistent for f.

There are two parts to the error made in approximating f, the deterministic part, err_{κ_n}, and the stochastic part due to the randomness in the estimates of the β_k, $\| \sum_{k \leq \kappa_n} (\widehat{\beta}_k - \beta_k) e_k \|$. The behavior of the stochastic part of the error is both well-understood and ineluctable.

There are results giving $\mathcal{O}(r_n)$ bounds on err_{κ_n}, the deterministic part of the error. Typically, the bounds hold for all f satisfying smoothness conditions, often bounds on the derivatives or the norm of the derivatives of f. Rather than trying to work with the variety of regularity conditions, this paper studies the set of all functions that are $\mathcal{O}(r_n)$ approximable.

Fix a sequence $r_n \to 0$. The set $C_n = \mathbf{span}\{e_1, \ldots, e_{\kappa_n}\}$ contains all the functions exactly representable when κ_n terms are used. The set $A_n^M := C_n + B(0, M \cdot r_n)$ contains all points at distance $M \cdot r_n$ or less from C_n. The set $\cup_{M\in\mathbb{N}}[A_n^M \text{ a.a.}]$ is the set of all functions for which the deterministic error is $\mathcal{O}(r_n)$. This set is shy.

Taking a countable union over $M \in \mathbb{N}$, it is sufficient to show that $[A_n^M \text{ a.a.}] = [C_n + B(0, M \cdot r_n) \text{ a.a.}]$ is shy. For this it is in turn sufficient to show that $[A_n^M \text{ i.o.}]$ is shy. The set C_n is a κ_n-dimensional subspace of $L^2(X)$. As a proportion of the unit ball in an ℓ dimensional subspace of $L^2(X)$ containing C_n, $C_n + B(0, \epsilon)$ is on the order of $\epsilon^{\ell-\kappa_n}$. When $\ell - \kappa_n$ is large, one suspects that $C_n + B(0, M \cdot r_n)$ "small" in $L^2(X)$, and this is the direct implication of Lemma 2.2.

These arguments remain valid when $L^2(X)$ is replaced by commonly used subspaces of smooth functions satisfying two requirements. First, the subspace must be complete so that consistency arguments work. Second, for every n and every $\ell > \kappa_n$, there must be an ℓ-dimensional subspace containing C_n. These conditions are satisfied by e.g. the Sobolev spaces of functions having integral bounds on the derivatives.

4.b. **Compactly generated two-way cones.**

Definition 4.1. *A set $C \subset \mathfrak{X}$ is a* **two-way cone** *if $C = \mathbb{R} \cdot C$, that is, if $x \in C$ implies that $r \cdot x \in C$ for all $r \in \mathbb{R}$. A two-way cone is* **compactly generated** *if there exists a compact $K \subset \partial U$, such that $C = \mathbb{R} \cdot K$.*

Define $\varphi : \mathfrak{X} \to \partial U \cup \{0\}$ by $\varphi(0) = 0$ and $\varphi(x) = x/\|x\|$ for $x \neq 0$. For any $E \subset \mathfrak{X}$, $\mathbb{R} \cdot E = \mathbb{R} \cdot \varphi(E)$. If K' is compact and $0 \notin K'$, then $K := \varphi(K')$ is a compact subset of ∂U, and $\mathbb{R} \cdot K' = \mathbb{R} \cdot K$. Thus, a two-way cone is compactly generated iff it is of the form $\mathbb{R} \cdot K'$ for some compact K' not containing 0.

Lemma 4.1. *Any compactly generated two-way cone is closed and has no interior. A two-way cone C is compactly generated iff $C \cap F$ is compact for every norm bounded, closed F.*

The role of K' not containing 0 is seen in the following.

Example 4.1. *If x_n is a countable dense subset of ∂U and K' is the closure of $\{x_n/n : n \in \mathbb{N}\}$, then K' is compact and $0 \in K'$. The two-way cone $\mathbb{R} \cdot K'$ is <u>not</u>*

compactly generated, not closed, and is dense, so that $\mathbb{R} \cdot K' + B(0, \epsilon) = \mathfrak{X}$ *for any* $\epsilon > 0$.

The genericity result is

Theorem 4.1. *For any sequence* C_n *of compactly generated two-way cones and any* $r_n \to 0$, $[A_n$ i.o.$]$ *is shy where* $A_n = C_n + B(0, r_n)$.

4.c. **Estimators as compactly generated two-way cones.** Most of the common nonparametric regression estimators pick points in a sequence of compactly generated two-way cones. The series estimators, Fourier series, wavelets, splines, and the various polynomial schemes, are the easiest, and are treated first. After this, the arguments are presented for kernel estimators and for other locally weighted regression schemes on compact domains, then for two broad classes of artificial neural networks.

4.c.1. *Series estimators.* Fourier series, wavelets, splines, and the various polynomial schemes specify a countable set $E = \{e_k : k \in \mathbb{N}\} \subset \partial U$ with the property that $\overline{\mathbf{span}}\, E = \mathfrak{X}$. (Descriptions of Fourier series the various polynomial schemes as linear subspaces are widely available in textbooks on functional analysis. For wavelets, see Debnath (2002), for splines see Eubank (1999).) The estimator based on n data points, $\hat{f}_n$, is a function of the form $\hat{f}(x) = \sum_{k \le \kappa_n} \hat{\beta}_k e_k(x)$ where $\kappa_n \uparrow \infty$.

These estimators belong to C_n, the span of $E_n = \{e_1, \ldots, e_{\kappa_n}\}$. Since C_n is a finite dimensional subspace of $\mathfrak{X}$, it is a cone. The set $K_n = C_n \cap \partial U$ is compact (being a closed and bounded subset of a finite dimensional subspace of $\mathfrak{X}$), and $C_n = \mathbb{R} \cdot K_n$.

4.c.2. *Kernel and locally weighted regression estimators.* Kernel estimators for functions on a compact domain typically begin with a function K supported on $[-1, +1]$ having its maximum at 0 and satisfying $\int_{-1}^{+1} K(u)\, du = 1$, $\int_{-1}^{+1} uK(u)\, du = 0$, and $\int_{-1}^{+1} u^2 K(u)\, du \neq 0$. Univariate kernel regression functions are (often) of the form

$$\hat{f}_n(x) = \sum_{i=1}^{n} \beta_i g_i(x \big| X_i, \lambda_n) = \sum_{i=1}^{n} \beta_i K((x - X_i)/\lambda_n)$$

where $\lambda_n \downarrow 0$ and the X_i are points in the compact domain, $K \subset \mathbb{R}$. Multivariate kernel regression functions are (often) of the form

$$\hat{f}_n(x) = \sum_{i=1}^{n} \beta_i g_i(x|X_i, \lambda_n) = \sum_{i=1}^{n} \beta_i K(\|x - X_i\|/\lambda_n)$$

where $\lambda_n \downarrow 0$ and the X_i are points in the compact domain $K \subset \mathbb{R}^\ell$. Locally weighted regressions have different functions $g_i(\cdot|\theta_{i,n})$, see e.g. Cleveland and Devlin (1988). In all of these cases, when the domain is compact, so are the sets of possible parameters for the functions g_i, and the mapping from parameters to functions is continuous. This implies that the $\hat{f}_n$ belong to the span of a compact set not containing 0.

4.c.3. *Artificial neural networks.* Single hidden layer feedforward (SLFF) network estimators often take E to be a set of the form $E = \{x \mapsto g(\gamma'\tilde{x}) : \gamma \in \Gamma\}$, $x \in \mathbb{R}^\ell$, $\tilde{x}^T = (1, x^T)^T$, Γ a compact subset of $\mathbb{R}^{n+1}$ with non-empty interior. The estimators are functions of the form $\hat{f}(x) = \sum_{k \leq \kappa_n} \hat{\beta}_k g(\gamma_k'\tilde{x})$, where the γ_k belongs to Γ. When g is continuous, E is compact, and for any of the common choices in the literature, $0 \notin E$. When using the L^p norms, g need not be continuous for E to be compact. Consistency is guaranteed in a number of contexts when g is non-polynomial and analytic (Stinchcombe and White (1992, 1998), Stinchcombe (1999)). With almost no changes to the above analysis, multiple hidden layer feedforward networks output functions are also expressible as the elements of the span of a compact set E. Consistency issues for multiple layer feedforward networks are addressed in Hornik, Stinchcombe, and White (1989, 1990)

Radial basis network estimators often take E_n to be a of the form $E_n = \{x \mapsto g((x-c)'\Sigma(x-c)/\lambda_n) : c \in C, \lambda_n \geq \underline{\lambda}_n\}$, C a compact subset of $\mathbb{R}^\ell$ containing the domain, Σ a fixed positive definite matrix, $\underline{\lambda}_n \downarrow 0$, g a continuous function. The continuity of g implies that the E_n have compact closure. For the common choices of g in the literature, $g(0) \neq 0$ so that $0 \notin E_n$. For the consistency properties of radial/elliptical basis neural networks, see Park and Sandberg (1991, 1993a, b).

4.c.4. *Removing the compact domain assumptions.* The analyses of kernels, locally weighted regressions, and artificial neural networks used the assumption that the domains, D, are compact. Replacing the domains with an increasing sequence D_n such that the probability that the data is in D_n a.a. removes this

restriction. This adds a "with probability 1" qualification to the result that only a shy set of functions are $\mathcal{O}(r_n)$ approximable.

4.d. **Interpretational issues.** For a nonparametric technique X, a rate of convergence result is of the form "Technique X can approximate any f in the set F at a rate $\mathcal{O}(r_n)$." Theorem 4.1 implies that the set F that appear in such a convergence result must be shy for any non-parametric regression scheme representable as the countable union of compactly generated cones. However, it is still possible that the reason that the set F must be shy is because the question is being asked in "too large" a space $\mathfrak{X}$. The cost of shrinking the space $\mathfrak{X}$ so that the shyness result disappears is the loss of the strength in the consistency results used to justify non-parametric regression analysis.

Theorem 4.1 applies when $\mathfrak{X}$ is any separable Banach space. In particular, imposing smoothness conditions by assuming that the target belongs to a Sobolev space or to a space of m times continuously differentiable functions, $C^m(X)$, $m \in \mathbb{N}$, X having a smooth boundary, does not change the conclusion. So, if it is the case that $\mathfrak{X}$ is "too large," it is either dimensionality or completeness that is to blame.

Dimensionality: If the space $\mathfrak{X}'$ of all possible target functions is finite dimensional with a known basis, there is no reason to use non-parametric techniques. If the finitely many basis functions are unknown, but belong to to some infinite dimensional Banach space $\mathfrak{X}$, we are back in essentially the same situation — convergence at any rate $\mathcal{O}(r_n)$ can only be had if the basis functions belong to a shy set.

Completeness: The other possibility is that the space $\mathfrak{X}'$ of all possible target functions is incomplete. The largest set of functions approximable at the rate $\mathcal{O}(r_n)$ by a nested sequence C_n of compactly generated two-way cones is the incomplete set $\mathfrak{X}' = \bigcup_M [C_n + B(0, M \cdot r_n) \text{ a.a.}]$.

Lemma 4.2. *If $C_n = \sum_{k=1}^{n} \mathbb{R} \cdot E(n)$, $E(n)$ a nested, increasing sequence of compact subsets of ∂U, and $\overline{\textbf{span}}\{E(n) : n \in \mathbb{N}\} = \mathfrak{X}$, then $\mathfrak{X}' = \bigcup_M [C_n + B(0, M \cdot r_n) \text{ a.a.}]$ is incomplete and contains a dense, linear subspace.*

There is more direct intuition about the shape of $\mathfrak{X}'$ in the case of Fourier series approximations. Here, $\mathfrak{X} = L^2$, $C_n = \mathbf{span}\{e_1, \ldots, e_n\}$, the e_k are an orthonormal spanning set, and $\mathfrak{X}' = \left\{ f \in L^2 : (\exists N)(\exists M) \left[\left(\sum_{n \geq N} |\langle f, e_n \rangle|^2 \right)^{\frac{1}{2}} \leq M \cdot r_n \right] \right\}$. This $\mathfrak{X}'$ contains the dense, convex set of f with only finitely many non-zero Fourier coefficients.

The incompleteness of $\mathfrak{X}'$ means that it is possible to arrive at a sequence of estimated $\hat{f}_n$ such that $\lim_{m,n\to\infty} \|\hat{f}_m - \hat{f}_n\| = 0$, yet $\hat{f}_n$ is not converging to anything in $\mathfrak{X}'$. Therefore, consistency proofs can only hold by assumption, by insisting that the target function is always in $\mathfrak{X}'$. In other words, the cost of removing the shyness of the set of targets approximable at rate $\mathcal{O}(r_n)$ is the loss of strength in the consistency results.

5. Nonparametric Instrumental Variable Regressions

Interest centers on nonparametrically estimating the functional relation between a random variable Y and a random vector $Z \in \mathbb{R}^p$ *after* conditioning on the random vector $W \in \mathbb{R}^q$ (the instruments). Following Ai and Chen (1999), Darolles *et. al.* (2001), or Newey and Powell (2002), an **instrumental regression for** Y is a function $\varphi(\cdot)$ such that

$$E\left(Y - \varphi(Z) \middle| W\right) = 0. \tag{5}$$

The genericity of the existence of a solution(s) to (5), and the generic properties of the proposed estimators are under study.

The maintained assumption is that all random variables are defined on a probability space $(\Omega, \mathcal{F}, P)$ with $\mathcal{F}$ countably generated and P non-atomic, and that $Y \in L^2(\mathcal{F}) = L^2(\Omega, \mathcal{F}, P)$.

5.a. **Existence and genericity.** For measurable R, let $L^2(R) = L^2(\sigma(R))$ and let π_R denote the projection of $L^2(\mathcal{F})$ onto $L^2(R)$ so that $\pi_R(X) = E\left(X \middle| R\right)$ for $X \in L^2(\mathcal{F})$. Note that projection is a continuous operator having operator norm $\|\pi_R\| := \sup\{\|\pi_R(f)\| : f \in \overline{U}\} = 1$.

Since $L^2(Z) = \{f(Z) : f \text{ is measurable and } \int [f(Z(\omega))]^2 \, dP(\omega) < \infty\}$, there exists an instrumental regression iff $\pi_W(Y) \in \pi_W(L^2(Z))$.

Theorem 5.1. *If $\pi_W(L^2(Z)) = L^2(W)$, then there exists an instrumental regression for all Y in L^2, if $\pi_W(L^2(Z)) \subsetneq L^2(W)$, then there exists an instrumental*

regression only for Y belonging to a shy subset of L^2, if $\overline{\pi_W(L^2(Z))} \subsetneq L^2(W)$, then there exists an instrumental regression only for Y belonging to a proper closed subspace of L^2.

The more difficult questions concern the genericity of the set of (W, Z) pairs such that $\pi_W(L^2(Z)) = L^2(W)$, or, for fixed W the genericity of the set of instruments Z such that $\pi_W(L^2(Z)) = L^2(W)$. If $Z = W$ or if the vector Z contains W as a subvector, then $\pi_W(L^2(Z)) = L^2(W)$, but this case is hardly interesting for instrumental variable regressions.

Lemma 5.1. *The set $R = \{X \in L^2(\mathcal{F}) : \sigma(X) = \mathcal{F}\}$ is Baire large.*

For all (W, Z) pairs in the Baire large set $R \times R$, $\sigma(W) = \sigma(Z) = \mathcal{F}$ so that $L^2(W) = L^2(Z) = L^2(\mathcal{F})$. An immediate Corollary is the existence of a function φ such that $Y = \varphi(W)$. In other words, using Baire largeness gives the rather odd conclusion that statistics is, generically, about the recovery of deterministic relations between observables. I conjecture that the set R in Lemma 5.1 and the set of (W, Z) pairs such that $\pi_W(L^2(Z)) = L^2(W)$ are both shy.[1]

5.b. **Generic properties of estimators.** A maintained assumption in the estimators of φ in nonparametric instrumental regression is that π_W is a compact operator from $L^2(Z)$ to $L^2(W)$ (e.g. Chen and Shen (1998), Ai and Chen (1999), Darolles, Florens, and Renault (2001), Newey and Powell (2002)). Use of instrumental variables that contain information independent of the regressors is known to be bad statistical practice.

Lemma 5.2. *If π_W is a compact operator from $L^2(Z)$ to $L^2(W)$ and W does not have finite range, then $\pi_W(L^2(Z)) \subsetneq L^2(W)$. If some non-constant function of W is independent of Z, then $\overline{\pi_W(L^2(Z))} \subsetneq L^2(W)$.*

Thus, consistency and rate results obtained using compact projection operators apply only Y in a shy subset of $L^2(\mathcal{F})$, and the use of bad instruments leads to non-existence of instrumental regressions unless Y belongs to a closed linear subspace with no interior.

[1]This problem is geometrically miserable because $\sigma(r \cdot X) = \sigma(X)$ for all $r \neq 0$. Therefore, the set of X such that $d_C(\sigma(X), \mathcal{F}) = c$ is a cone without the origin where $d_C(\cdot, \cdot)$ is *any* metric on sub-σ-fields. There seems to be no clear relation between the vector space structure of $L^2(\mathcal{F})$ and the properties of the sub-σ-fields by the elements of $L^2(\mathcal{F})$.

Suppose that no non-constant function of the regressors is independent of the instruments. Even if one assumes, in order to have an easier estimation strategy, that the Y belongs to a shy set, one should examine the primitive assumptions used to guarantee that the projection be a compact operator. Of particular interest is the role of an informational "continuity" assumption.

The variables (W, Z) may have components in common, let $\widetilde{X} = (\widetilde{X})_{i=1}^{r} \in \mathbb{R}^r$ denote the random vector of components in (W, Z) after duplicates have been removed. Let $\widetilde{Q}$ (resp. $\widetilde{Q}_i$) denote the distribution (resp. the i'th marginal distribution) of $\widetilde{X}$.

Definition 5.1. *The information in $\widetilde{Q}$ is* **continuous** *if $\times_i \widetilde{Q}_i \succ \widetilde{Q}$, that is, if there exists a measurable $f \geq 0$ such that, for all $E \in \mathcal{B}^r$, $\widetilde{Q}(E) = \int_E f \, d \times_i \widetilde{Q}_i$.*

Generically, information is dis-continuous, and this is equivalent to the existence of a subtle kind of commonality among the components of $\widetilde{X}$.

Let $\Delta(\mathbb{R}^r)$ denote the set of distributions on $\mathcal{B}^r$, the Borel σ-field on $\mathbb{R}^r$, and for $I \subset \{1, \ldots, r\}$, let $\mathcal{B}^I$ denote the product σ-field generated by the i'th marginal sub-σ-fields, $i \in I$.

Definition 5.2. *A set $B \in \mathcal{B}^r$ is $\widetilde{Q}$-smooth if $\widetilde{Q}(B) > 0$ and $\widetilde{Q}_B(\cdot) := \widetilde{Q}(\cdot|B)$ is non-atomic. A $\widetilde{Q}$-smooth $B \in \mathcal{B}^r$ has* **measurable commonality** *if for non-empty, disjoint $I, J \subset \{1, \ldots, r\}$, there exists a measurable $\psi : B \to (0, 1]$ such that $g_I := E^{\widetilde{Q}_B}(\psi|\mathcal{B}^I) = g_J := E^{\widetilde{Q}_B}(\psi|\mathcal{B}^J)$ $\widetilde{Q}_B$-a.e., and $g_I(\widetilde{Q}_B) = g_J(\widetilde{Q}_B)$ is non-atomic. The probability $\widetilde{Q}$ has* **measurable commonality** *if there exists a $\widetilde{Q}$-smooth $B \in \mathcal{B}$ that has measurable commonality.*

Theorem 5.2. *A prevalent subset of $\widetilde{Q}$ have dis-continuous information and measurable commonality.*

Measurable commonality means that there is some event B and some function ψ on B such $\psi(\widetilde{Q}(\cdot|B))$ is non-atomic, and disjoint sets of components of the vector $\widetilde{X}$, I, J, such that conditioning on B and either of those sets of components perfectly reveals the value of ψ. Given the generic equivalence of measurable commonality and dis-continuous information, it is no surprise that the simplest kind of measurable commonality is a diagonal concentration.

Example 5.1. *With s and t independent and uniformly distributed on $(0, 1]$, take $\widetilde{Q}$ to be the distribution of $(\widetilde{X}_1, \widetilde{X}_2) = (s, s)$ if $s \in (0, \frac{1}{2}]$ and (s, t) if $s \in (\frac{1}{2}, 1]$.*

Let B be the event $\{(s,s) \in (0,\frac{1}{2}] \times (0,\frac{1}{2}]\}$, and define $\psi = s$ on the set B to see that $\widetilde{Q}$ has measurable commonality.

Stinchcombe (2002) provides more examples and an examination the role of measurable commonality as conditional common knowledge in game theoretic models.

5.c. **Interpretational issues.** It is possible that the shyness results for Y are appearing because $L^2(\mathcal{F})$ is "too large." When the instruments are badly chosen, this is clearly not the case. However, when the (W,Z) are such that the projection operator is compact but has dense range, the set of Y for which an instrumental regression exists is shy, but dense. Essentially the same arguments as appeared in §4.d show that restricting attention to this dense set by assumption eviscerates the strength of the consistency and rate of approximation results.

Finite dimensional questions are not interesting for nonparametric regression. The situation is quite different for nonparametric instrumental variable regressions. The subspaces $L^2(W)$ and $L^2(Z)$ are finite dimensional iff W and Z have finite range. Let $\mathbb{P}(W)$ and $\mathbb{P}(Z)$ denote partitions of Ω generated by $\sigma(W)$ and $\sigma(Z)$ respectively. There are two observations:

1. An instrumental regression exists iff the columms of the $\#\mathbb{P}(Z) \times \#\mathbb{P}(W)$ matrix $M = (P(A\big|B))_{A\in\mathbb{P}(Z),\, B\in\mathbb{P}(W)}$ are linearly independent (this is an implication of Theorem 5.1, but much more direct proofs can be found in Newey and Powell (2002), Darolles *et. al.* (2001)). A necessary condition for this is that Z take on at least as many values as W.
2. The existence of a non-constant function of W that is independent of Z corresponds to the existence of at least 2 columns of non-zero constants in M (a failure of the linear independence condition).

6. The Generic Consistency of ICM Tests

Under study are the generic properties of integrated conditional moment (ICM) tests of regression specification, and ICM tests of conditional and unconditional distributional specifications. ICM tests can be understood as estimators of the continuous seminorm of a function ε, typically a residual. If the estimated value of the seminorm is too large, the null hypothesis of correct specification is rejected.

Continuous seminorms can be identified with their compact, absolutely convex polars (defined below). The genericity analysis of ICM tests is carried out in the vector space of compact, absolutely convex sets. The set of consistent ICM tests is both prevalent and Baire large. Within the class of ICM tests that are finitely parametrized, consistency is Baire large, and is conjectured to be shy.

6.a. **ICM's as seminorms with compact polars.** For this section, $\mathfrak{X}$ is a separable Banach space with a separable dual, $\mathfrak{X}^*$, e.g. the separable L^p spaces or Sobolev spaces $S_m^p(\mathbb{R}^r, \mu)$, $p \in (1, \infty)$. The canonical bilinear form on $\mathfrak{X} \times \mathfrak{X}^*$ is $\langle \cdot, \cdot \rangle$.

The null hypothesis of interest is that a function ε, usually regression residuals as a function of the regressors, is equal to 0. The null hypothesis is rejected if the sample version of $T(\varepsilon)$ is too large for a seminorm $T(\cdot)$.

Definition 6.1. *A continuous function $T : \mathfrak{X} \to \mathbb{R}_+$ is a* **continuous seminorm** *(or just* **seminorm** *here) if*

1. $T(\alpha x) = |\alpha| T(x)$ *for every* $\alpha \in \mathbb{R}$ *and* $x \in \mathfrak{X}$*, and*
2. $T(x + y) \leq T(x) + T(y)$ *for every* $x, y \in \mathfrak{X}$*.*

ICM continuous seminorms are often of the form $T_{\mu,r}(\varepsilon) = \left[\int_C |\langle \varepsilon, g \rangle|^r \, d\mu(g)\right]^{1/r}$ for a probability μ satisfying $\mu(C) = 1$ for some compact $C \subset \mathfrak{X}^*$ and some $r \in [1, \infty]$.

For any (continuous) seminorm T, whether or not it is of the form $T_{\mu,r}$, the set $D = \{x \in \mathfrak{X} : T(x) \leq 1\}$ absolutely convex and absorbent. (The absolutely convex hull of a set D is denoted $\mathsf{aco}\,(D)$ is defined as the convex hull of $D \cup -D$. A set D is **absolutely convex** if $D = \mathsf{aco}\,(D)$. A set D is **absorbent** if for every $x \in \mathfrak{X}$, there exists an $\alpha > 0$ such that $x \in \alpha D$.) A given T can be identified with D in the sense that for every $x \in \mathfrak{X}$, $T(x) = \inf\{\alpha > 0 : x \in \alpha D\}$ (Robertson and Robertson, Proposition I.4.7, p. 14). Any closed absolutely convex D can in turn be identified with its polar $D^\circ := \{x^* \in \mathfrak{X}^* : |\langle D, x^* \rangle| \leq 1\}$ in the sense that $D = (D^\circ)^\circ$. Because $D = \{T \leq 1\}$ is absorbent, D° is compact (Robertson and Robertson Proposition II.4.9 *et seq.*). Directly from the definitions, T and D° are related by $T(x) = \sup\{|\langle x, x^* \rangle| : x^* \in D^\circ\}$ for every $x \in \mathfrak{X}$.

For a compact, absolutely convex $E \subset \mathfrak{X}^*$, $T_E(x) := \sup\{|\langle x, x^* \rangle| : x^* \in E\}$. defines a continuous seminorm. The previous paragraph can be summarized as

saying "T is a continuous seminorm iff it is of the form T_{D° with D° being compact and absolutely convex."

For the ICM seminorms $T_{\mu,r}$, we have

Lemma 6.1. *If C is a compact subset of $\mathfrak{X}^*$, and μ is a probability measure on $\mathfrak{X}^*$ with $\mu(C) = 1$, then for each $r \in [1, \infty]$, there exists an absolutely convex subset $E_r \subset \mathsf{aco}\,(C)$ such that for all $x \in \mathfrak{X}$, $T_{E_r}(x) = T_{\mu,r}(x)$.*

The class of compact absolutely convex subsets of $\mathfrak{X}^*$ is denoted $\mathbb{K}$. Addition and scalar multiplication of elements of $\mathbb{K}$ are continuous in the Hausdorff metric, d_H, and $(\mathbb{K}, d_H)$ is a complete, separable locally convex topological vector space.

Definition 6.2. *A test $E \in \mathbb{K}$ is* **consistent** *if $\bigcap_{r>0}\{x \in \mathfrak{X} : T_E(x) < r\} = \{0\}$.*

The equivalence of the following comes directly from standard concepts in the study of topological vector spaces: E is consistent; the seminorm topology generated by E is Hausdorff; the span of E is $\sigma(\mathfrak{X}^*, \mathfrak{X})$ dense in $\mathfrak{X}^*$ (see Stinchcombe and White (1998) for details).

Generically, ICM tests are consistent.

Theorem 6.1. *The set of consistent $E \in \mathbb{K}$ is prevalent and Baire large.*

6.b. **Examples.** Many tests of the correctness of parametric specification of regression models, most tests of the equality of unconditional distributions and of the equality of parametrized conditional distributions can be formulated as estimators of a T_E, $E \in \mathbb{K}$.

6.b.1. *Tests of parametric regression models.* A parametric model for a conditional mean is a collection of functions on $\mathbb{R}^\ell$, $\{x \mapsto f(x,\theta) : \theta \in \Theta\}$, $x \in \mathbb{R}^\ell$, $\Theta \subset \mathbb{R}^k$. The null hypothesis that the model is correctly specified for $E(Y\big|X)$ is

$$H_0 : (\exists \theta_0 \in \Theta)[u = Y - f(X,\theta_0), \text{ and } \varepsilon = E(u\big|X) = 0)]. \tag{6}$$

Pick $p, q \in (1, \infty)$ such that $\frac{1}{p} + \frac{1}{q} = 1$. The set of deviations, ε, from the null is $L^p(X) := L^p(\Omega, \sigma(X), P)$ with $\sigma(X)$ being the minimal σ-field making X measurable. Specifically, the set of possible deviations from H_0 is

$$\{\varepsilon \in L^p(X) : (\exists Y \in L^p(\Omega, \mathcal{F}, P))[\varepsilon \in E(Y - f(X,\Theta)\big|X)]\}. \tag{7}$$

The set of test functions is $L^q(X)$. The literature has used sample versions of T_E for many $E \in \mathbb{K}$.

1. If $E = [-f, f] = \mathsf{aco}(\{f\})$, f a function in $L^q(X)$, Hausmann (1978), Newey (1985), Tauchen (1985), White (1987), White (1994). Such tests are not consistent, having no power against anything in the "orthogonal" complement of E.
2. For X taking values in compact subsets of $\mathbb{R}^\ell$, $\ell \geq 1$, Bierens (1990) analyzes $E = \mathsf{aco}(\{x \mapsto \exp(x'\tau) : \tau \in T\})$, T a compact cube in $\mathbb{R}^\ell$.
3. Again for X taking values in compact subsets of $\mathbb{R}^\ell$, Stinchcombe and White (1998) analyzes $E = \mathsf{aco}(\{x \mapsto f(\tilde{x}'\tau) : \tau \in T\})$, T a compact subset of $\mathbb{R}^{\ell+1}$ with non-empty interior, where $\tilde{x} = (1, x')' \in \mathbb{R}^{\ell+1}$, and f is any non-polynomial analytic function.
4. For $X \in \mathbb{R}^1$, Stute (1997) analyzes $E = \mathsf{aco}(\{x \mapsto 1_{(-\infty,a]}(x) : a \in \mathbb{R}\})$.

By Lemma 6.1, the class of T_E also contains ICM tests of Bierens (1982), Bierens (1990), White (1989), Stinchcombe and White (1998), and Bierens and Ploberger (1997). Theorem 6.1 shows that, in the class of all ICM tests for correct specification, the consistent ones are generic.

6.b.2. *Tests of the equality of unconditional distributions.* The null hypothesis for the Kolmogorov-Smirnov test is the the equality of two distributions on $[0, 1]$ with Lebesgue densities f and h. The densities f and h are elements of $\mathfrak{X} = L^p([0, 1], \mathcal{B}, \lambda)$, $p \in (1, \infty)$, $\mathcal{B}$ being the usual Borel sets λ being Lebesgue measure. The null hypothesis is that $\varepsilon := f - h = 0$ where $f, h \in \mathcal{C}$, $\mathcal{C}$ the convex, norm closed set of densities. The set of deviations from the null hypothesis is the set $\mathcal{C} - \mathcal{C}$. The test functions belong to $L^q([0, 1], \mathcal{B}, \lambda)$, $\frac{1}{p} + \frac{1}{q} = 1$.

When E is the compact subset of L^q, $\mathsf{aco}(\{1_{[0,a]}(\cdot) : a \in (0, 1)\})$, the sample version of $T_E(\varepsilon) = \sup_{x^* \in E} |\langle \varepsilon, x^* \rangle|$ is the Kolmogorov-Smirnov statistic. If F^n and H^n are empirical cdf's from two iid. samples drawn with densities f and h, the sample version of the statistic is $\sup_{a \in (0,1)} |F^n(a) - H^n(a)|$.

The Cramer-von Mises test for $\varepsilon = 0$ is the the sample version of

$$T_{\mu,2}(\varepsilon) = \left[\int_C |\langle \varepsilon, x^* \rangle|^2 \, d\mu(x^*) \right]^{\frac{1}{2}} \tag{8}$$

where μ is the compactly supported distribution on $\mathfrak{X}^*$ induced by picking an $a \in [0, 1]$ according to the uniform distribution and then integrating ε against $x^* = 1_{[0,a]}$. Note that C is a compact subset of L^q, and by Lemma 6.1, there is a $E \subset \mathsf{aco}(C)$ such that $T_E(\varepsilon) = T_{\mu,2}(\varepsilon)$ for all $\varepsilon \in L^p$.

Theorem 6.1 implies that for generic E, the test based on sample versions of T_E is consistent.

6.b.3. *Tests for parametric specifications of conditional distributions.* This follows Andrews (1997) and Stinchcombe and White (1998). Let the conditional likelihood function with respect to a σ-finite measure ν and parameterized by $\theta \in \Theta$, Θ an open subset of $\mathbb{R}^p$, be given by a measurable function, $f : \mathbb{R} \times \mathbb{R}^k \times \Theta \to \mathbb{R}_+$, with the property that for all $x \in \mathbb{R}^k$ and $\theta \in \Theta$, $\int_{\mathbb{R}} f(y|x,\theta) d\nu(y) = 1$. Let $\mathcal{S}$ denote $\{f(\cdot|\cdot,\theta) : \theta \in \Theta\}$. $\mathcal{S}$ is **correctly specified for Y conditional on X** when for P_X-almost all x, $f(\cdot|x,\theta_0)$ is a version of the true conditional density of Y given $X = x$ with respect to ν for some θ_0 in Θ.

For $\theta \in \Theta$, define Q_θ to be the distribution on $\mathbb{R}^1 \times \mathbb{R}^\ell$,

$$Q_\theta(A) = \int_A f(y|x,\theta)\, d\nu(y) dP_X(x), \ A \text{ a Borel subset of } \mathbb{R} \times \mathbb{R}^k. \tag{9}$$

With P denoting the true distribution of (Y, X), the null hypothesis is that $0 \in P - \{Q_\theta : \theta \in \Theta\}$.

Assume that P has a density with respect to $\nu \times P_X$. This is not innocuous, but it is a minimal assumption needed for entertaining the possibility that the model is correctly specified. A space containing the densities of P and the Q_θ is

$$\mathfrak{X} = L^p(\mathbb{R}^{1+\ell}, \mathcal{B}^{1+\ell}, \nu \times P_X), \tag{10}$$

$p \in (1, \infty)$. The test functions belong to the dual space $\mathfrak{X}^* = L^q$. Stinchcombe and White (1998) used $E = \mathsf{aco}\,(\{x \mapsto f(\tilde{x}'\gamma) : \gamma \in \Gamma\})$, Γ a compact subset of $\mathbb{R}^{\ell+1}$ with non-empty interior, f any non-polynomial analytic function. Andrews (1997) uses the L^q compact set $E = \mathsf{aco}\,(\{1_{(-\infty,z]}(\cdot) : z \in \mathcal{Z}\})$ where $\mathcal{Z}$ is a support set for the random vector (Y, X) in $\mathbb{R}^{1+\ell}$.

Theorem 6.1 implies that for generic E, the test based on sample versions of T_E is consistent.

6.c. **Finitely parametrized ICM's.** As seen in the examples, the E that arise in practice are often smoothly parametrized by finite dimensional sets. Such E are the smooth images of a compact Θ.

This section supposes that Θ is an infinite, compact metric space. Let $C(\Theta; \mathfrak{X})$ denote the set of continuous functions from Θ to a separable Banach space $\mathfrak{X}$. A vector subspace $C' \subset C$ is **full** if for all finite collections $\{\theta_n : n \leq N\} \subset \Theta$, and

all collections $\{h_n : n \leq N\} \subset \mathfrak{X}$, there exists an $f \in C'$ such that $f(\theta_n) = h_n$ for $n \leq N$. When Θ is a manifold, the vector subspaces of smooth functions in $C(\Theta; \mathfrak{X})$ are typically full, typically dense, and they are typically G_δ's, hence topologically complete.

Theorem 6.2. *If Θ is a compact metric space, and C' is a dense, full vector subspace that is a G_δ in $C(\Theta; \mathfrak{X})$, then the set $\{f \in C' : \mathsf{aco}\,(f(\Theta))$ is consistent$\}$ is Baire large in C'.*

I conjecture that this result is true with "prevalent" replacing "Baire large," but this is an open (and vexing) problem.

For Θ an infinite compact space, $C' = C(\Theta; \mathfrak{X})$ satisfies the remaining assumptions in Theorem 6.2. This means that for any infinite, compact parameter set, a dense set of continuous parametrizations spans $\mathfrak{X}$. In some directions, this is a generalization of the universal approximation results for neural networks in Cybenko (1989), Funahashi (1989), Hornik, Stinchcombe and White (1989, 1990).

6.d. **Interpretational issues.** It is possible that the conclusion that the set of consistent ICM's being "large" is due to the problem being imbedded in "too small" a set of ICM tests. However, it seems quite difficult to imagine a larger space of non-adaptive tests than the space $\mathbb{K}$ considered here.

7. Concluding Remarks

A genericity analysis becomes nonsensical if the wrong setting is chosen — if the statistically relevant cases are two dimensional, then a three dimensional genericity analysis hides more than it reveals. The challenge then, is to understand the results found here in this light.

Bayesian updating: The generic inconsistency of Bayes updating in plausible infinite contexts is surprising. Small details of the prior turn out to matter quite sharply. It seems to mean that we should calculate out which observation(s) have the most effect on our posterior distribution and ask if we trust those data enough to tolerate them moving the posterior as much as they do. There are a number of possible heuristics for these procedures, but almost certainly there is no best procedure.

Limiting the set of priors can, in many cases, provide computationally tractable, parametrized statistical models. As a general model of optimizing behavior, such

a step is clearly unsatisfactory. Further, even in these parametrized models, Bayesian updating can be inconsistent. These considerations lead to the conclusion that the genericity of inconsistency is not an artifact of the wrong setting being chosen.

Nonparametric regression: Rates of convergence analyses are, hopefully, guides to finite sample behavior. Quick rates of convergence are often achieved at the expense of maintaining smoothness assumptions. It is difficult to give a satisfactory general argument about e.g. the appropriateness of smoothness conditions that improve rates. For example, if the errors in the X's in a non-linear regression are plausibly smoothly distributed, the best possible recoverable mean of the Y conditional on observed X is likely to be a smooth function, whether or not the true conditional mean is. The present work shows that whatever the smoothness assumptions are, they do not cover a generic set of targets. If one is not willing to firmly declare for a non-generic set, then rate of convergence results only hold at the expense of consistency results.

One version of the lessons to be drawn from the genericity analysis of nonparametric regression is that technique without insight will not carry us very far. By being "without insight" into a particular regression function, I mean that "all of $\mathfrak{X}$ must be considered." The results here reinforce the lesson that $\mathfrak{X}$ is a very large place to search.

Instrumental variable regressions: The analysis of nonparametric instrumental regressions reinforces one's convictions that it is a difficult business. The required assumptions are generically not satisfied under the known conditions needed for estimation, and the target regression functions are, generically, not fully recoverable. In particular, there is a clear formulation of functional relations not being recoverable if the instruments are badly chosen.

A further benefit of the present analysis is that it shows that one of the non-generic required assumptions is equivalent to there being no measurable commonality. Measurable commonality is generically equivalent to dis-continuous information. This provides a bit more shape to the question of whether the continuous information assumption is acceptable.

Specification testing: Once Bierens (1982, 1990) gave us the initial insights, consistent ICM tests became relatively easy to write down. The results here suggest that we should not be surprised by this because consistency is a generic

property of ICM tests. This is so pleasant a conclusion that I cannot find it in myself to question whether it arises from wrong-headedness.

8. References

Ai, Chunrong and Xiaohong Chen (1999). "Efficient Estimation of Models with Conditional Moment Restrictions Containing Unknown Functions", working paper, University of Florida and London School of Economics.

Anderson, Robert M. and William R. Zame (2001). "Genericity with Infinitely Many Parameters," *Advances in Theoretical Economics*: Vol. 1: No. 1, Article 1.

Andrews, Donald W. (1997). "A Conditional Kolmogorov Test. *Econometrica*," **65**, 1097-1128.

Arnold, Barry C., Patrick L. Brockett, William Torrez, and A. Larry Wright (1984). "On the Inconsistency of Bayesian Non-Parametric Estimators in Competing Risk/-Multiple Decrement Models," *Insurance, Mathematics and Economics*, **3**, 49-55.

Baire, Rene. (1899). "Sur les Fonctions de Variables Réelles," *Annali di Matematica Pura ed Applicada* Series 3, Vol. 3, 1-122.

Bierens, Hermann B. (1982). "Consistent Model Specification Tests," *Journal of Econometrics*, **20**, 105-134.

Bierens, Hermann B. (1990). "A Consistent Conditional Moment Test of Functional Form," *Econometrica* **58**, 1443-1458.

Bierens, Hermann B. and Werner Ploberger (1997). "Asymptotic Theory of Integrated Conditional Moment Tests." *Econometrica*, **65**, 1129-1152.

Chen, X. and X. Shen (1998). "Sieve Extremum Estimates for Weakly Dependent Data," *Econometrica*, **66**, 289-314.

Christensen, Jens Peter Reus (1974). *Topology and Borel Structure.* Amsterdam: North-Holland Publishing Company.

Cleveland, William and Susan Devlin (1988). "Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting," *Journal of the American Statistical Association*, **83**(403), 596-610.

Cybenko, G. (1989). "Approximation by Superpositions of a Sigmoidal Function." *Mathematics of Control, Signals and Systems* **2**, 303-14.

Cotter, Kevin D. (1986). "Similarity of Information and Behavior with a Pointwise Convergence Topology." *Journal of Mathematical Economics* **15**(1) 25-38.

Darolles, Serge, Jean-Pierre Florens, and Eric Renault (2001). "Nonparametric Instrumental Regression," Manuscript, GREMAQ, University of Toulouse.

Dellacherie, C. and P.-A. Meyer (1978), *Probabilities and Potential*. Amsterdam: North Holland Publishing Co.

Diaconis, Persi and David Freedman (1986a). "On the Consistency of Bayes Estimates," *Annals of Statistics* **14**(1), 1-26.

Diaconis, Persi and David Freedman (1986b). "On Inconsistent Bayes Estimats of Location," *Annals of Statistics* **14**(1), 68-87.

Eubank, Randall L. (1999). *Spline smoothing and nonparametric regression*, 2'nd ed. M. Dekker, New York.

Ferguson, T. (1973). "A Bayesan Analysis of Some Nonparametric Problems," *Annals of Statistics* **1**, 209-230.

Freedman, David (1963). "On the Asymptotic Behavior of Bayes Estimates in the Discrete Case I," *Annals of Mathematical Statistics* **34**, 1386-1403.

Freedman, David (1965). "On the Asymptotic Behavior of Bayes Estimates in the Discrete Case II," *Annals of Mathematical Statistics* **36**, 454-456.

Funahashi, K. (1989). On the Approximate Realization of Continuous Mappings by Neural Networks. *Neural Networks* **2**, 183-92.

Grenander, Ulf (1981). *Abstract Inference*. Wiley, New York.

Halmos, Paul (1957). *Introduction to Hilbert space and the theory of spectral multiplicity*, 2'nd ed. Chelsea Pub. Co., New York.

Hausman, J. (1978). "Specification Tests in Econometrics," *Econometrica* **46**, 1251-1272.

Hornik, Kurt (1991). "Approximation capabilities of multilayer feedforward networks," *Neural Networks* **4**(2), 251-257.

Hornik, K. (1993). "Some New Results on Neural Network Approximation," *Neural Networks* **6**(8), 1069-1072.

Hornik, Kurt, Maxwell Stinchcombe, and Halbert White (1989). "Multi-layer Feedforward Networks are Universal Approximators," *Neural Networks* **2**, 359-366 (1989).

Hornik, Kurt, Maxwell Stinchcombe, and Halbert White (1990). "Universal Approximation of an Unknown Mapping and its Derivatives using Multilayer Feedforward Networks," *Neural Networks* **3**, 551-560 (1990).

Hunt, B. R., T. Sauer, and J. A. Yorke (1992). "Prevalence: A Translation-Invariant 'Almost Every' on Infinite-Dimensional Spaces," *Bulletin (New Series) of the American Mathematical Society* **27**, 217-238.

Debnath, Lokenath (2002). *Wavelet transforms and their applications*. Birkhauser, Boston.

Nachbar, John (1997). "Prediction, Optimization, and Learning in Repeated Games," *Econometrica*, **65**(2), 275-309.

Newey, Whitney (1985). "Maximum Likelihood Specification Testing and Conditional Moment Tests," *Econometrica* **53**, 1047-70.

Newey, Whitney and James Powell (2002). "Instrumental Variable Estimation of Nonparametric Models," forthcoming, *Econometrica*.

Park, J. and I. W. Sandberg (1991). "Universal Approximation Using Radial Basis-Function Networks," *Neural Computation*, **3**(2), 246-257.

Park, J. and I. W. Sandberg (1993a). "Approximation and Radial-Basis Function Networks," *Neural Computation*, **5**(2), 305-316.

Park, J. and I. W. Sandberg (1993b). "Nonlinear Approximations Using Elliptic Basis Function Networks," *Circuits, Systems and Signal Processing*, **13**(1), 99-113.

Robertson, Alex P. and Wendy J. Robertson (1973). *Topological Vector Spaces*. Cambridge University Press, Cambridge.

Rudin, Walter (1973). *Functional Analysis*. McGraw-Hill, New York.

Stinchcombe, Maxwell (1990). "Bayesian Information Topologies," *Journal of Mathematical Economics* **19**, 3, 233-254.

Stinchcombe, Maxwell (1993). "A Further Note on Bayesian Information Topologies," *Journal of Mathematical Economics* **22**, 189-193.

Stinchcombe, Maxwell (1999). "Neural Network Approximation of Continuous Functionals and Continuous Functions on Compactifications," *Neural Networks* **12**, 467-477.

Stinchcombe, Maxwell (2001). "The Gap Between Probability and Prevalence: Loneliness in Vector Spaces," *Proceedings of the American Mathematical Society* **129**, 451-457.

Stinchcombe, Maxwell (2002). Finiticity in Measurable-Continuous Games. Department of Economics, University of Texas at Austin.

Stinchcombe, Maxwell and Halbert White (1992). "Using Feedforward Networks to Distinguish Multivariate Populations," with Halbert White, Proceedings of the International Joint Conference on Neural Networks, IEEE Press, New York, I:788-793.

Stinchcombe, Maxwell and Halbert White (1998). "Consistent Specification Testing with Nuisance Parameters Present Only Under the Alternative," *Econometric Theory* **14**, 295-325.

Stute, W. (1997). Nonparametric Model Checks for Regression. *Annals of Statistics*, Vol. 25, 613-641.

Tauchen, G. (1985). "Diagnostic Testing and Evaluation of Maximum Likelihood Models," *Journal of Econometrics* **30**, 415-444.

White, Halbert (1987). "Specification Testing in Dynamic Models," in T. Bewley, (ed.), *Advances in Econometrics - Fifth World Congress*, Vol 1. New York: Cambridge University Press, pp. 1-58.

White, Halbert (1989). "An Additional Hidden Unit Test for Neglected Nonlinearity in Multilayer Feedforward Networks," *Proceedings of the International Joint Conference on Neural Networks, Washington D.C.*. San Diego: SOS Printing, II:451-455.

White, Halbert (1994). *Estimation, Inference and Specification Analysis*. Cambridge University Press, New York.

9. Proofs

Proof of Lemma 2.1: The equivalence of (1) and (2) is immediate. The following proof of the equivalence of (1) and (3) is directly from Hunt, Sauer, and Yorke (1992), and is reproduced here for completeness.

If $\Lambda^k(S) = 0$, take $\eta = U^k$, the uniform distribution on $[0,1]^k$. If there is a compactly supported η such that $\eta(S+x) \equiv 0$, then $\int_{\mathbb{R}^k} \eta(S+x)\, d\Lambda^k(x) = 0$, so that $\int_{\mathbb{R}^k} \left[\int_{\mathbb{R}^k} 1_{(S+x)}(y)\, d\eta(y)\right]\, d\Lambda^k(x) = 0$, implying $\int_{\mathbb{R}^k} \left[\int_{\mathbb{R}^k} 1_{(S+x)}(y)\, d\Lambda^k(x)\right]\, d\eta(y) = 0$. Since $1_{(S+x)}(y) \equiv 1_{(S-y)}(x)$ and $\int_{\mathbb{R}^k} 1_{(S-y)}(x)\, d\Lambda^k(x) = \Lambda^k(S-y)$, we have $\int_{\mathbb{R}^k} \Lambda^k(S-y)\, d\eta(y) = 0$. Since Λ^k is translation invariant and non-negative, $\Lambda^k(S-y) \equiv \Lambda^k(S) \geq 0$, $\int_{\mathbb{R}^k} \Lambda^k(S)\, d\eta(y) = 0$, implying that $\Lambda^k(S) = 0$. ∎

Proof of Lemma 2.3: If necessary, translate C so that $0 \in \operatorname{rel\,int}_{\overline{\mathbf{co}}\, S}(\overline{\mathbf{co}}\, S)$. For each $n \in \mathbb{N}$, $[n \cdot \overline{\mathbf{co}}\, S] \cap C$ is a closed set with no interior in the csm space C. By Baire's theorem, the complement of $\cup_n [n \cdot \overline{\mathbf{co}}\, S] \cap C$ is residual, hence non-empty. Pick arbitrary v in this residual set, and take η to be the uniform distribution on the line $L_v = [0, v]$.

For any $x \in \mathfrak{X}$, $[L_v + x] \cap \overline{\mathbf{co}}\, S$ is a compact convex interval, possibly empty or degenerate. Suppose, for the purposes of contradiction, that this interval has positive η mass. This means that it is a line segment $[y, y']$, $y \neq y'$. Since $0 \in \operatorname{rel\,int}_{\overline{\mathbf{co}}\, S}(\overline{\mathbf{co}}\, S)$ and $\overline{\mathbf{co}}\, S$ is convex, there exist $k_y, k_{y'} > 0$ such that $-k_y \cdot y, -k_{y'} \cdot y' \in \overline{\mathbf{co}}\, S$. Therefore, the plane spanned by $[L_v + x] \cap \overline{\mathbf{co}}\, S$ and 0 belongs to $\cup_n [n \cdot \overline{\mathbf{co}}\, S] \cap C$. But v also belongs to this set, a contradiction. ∎

Proof of Theorem 3.1: Fix a full support θ and a countable collection of $f_n \in C^{\lambda}_{++}$ such that the $\theta_n := \theta_{f_n}$ are dense in $\Delta(X)$, and $\theta_n \neq \theta$ for all $n \in \mathbb{N}$.

Abuse notation with $f_n(h^t) := \Pi_{i=1}^t f_n(x_i)$ for partial histories $h^t = (x_1, \ldots, x_t)$. Let $U_{n,m}$ be a nested sequence of open neighborhoods of θ_n with the diameter of $U_{n,m}$ less than $1/m$. Let $v_{n,m} : \Delta(X) \to [0,1]$ be a continuous function taking the value 1 on $U_{n,m}$ and 0 on the complement of $U_{n,m-1}$. For each h^t, define the continuous function $m_{n,m}(\cdot, h^t)$ on $\mathbb{M}_B^\lambda$ by

$$m_{n,m}(\mu, h^t) = \left(\int_{C_+^\lambda} v_{n,m}(f) f(h^t)\, d\mu(f)\right) / \left(\int_{C_+^\lambda} f(h^t)\, d\mu(f)\right).$$

This is the expected value of $v_{n,m}$ conditional on h^t when beliefs are μ. If posterior beliefs along a sequence of histories h^t converge to δ_{θ_n}, then $\lim_t m_{n,m}(\mu, h^t) = 1$.

For $\epsilon > 0$ and $t \in \mathbb{N}$, define $S_{n,m}(\epsilon, t) = \{\mu \in \mathbb{M}_B^\lambda : \int_{X^\infty} m_{n,m}(\mu, h^t)\, d\theta^\infty(h^t) \leq \epsilon\}$. By continuity, $S_{n,m}(\epsilon, t)$ is a closed subset of $\mathbb{M}_B^\lambda$.

Outline:

1. For all $\epsilon > 0$ and for all t, $S_{n,m}(\epsilon/2, t) \subset \overline{\mathbf{co}}\, S_{n,m}(\epsilon/2, t) \subset S_{n,m}(\epsilon, t)$. This intermediate result leads to
2. For all T, $\bigcap_{t \geq T} S_{n,m}(\epsilon, t)$ is shy, equivalently, $\bigcup_{t \geq T} S_{n,m}(\epsilon, t)^c$ is prevalent.
3. $\mathrm{err}(\theta) = \bigcap_{\epsilon,n,m,T} \bigcup_{t \geq T} S_{n,m}(\epsilon, t)^c$ (the intersection taken over rational ϵ in $(0,1)$). Since the intersection of countably many prevalent sets is prevalent, this and the second step complete the proof.

Details:

1. For all $\epsilon > 0$ and all t, $S(\epsilon/2, t) \subset \overline{\mathbf{co}}\, S(\epsilon/2, t) \subset S(\epsilon, t)$.

   Since each $S(\epsilon, t)$ is closed, showing $\mathbf{co}\, S(\epsilon/2, t) \subset S(\epsilon, t)$ is sufficient. Pick $\mu, \mu' \in S(\epsilon/2, t)$ and $0 < \alpha < 1$. Let $\mu\alpha\mu' = \alpha\mu + (1-\alpha)\mu'$. What must be shown is $\int m(\mu\alpha\mu', h^t)\, d\theta^\infty(h^t) \leq \epsilon$. For numbers $s, s' > 0$ and $r, r' \geq 0$ $\frac{\alpha r + (1-\alpha) r'}{\alpha s + (1-\alpha) s'} \leq \max\{\frac{r}{s}, \frac{r'}{s'}\} \leq \frac{r}{s} + \frac{r'}{s'}$. This delivers

   $$\begin{aligned}
   &\int m(\mu\alpha\mu', h^t)\, d\theta^\infty(h^t) \\
   &\quad = \int \frac{\int v_{n,m}(f) f(h^t)\, d\mu\alpha\mu'(f)}{\int f(h^t)\, d\mu\alpha\mu'(f)}\, d\theta^\infty(h^t) \\
   &\quad \leq \int \frac{\int v_{n,m}(f) f(h^t)\, d\mu(f)}{\int f(h^t)\, d\mu(f)}\, d\theta^\infty(h^t) + \int \frac{\int v_{n,m}(f) f(h^t)\, d\mu'(f)}{\int f(h^t)\, d\mu'(f)}\, d\theta^\infty(h^t) \\
   &\quad \leq \epsilon/2 + \epsilon/2 = \epsilon,
   \end{aligned}$$

   completing the proof of the first step.
2. For all T, $\bigcap_{t \geq T} S_{n,m}(\epsilon, t)$ is shy.

   We have $\bigcap_{t \geq T} S_{n,m}(\epsilon, t) \subset \bigcap_{t \geq T} \overline{\mathbf{co}}\, S_{n,m}(\epsilon, t) \subset \bigcap_{t \geq T} S_{n,m}(2\epsilon, t)$. by the definition of $\overline{\mathbf{co}}$ and Step 1. The set $\bigcap_{t \geq T} \overline{\mathbf{co}}\, S_{n,m}(\epsilon, t)$ is convex and closed.

Therefore, it is sufficient to show that the closed set $\bigcap_{t\geq T} S_{n,m}(2\epsilon, t)$ has no interior. The proof will be completed by showing the existence of a dense set $\mathbb{D}_n$ with the property that $\mathbb{D}_n \cap \bigcap_{t\geq T} S_{n,m}(\epsilon, t) = \emptyset$. Specifically, the set $\mathbb{D}_n$ will have the property that for every $\mu \in \mathbb{D}_n$, $\mu(\cdot\big|h^t) \to_{w^*} \delta_{\theta_n}$ θ^∞-a.e.

Let $M \subset \mathbb{M}$ denote the set of probabilities for which there exists some non-empty open set in $\Delta(X)$ receiving mass 0. Let $\mathbb{D}_n$ be the set of μ's in $\mathbb{M}_B^\lambda$ putting positive mass on θ_n and on finitely many points in M. $\mathbb{D}_n$ is easily seen to be dense. Since θ is full support, every non-empty open set will be visited infinitely often θ^∞ a.e. Therefore, for every $\mu \in \mathbb{D}_n$, only the full support θ_n can have positive posterior weight in the limit, that is, $\mu(\cdot\big|h^t) \to_{w^*} \delta_{\theta_n}$.

3. $\text{err}(\theta) = \bigcap_{\epsilon,n,m,T} \bigcup_{t\geq T} S_{n,m}(\epsilon, t)^c$ (the intersection taken over rational ϵ in $(0, 1)$).

   Since the θ_n are dense in $\Delta(X)$ and the diameters of the $U_{n,m}$ converge to 0, every non-empty open U contains a $U_{n,m}$. Therefore, $\mu \in \bigcap_{\epsilon,n,m,T} \bigcup_{t\geq T} S_{n,m}(\epsilon, t)^c$ iff for all rational ϵ in $(0, 1)$, all non-empty open U, and all T, there exists a $t \geq T$ such that $\mu \notin S_{n,m}(\epsilon, t)$, that is, iff $\mu \in \text{err}(\theta)$. ∎

Proof of Lemma 4.1: Let $x_n = r_n \cdot k_n$, $k_n \in K'$, be a sequence in the compactly generated two-way cone $C = \mathbb{R} \cdot K'$, K' a compact subset of ∂U, and suppose that $\|x_n - x\| \to 0$. Then $r_n := \|x_n\| \to \|x\|$, and $k_n \to k$ for some $k \in K'$. Therefore $x = \|x\|k \in \mathbb{R} \cdot K'$, so that C is closed.

Suppose that $C = \mathbb{R} \cdot K$ for some compact $K \subset \partial U$. Let F be closed and suppose that $\|F\| \leq B$. Then $C \cap F$ is closed by the first step. Further, $C \cap F \subset [-B, B] \cdot K \cap F$ expresses $C \cap F$ as a closed subset of the compact set $[-B, B] \cdot K$ so that $C \cap F$ is compact.

If $C \cap F$ is compact for every closed, norm bounded F, then $K' := C \cap \partial U$ is compact, and $C = \mathbb{R} \cdot K'$.

Finally, if C has non-empty interior, then $C \cap \overline{U}$ has non-empty interior. But $C \cap \overline{U}$ is compact, and compact subsets of $\mathfrak{X}$ must have empty interior. ∎

Proof of Theorem 4.1: Fix arbitrary $R > 0$. Taking the union over some sequence $R_n \uparrow \infty$ shows that it is sufficient to prove that $(R \cdot U) \cap [A_n \text{ i.o.}]$ is shy.

Fix arbitrary $\epsilon > 0$. $R \cdot U$ is a subset of the closed, norm bounded set $R \cdot (1+\epsilon)\overline{U}$. For each n, let F_n be the compact, hence approximately flat set $C_n \cap (R \cdot (1+\epsilon)\overline{U})$. By Lemma 2.2, $S = [(F_n + B(0, r_n)) \text{ i.o.}]$ is shy. By construction, $[(R \cdot U) \cap [A_n \text{ i.o.}]] \subset S$. ∎

Proof of Lemma 4.2: $\mathfrak{X}'$ contains the dense set $\bigcup_n C_n$. The completion of a dense subset of any complete metric space is the space itself. Therefore, if $\mathfrak{X}'$ was complete, it would be equal to $\mathfrak{X}$, which contradicts Theorem 4.1.

If $f, g \in \bigcup_n C_n$, there there exists an N such that $f, g \in C_N$. Since C_N is a two way cone, for all $\alpha, \beta \in \mathbb{R}$, $\alpha f, \beta g \in C_N = \sum_{k=1}^{N} \mathbb{R} \cdot E_k$. Since $E_N \subset E_{2 \cdot N}$, $\alpha f + \beta g \in C_{2 \cdot N} \subset \bigcup_n C_n$. ∎

Proof of Theorem 5.1: If $S = \overline{\pi_W(L^2(Z))} \subsetneq L^2(W)$, then it is immediate that the only Y for which an instrumental regression can exist belong to the closed, proper subspace $\overline{\pi_W^{-1}(S)}$), proving the last part of the Theorem.

Now treat π_W as a continuous operator from $L^2(Z)$ to $L^2(W)$ (it still has operator norm 1), and suppose that $\pi_W(L^2(Z)) \subsetneq L^2(W)$. There are two cases, when π_W is bounded below, $\inf\{\|\pi_W(f)\| : f \in L^2(Z) \cap \partial U\} = \delta > 0$, and $\inf\{\|\pi_W(f)\| : f \in L^2(Z) \cap \partial U\} = 0$.

When π_W is bounded below, it is invertible on the closure of its range (e.g. Halmos (1957, p. 38)), implying that the range is closed. A closed proper subspace of $L^2(W)$ is shy, completing the proof in this case.

If $\inf\{\|\pi_W(f)\| : f \in L^2(Z) \cap \partial U\} = 0$, then there exists a sequence $f_n \in L^2(Z) \cap \partial U$ such that $\|\pi_W(f_n)\| \to 0$. Let $\mathsf{N} = \overline{\mathbf{span}}\{f_n : n \in \mathbb{N}\}$. Expressing $L^2(Z)$ as the direct sum $\mathsf{N}^{\perp} \oplus \mathsf{N}$ shows that $\pi_W(L^2(Z)) \subset (\pi_W(\mathsf{N}))^{\perp} \oplus \pi_W(\mathsf{N})$. For any shy set S, $S^{\perp} \oplus S$ is shy, so it is sufficient to show that $\pi_W(\mathsf{N})$ is a shy subset of $L^2(W)$. For this, it is in turn sufficient to show that $\pi_W(\mathsf{N})$ is a countable union of compact sets.

$\mathsf{N} \cap \overline{U}$ is a weakly compact subset of $L^2(Z)$, so it is sufficient to show that, restricted to N, π_W is a compact operator (i.e. continuous when the range, $L^2(W)$, is given the norm topology and the domain, $\mathsf{N} \subset L^2(Z)$, is given the weak topology). This follows from $\|\pi_W(f_n)\| \to 0$. ∎

Proof of Lemma 5.1: The relevant metric on sub-σ-fields is

$$d_C(\mathcal{H}, \mathcal{G}) = \sum_{n \in \mathbb{N}} 2^{-n} \min\{1, \|E(f_n|\mathcal{H}) - E(f_n|\mathcal{G})\|\} \tag{11}$$

where f_n is a countable dense subset of $L^2(\mathcal{F})$. This metric and its basic properties are due to Cotter (1986), the Bayesian interpretation of this metric can be found in Stinchcombe (1990, 1993). Basic properties of projections show that for any $r > 0$, $S_{r+} = \{X \in L^2(\mathcal{F}) : d_C(\sigma(X), \mathcal{F}) \geq r\}$ is closed, so that its complement, S_{r-}, is open. $R = \bigcap_{\epsilon} S_{\epsilon-}$, so it is sufficient to show that each $S_{\epsilon-}$ is dense in $L^2(\mathcal{F})$.

Pick arbitrary $g \in L^2(\mathcal{F})$ and $\epsilon > 0$. It is sufficient to show that there exists an f such that $\|f - g\| < \epsilon$ and $d_C(\sigma(f), \mathcal{F}) < \epsilon$.

Enumerate a countable field $\mathcal{F}^{\circ}$ generating $\mathcal{F}$ as $\{E_n : n \in \mathbb{N}\}$. Let $\mathcal{F}_n^{\circ} = \sigma\{E_k : k \leq n\}$. $d_C(\mathcal{F}_n^{\circ}, \mathcal{F}) \downarrow 0$. Pick N_1 such that $d_C(\mathcal{F}_{N_1}^{\circ}, \mathcal{F}) < \epsilon$.

Define $g_n^- = \sum_{k \in \mathbb{Z}} \frac{k}{2^n} 1_{f^{-1}([k/2^n, (k+1)/2^n))}$ and $g_n^+ = g_n^- + \frac{1}{2^n}$. $\|g_n^- - g\| \downarrow 0$ and $\|g_n^+ - g\| \downarrow 0$. Pick N_2 such that $\|g_{N_2}^- - g\| < \epsilon$ and N_3 such that $\|g_{N_3}^+ - g\| < \epsilon$. For any $n \geq \max\{N_2, N_3\}$ and g' such that $g_n^- \leq g \leq g_n^+$, $\|g' - g\| < \epsilon$.

Pick $n \geq \max\{N_1, N_2, N_3\}$ and enumerate the partition of Ω generated by $\mathcal{F}_n^\circ$ as $\{E_j : j \leq m\}$. Define $f = g_n^- + \sum_{j \leq m} \frac{j}{m \cdot 2^n + 1} 1_{E_j}$. By construction, $g_n^- \leq f \leq g_n^+$ so that $\|f - g\| < \epsilon$. Also by construction, $\sigma(f)$ is at least as large as $\mathcal{F}_n^\circ$ so that $d_C(\sigma(f), \mathcal{F}) < \epsilon$. ∎

Proof of Lemma 5.2: The only closed subspaces of the range of a compact operator are finite dimensional (e.g. Halmos (1982, Ch. 20, #180, p. 96)). Therefore, $\pi_W(L^2(Z)) \subsetneq L^2(W)$ unless the range of W is finite. Suppose that $g \in L^2(W)$ is non-constant and g is independent of Z. Because g is non-constant, $L^2(g)$ is a closed linear subspace of $L^2(W)$ strictly containing the constant functions. By independence, any $f \in L^2(Z)$ projects onto a constant function in $L^2(g)$. ∎

Proof of Theorem 5.2: Stinchcombe (2002, Theorem 4.4) proves this result when I and J have one dimension. The general case follows from the result that all uncountable Borel subsets of complete separable metric spaces are measurably isomorphic (e.g. Dellacherie and Meyer (1978, Theorem III.19, p. 48)). ∎

Proof of Lemma 6.1: Let $D_r = \{x \in \mathfrak{X} : T_{\mu,r}(x) \leq 1\}$ and let E_r be the polar of D_r. $E_r \subset \mathsf{aco}\,(C)$ because $T_{\mu,r}(x) \leq T_{\mu,\infty}(x) \leq \sup\{|\langle x, x^* \rangle| : x^* \in C\} = \sup\{|\langle x, x^* \rangle| : x^* \in \mathsf{aco}\,(C)\}$. ∎

Proof of Theorem 6.1: Baire largeness: For each f in a countable dense subset of $\mathfrak{X}$ and each rational $\epsilon > 0$, let $\mathcal{E}(f, \epsilon) \subset \mathbb{K}$ denote the closed set of E such that $d(f, \overline{\mathbf{span}}\,(E)) \geq \epsilon$. The set of non-consistent elements of $\mathbb{K}$ is the countable union the $\mathcal{E}(f, \epsilon)$, so it is sufficient to show that each $\mathcal{E}(f, \epsilon)$ has empty interior.

Pick $E \in \mathcal{E}(f, \epsilon)$ and arbitrary $\delta > 0$. It is sufficient to show that no δ-ball around E fails to contain a set that spans f. Pick $\gamma > 0$ such that $d_H(\mathsf{aco}\,(E \cup \{\gamma f\}), E) < \delta$. For all $\gamma > 0$, $f \in \overline{\mathbf{span}}\,(E \cup \{\gamma f\})$, completing the proof of Baire largeness.

Prevalence: Since the set of consistent elements is Baire large, it is not empty. Let E be one of the consistent elements. Let $V = \mathbb{R} \cdot E \subset \mathbb{K}$ be the one dimensional span of E. For any $E' \in \mathbb{K}$ and any $r \neq 0$, $E' + rE$ is consistent because $0 \in E'$ so that $\overline{\mathbf{span}}\,(E' + rE) \supset \overline{\mathbf{span}}\,(E)$. ∎

Proof of Theorem 6.2: The result is immediate when Θ is finite. The rest of the proof covers the case that Θ is infinite.

For each g in a countable dense subset of $C(\Theta; \mathfrak{X})$ and each rational $\epsilon > 0$, let $\mathcal{E}(g, \epsilon)$ be the set of $f \in C'$ such that $d(g, \overline{\mathbf{span}}\,(f(\Theta))) \geq \epsilon$. Taking the countable union over the g's and ϵ's, it is sufficient to show that $\mathcal{E}(g, \epsilon)$ is closed and has empty interior.

Closure: Let $f_k \to f$, $f_k \in \mathcal{E}(g, \epsilon)$, but suppose, for the purposes of contradiction, that $d(g, \overline{\mathbf{span}}\,(f(\Theta))) < \epsilon$. This implies that there exist $N \in \mathbb{N}$ and $(\beta_n, \theta_n)_{n \leq N}$ such that $d(g, \sum_{n \leq N} \beta_n f(\theta_n)) < \epsilon$. Because $f_k \to f$, we know that

$\sum_{n\leq N}\beta_n f_k(\theta_n)) \to \sum_{n\leq N}\beta_n f(\theta_n))$. For large k, $d(g, \sum_{n\leq N}\beta_n f_k(\theta_n)) < \epsilon$, contradicting the assumption that $f_k \in \mathcal{E}(g,\epsilon)$.

Empty interior: Pick arbitrary $f \in \mathcal{E}(g,\epsilon)$ and arbitrary $\delta > 0$. It is sufficient to show the existence of an $h \in B(f,\delta)$ such that $d(g, \overline{\mathbf{span}}\,(h(\Theta))) < \epsilon$.

Since Θ is an infinite, compact metric space, it has an accumulation point θ_0. Since f is continuous, there exists a $\delta' > 0$ such that for $d(f(\theta_0), f(B(\theta_0,\delta'))) < \delta/2$. Pick arbitrary $\theta_1 \in B(\theta_0,\delta')$ so that $d(f_0,f_1) < \delta/2$ where $f_0 = f(\theta_0)$ and $f_1 = f(\theta_1)$.

Pick $r \neq 0$ so that $d(rg,0) < \delta/2$. Because C' is dense and full, it contains an $h \in B(f,\delta)$ such that $h(\theta_0) = f_1 + rg$ and $h(\theta_1) = f_1$. Since $\beta(h(\theta_0) - h(\theta_1)) \in \mathbf{span}\, h(\Theta)$, setting $\beta = 1/r$ shows that $g \in \mathbf{span}\, h(\Theta)$. ∎

Department of Economics, University of Texas, Austin, TX 78712-1173 USA, e-mail: maxwell@eco.utexas.edu