ON THE RECOVERABILITY OF FORECASTERS' PREFERENCES

ROBERT P. LIELI & MAXWELL B. STINCHCOMBE

Abstract. We provide a nearly complete analysis of the problem of recovering an expected utility maximizing forecasters' preferences from a sequence of forecasts, realizations, and covariates used in producing the forecasts. There are essentially different expected utility functions that lead to the same forecasts in all situations. We show that these are non-generic in a very strong sense, and the "nearly complete" aspect of the analysis arises only from the non-generic cases. We also give set identification results when the forecaster uses covariates not observed by the econometrician.

Date: November, 2008

Preliminary and incomplete. Please do not cite or circulate.

Contents

1. Introduction 3
1.1. Testing the Rationality of Forecasts 3
1.2. Outline and Summary 4
2. Basics 4
2.1. The Model of Forecaster Behavior 5
2.2. The Binary Case 6
2.2.1. The Most That Can Be Recovered 6
2.2.2. The Requisite Variability 7
2.2.3. When Some Covariates Are Not Observed 8
2.3. The Utility Functions 9
2.4. Useful Notation and Definitions 10
2.5. Preliminary Results 12
3. Nonparametric Identification 12
3.1. Potential Identifiability 13
3.2. A Totally Large Set of Preferences are Potentially Identifiable 13
3.3. The Observational Requirements 15
3.3.1. Necessary Variability in the Set of Conditional Distributions 15
3.3.2. Differences on Smaller Sets of Probabilities 16
3.3.3. Variability in the Set of Allowable Forecasts 17
4. Identification for Parametric Classes 18
4.1. Parametrizations 19
4.2. Identifying Parameters Within a Parametrization 19
4.3. Potential Identifiability of Parametrizations 20
4.4. Distance Minizing Identification 20
4.5. FOC Local Identification 22
Appendix: Proofs 23
C. Proof of Theorem 1 25

1. Introduction

The purpose of this paper is to examine how much can be learned about an expected utility maximizing forecaster's utility function from a sequence of forecasts, the corresponding sequence of realizations, and a sequence of covariates used in producing the forecasts. In the non-parametric case, a necessary condition for forecasters with different utility functions to be distinguishable from each other is that there be an immense amount of variability both in the conditional distributions of the variable to be forecast and set of available forecasts. Outside of a small (i.e. non-generic) set of utility functions, these conditions are also sufficient. By contrast, for all but a small (i.e. non-generic) set of parametrizations, it takes a minimal amount of variability in the conditional distributions and none in the set of available forecasts to be able to distinguish different utility functions.

1.1. Testing the Rationality of Forecasts.

The existing econometric literature on the problem of recovering forecaster preferences is rather slim. The only papers we know of that address this question explicitly are Elliott et al. (2003, 2007) and, partly, Patton and Timmermann (2005). There is a much larger related literature concerned with testing the rationality of forecasts that dates back to at least Mincer and Zarnowitz (1969).

Empirical work in this area typically relies on the assumption that the forecaster's objective is to minimize mean squared error loss.[1] Indeed, the square loss function is technically convenient and has very sharp observable implications concerning the properties of optimal forecasts, including unbiasedness, uncorrelatedness of one-step-ahead forecast errors, increasing forecast error variance as the forecast horizon expands, etc.

However, as argued by Granger (1969), economic forecasts are often produced in an environment where the square loss function or, more generally, any symmetric loss function, does not adequately capture the costs resulting from overprediction vs. underprediction. Under general loss functions, all of the optimality properties listed above can be lost, and, as pointed out by Elliott et al. (2003), purported tests of forecast rationality based on the implications of minimizing mean squared error are more appropriately viewed as joint tests of forecaster rationality and the mean square loss function.

[1] Patton and Timmermann (2005) cite the following URL for a collection of papers concerned with testing the rationality of economic forecasts: `www.Phil.frb.org/econ/spf/spfbib.html`

Given that the notion of forecast rationality is inextricably linked to the objective that the forecaster is presumably trying to achieve, there are two ways to study individual forecasters' behavior. If interest continues to center on testing for optimizing (rational) behavior, then one needs to explore further the properties of optimal forecasts under general classes of loss functions that allow for asymmetries and functional forms other than square loss. This approach is outlined by Patton and Timmermann (2005). Alternatively, one can focus on the inverse problem: maintain the assumption of optimizing behavior, identify and estimate a loss function, or a class of loss functions, consistent with the properties of the observed forecasts. This is the approach taken by Elliott et al. (2003) and this paper.

1.2. **Outline and Summary.** The next section sets notation and assumptions. The subsequent section studies the identification of nonparametric utility functions, showing that a weak form of identification is all that is generally achievable, and that for all but a small (i.e. non-generic) set of utility functions, this form of identification is achievable. The form of identification is weak because it requires, first, that the conditional distributions of the predicted variable that the forecaster faces be extremely rich, and second, that there also be nearly complete variability in the set of allowable forecasts.

The penultimate section studies 'identification by parametric assumption,' and the results are much more positive. For all but a small (i.e. non-generic) set of parametrizations, identification is possible under minimal conditions on the set of conditional distributions of the predicted variable that the forecaster faces. Throughout, we first analyses the case of a compact set of possible realizations and forecasts, and then treat the generalizations.

The last section contains conclusions and sketches possible extensions of the results given here.

2. Basics

We begin with the model of forecaster behavior, then examine the possibilities and difficulties of recovering forecaster preferences in the binary case. Of particular import is our assumption that the forecasters are happiest when they make no error, a "no bias in case of certainty" assumption. In the general case, this leads to a canonical form of the utility functions. After we give this, we gather some notation and preliminary results.

2.1. **The Model of Forecaster Behavior.** The variable to be forecast at time t is a random scalar Y_{t+1} that takes values in a set $D \subset \mathbb{R}$. The forecaster possesses a jointly continuous utility function on $D \times D$, $(\hat{y}, y) \mapsto u(\hat{y}, y)$. The first argument, $\hat{y}$, denotes the value of the forecast, and the second, y, the actual realization of Y_{t+1}.

$\Delta(D)$ denotes the set of (countably additive Borel) probabilities on D with the Prokhorov metric[2] and the associated σ-field, and let $\Delta^2(D) = \Delta(\Delta(D))$ denote the (countably additive Borel) probabilities on $\Delta(D)$. Let $p_t \in \Delta(D)$ be the conditional distribution of Y_{t+1} given the information held by the forecaster at time t, and let f_t denote forecast of Y_{t+1} made at time t. We assume that f_t is an element of the set

$$(1) \qquad Br\left(p_t \mid F, u\right) := \arg\max_{\hat{y} \in F} \int u(\hat{y}, y) p_t(dy),$$

where F is a subset of D. A leading case is $F = D$, but we will need more generality.

We assume that p_t depends on the information available at time t only through a finite dimensional vector $(X_t, X_t') \in \mathbb{R}^{\ell+m}$ where X_t both X_t and X_t' are observed by the forecaster, but X_t is all that is observed by the econometrician. The vector X_t may contain current and lagged values of Y, e.g. $X_t = (Y_t, \ldots, Y_{t-k})$, and X_t' may be constant, i.e. $m = 0$.

We also assume that $(Y_{t+1}, (X_t, X_t'))$ is a strictly stationary process, defined on a probability space $(\Omega, \mathcal{F}, P)$, with limit distribution defined by $Q(A \times B) = \lim_T \frac{1}{T} \sum_{t=1}^{T} 1_A(Y_{t+1}) \cdot 1_B(X_t, X_t')$. The marginal distributions of Q are denoted Q_Y and $Q_{X,X'}$. Thus, $p_{X_t, X_t'}(\cdot) = Q(\cdot | X_t, X_t') \in \Delta(D)$ is the (regular) conditional distribution, and $\mathfrak{Q} \in \Delta^2(D)$ denotes the asymptotic distribution over conditional distributions that the forecaster faces.

If, in addition, the maximum in (1) is unique, or if the forecaster uses a fixed tie-breaking rule, then there will be a time-invariant, non-random mapping linking the possible values of (X_t, X_t') and f_t; denote this behavioral mapping by $f(\cdot)$, $f_t = f(X_t, X_t')$. If X_t' is constant,

[2]The Prokhorov metric is $\rho(p, q) = \inf\{\epsilon \geq 0 : \forall A,\ p(A) \leq q(A^\epsilon) + \epsilon,\ \text{and}\ q(A) \leq p(A^\epsilon) + \epsilon\}$. The distance between point masses, $\rho(\delta_y, \delta_{y'})$, is equal to $\min\{|y - y'|, 1\}$, and $\rho(p_n, p) \to 0$ iff $\int h\, dp_n \to \int h\, dp$ for all bounded continuous h. The Borel σ-field on $\Delta(D)$ generated by ρ is the smallest one containing all sets of the form $\{Q \in \Delta(D) : Q(E) \leq r\}$, E measurable and $r \in [0, 1]$.

then, over time, it is in principle possible to recover the function f over the support of X_t.[3]

The function $f(\cdot)$ is the composition of two other functions. First, (X_t, X_t') is mapped into $p_{X_t,X_t'}$. Second, $p_{X_t,X_t'}$ is mapped into the forecast f_t in a way that depends on the forecaster's utility function u, i.e. $f_t = Br\left(p_{X_t,X_t'} \mid F, u\right)$. Since (Y_{t+1}, X_t) is observed and stationary, $X_t \mapsto p_{X_t}$ is nonparametrically identifiable, but contains no information on u. In the case $m = 0$ where the econometrician observes all of the covariates used by the forecaster, it is the mapping $p_{X_t} \mapsto Br\left(p_{X_t} \mid F, u\right)$ from which one can hope to learn about u. In the case $m > 0$ where the forecaster uses information not observed by the ecnometrician, it is the mapping $p_{X_t} \mapsto \Delta(Br\left(p_{X_t,X_t'} \mid F, u\right)$ from which one can hope to learn about u.

2.2. **The Binary Case.** Many of the basic results can be seen in trying to recover the preferences of a forecaster who forecasts a binary random variable. The questions are,

(1) What is the maximal information about a utility function that can be recovered when we observe all of the variables used in making the forecast?
(2) What variability of the data is required in order to recover this maximal amount of information?
(3) What about when we do not observe all of the variables used in making the forecast?

2.2.1. *The Most That Can Be Recovered.* Modulo a dominance requirement, the most that can be recovered about a utility function is which generalized affine equivalence class it belongs to.

Example 1 Suppose that $D = \{0, 1\}$, $X_t = Y_t$, that the $\{Y_t : t \in \mathbb{N}\}$ are i.i.d. with $P(Y_t = 1) = r$, $P(Y_t = 0) = 1 - r$ for some $r \in [0, 1]$, that the forecaster starts with e.g. a uniform $[0, 1]$ prior for r, and forms p_t by updating. At time t, the posterior over $[0, 1]$ is a Beta distribution, $Beta(N_t + 1, M_t + 1)$ where $N_t = \sum_{s \le t} X_t$ is the number of 1's that have been observed at times $1, 2, \ldots, t$, and $M_t = t - N_t$. Thus, $p_t(Y_{t+1} = 1) = p_t(1) = (N_t + 1)/(t + 2)$.

Given utility function $u(\hat{y}, y)$, $\hat{y}, y \in \{0, 1\}$, the forecaster solves

$$\max\left\{\int u(0, y)\, dp_t(y), \int u(1, y)\, dp_t(y)\right\} \tag{2}$$

for the optimal forecast, f_t. ∎

[3]The support of a random variable X is the smallest closed set C with $P(X \in C) = 1$.

Throughout, we will make the dominance assumption that the forecaster is happiest being correct. In this binary case, this corresponds to $u(0,0) > u(1,0)$ and $u(1,1) > u(1,0)$. These two assumptions are equivalently formulated as there being **no bias in the case of certainty (nbcc)**, i.e. $Br(\delta_y \mid D, u) = \{y\}$ where δ_y is point mass on y. In this, the binary case, this reduces to $[p_t(Y_{t+1} = 1) = 1] \Rightarrow [f_t = 1]$ and $[p_t(Y_{t+1} = 0) = 1] \Rightarrow [f_t = 0]$.

Given nbcc, there is a critical value $c \in (0,1)$ for the problem in (2). Specifically, $f_t = 1$ if $p_t(1) > c$ and $f_t = 0$ if $p_t(1) < c$ where $c = \frac{[u(0,0)-u(0,1)]}{[u(0,0)-u(0,1)]+[u(1,1)-u(1,0)]}$. Further, this critical value c is **all** that can be recovered from the utility function $u(\cdot,\cdot)$.

The utility function v is a **generalized affine transformation** of u if for some $r > 0$ and function $y \mapsto g(y)$, $v(\hat{y}, y) = r \cdot u(\hat{y}, y) + g(y)$. In the binary case, v is a generalized affine transformation of u iff u and v have the same critical value.[4]

2.2.2. *The Requisite Variability.* The extent to which c can be recovered depends on the random sequence $\{p_t : t \in \mathbb{N}\}$ and the relation between true probability r and the critical value c:

1. If $r = c$, then for all $\epsilon > 0$, $p_t(1) \in (c - \epsilon, c)$ infinitely often and $p_t(1) \in (c, c + \epsilon)$ infinitely often with probability 1. In this case, c is identified and as much as can be possibly recovered of the utility function is recovered.
2. If $r \neq c$, then with probability 1, $d(c, \{p_t(1) : t \in \mathbb{N}\}) > 0$, and all that will be recovered is some random interval containing the true value of c.

What is needed to recover the critical value is sufficient variability of the set of p_t's.

Example 2 Now suppose in Example 1 that there is a sequence of covariates X_t such that (X_t, Y_{t+1}) is i.i.d., Y_t is Bernoulli(r) as above, and the support of the random variable $P(Y_{t+1} = 1|X_t)$ is $[0,1]$. ∎

Here, for all c, $d(c, \{p_t(1) : t \in \mathbb{N}\}) = 0$ with probability 1, and as much information about u as is possible to recover is recovered with probability 1. More generally, c needs to be in the interior of the support of $P(Y_{t+1} = 1|X_t)$ for complete recoverability.

[4]The set of utility functions on $\{0,1\} \times \{0,1\}$ is 4-dimensional. The reduction to a 1-dimensional scale c happens as follows: one dimension is lost to the positive affine rescaling by r; and another two are lost by the addition of g, which belongs to the 2-dimensional set of functions on $\{0,1\}$.

When D is finite, non-parametric recovery of the utility function is a finite dimensional problem, when D is infinite, recoverability is no longer finite dimensional.

2.2.3. *When Some Covariates Are Not Observed.* We now suppose that the forecaster uses a vector of variables $(X, X') \in \mathbb{R}^{\ell+m}$ in producing forecasts, but that the econometrian observes only $X \in \mathbb{R}^{\ell}$.

Example 3 Now suppose in Example 1 that there is a sequence of covariates X_t such that $((X_t, X_t'), Y_{t+1})$ is i.i.d., Y_t is Bernoulli(r) as above, and the support of the random variable $P(Y_{t+1} = 1|(X_t, X_t'))$ is $[0, 1]$. ∎

For each value $p_x = P(Y_{t+1} = 1|X_t = x)$, the proportion of the time, Q_x, that the forecaster chooses $\hat{y} = 1$ is identified. For each such p_x, let $R_x \in [0, 1]$ be the random variable $P(Y_{t+1} = 1|(x, X_t'))$, so that iterated expectations delivers $E\, R_x = p_x$ a.e.

If $R_x > c$, the Forecaster will choose $\hat{y} = 1$, if $R_x < c$, they will choose $\hat{y} = 0$. The bounds on the probability Q_x that $\hat{y} = 1$ when $X_t = x$ that are implied by the value of the unknown c arise from finding

$$a_{p_x} = \inf P(R_x > c) \text{ and } A_{p_x} = \sup P(R_x > c) \tag{3}$$

subject to the conditions that $R_x \in [0, 1]$ and $E\, R_x = p_x$. For $p_x \leq c$, these bounds are $[a_{p_x}, A_{p_x}] = [0, \frac{p}{c}]$, and for $p_x > c$, they are $[a_{p_x}, A_{p_x}] = [\frac{p-c}{1-c}, 1]$.

Let L_c denote the set of (p, Q) pairs with $0 \leq Q \leq \frac{p}{c}$ for $p \leq c$ and $\frac{p-c}{1-c} \leq Q \leq 1$ for $p > c$. The set of (p_x, Q_x) pairs that arise is observable, and the **identification set** is the interval $\{c \in (0, 1) : (p_x, Q_x) \in L_c \text{ a.e.}\}$.

In the extreme case that X_t is stochastically idependent of (X_t', Y_{t+1}), then $\sigma(p_X)$ is the trivial σ-field, and the covariates that the econometrician observes are entirely useless for producing forecasts. Even in this case, there may be some identifying power. There will be none if the single observed pair (p_x, Q_x) is on the diagonal, because the diagonal is a subset of every L_c. However, the further the pair is from the diagonal, the smaller is the identification set.

With more variability, it is even possible that the identification set is a singleton. If there are two (or more) observed (p_x, Q_x) pairs, then $\sigma(p_X)$ contains at least a non-trivial partition of Ω. In this case, if one of the $(p_x, Q_x$ takes its value on the upper sloped part of the boundary of the true L_c and another takes its value on the lower sloped part of the

boundary, then there is a unique c with L_c containing the (p, Q) pairs a.e. (See Figure 2.) In general, the richer $\sigma(p_X)$ and the greater the variability of the random variables R_x, the smaller is the identification set.

2.3. **The Utility Functions.** $C = C(D \times D)$ is the set of continuous functions on $D \times D$. To rule out some mathematically perverse cases such as D being the set of rationals, we assume throughout that D is a countable union of open sets having compact closure, i.e. is a strongly σ-compact subset of $\mathbb{R}$ (see e.g. Corbae, Stinchcombe, Zeeman (2009, §6.10.b) for properties of such set).

One cannot determine the tradeoffs between different possible forecasts if one, or both, of the forecasts is never made. The following condition rules out the existence of completely dominated forecasts.

Definition 1. *We say that $u(\hat{y}, y) \in \mathcal{C}$ has* ***no bias in case of certainty (nbcc)*** *if for all $y \in D$, $Br_u(\delta_y \mid D) = \{y\}$. The set of utility functions in $\mathcal{C}$ with the nbcc property will be written as $\mathcal{C}_{nbcc} = \mathcal{C}_{nbcc}(D \times D)$.*

Equivalently, $u \in \mathcal{C}_{nbcc}$ means that for any fixed $y \in D$, the function $\hat{y} \mapsto u(\hat{y}, y)$ has a unique global maximum at $\hat{y} = y$, i.e. $u(\hat{y}, y) \leq 0$ with equality if and only if $\hat{y} = y$. If the forecaster is certain that a given value of Y will be realized, then *nbcc* requires that the unique optimal point forecast coincides with that value. Granger and Machina (2005) show that if $u(\hat{y}, y)$, the utility associated with a forecast-realization pair, derives from an underlying decision problem, then $u(\hat{y}, y)$ will possess the *nbcc* property almost automatically.[5]

Definition 2. $u, v \in C(D \times D)$

(a) *are* ***affinely equivalent****, written $u \sim_{aff} v$, if there exists a continuous $y \mapsto g(y)$ and $r > 0$ such that $v(\hat{y}, y) = r \cdot u(\hat{y}, y) + g(y)$, and*

(b) *are* ***forecast equivalent****, written $u \sim_{Br} v$. if the have the same forecasts on compact sets, $(\forall$ compact $F \subset D)(\forall p \in \Delta(F)[Br\,(p \mid F, u) = Br\,(p \mid F, v)]$.*

Lemma 1. *Affine equivalence implies forecast equivalence, but best response equivalence does not imply affine equivalence.*

[5] At the cost of some confusing relabelling of the set of forecasts, we could replace this condition with the Moriss and Ui (2004, Proposition 2) condition that every forecast be strictly optimal for some δ_y. Note that *nbcc* is *not* satisfied if the forecaster has an interest in the decisions that the consumer of the forecast will make as in the study of cheap talk games begun by Crawford and Sobel (1982).

Proof: $\int_F v(\hat{y},y)p(dy) = \int_F [r\cdot u(\hat{y},y)+g(y)]\,p(dy) = r\cdot\left[\int_F u(\hat{y},y)\,p(dy)\right]+\int_F g(y)\,p(dy)$, where $\int_F g(y)\,p(dy)$ exists, as $g(y)$ is continuous and hence bounded on the compact set F. As this term does not depend on $\hat{y}$, it is clear that $\hat{y}^* \in D$ solves $\max_{\hat{y}} \int u(\hat{y},y)\,p(dy)$ iff it also solves $\max_{\hat{y}} \int v(\hat{y},y)\,p(dy)$. For the failure of the reverse implication, see Example 4. ∎

Each affine equivalence class contains a canonical form.

Definition 3. *The* ***canonical form*** *of a $u \in C(D\times D)$ is defined as $u^c(\hat{y},y) = u(\hat{y},y) - u(y,y)$. The set of utility functions in canonical form is denoted $\mathcal{C}(D\times D)$ or simply $\mathcal{C}$.*

The canonical form of a utility function that has *nbcc* is characterized by the property that $u^c(\hat{y},y) \le 0$ and is equal to 0 iff $\hat{y} = y$. $\mathcal{C}(D\times D)$ is a convex cone not containing its vertex, i.e. for all $\alpha,\beta > 0$ and $u,v \in \mathcal{C}(D\times D)$, $\alpha u + \beta v \in \mathcal{C}(D\times D)$ and $0 \notin \mathcal{C}(D\times D)$. If $\#D = M < \infty$, then $\mathcal{C}$ is isomorphic to the strictly negative elements of $\mathbb{R}^{M(M-1)}$ and non-parametric preference recovery is a finite dimensional problem. If D is an infinite set then $\mathcal{C}(D\times D)$ is infinite dimensional.

Clearly, $u \sim_{aff} u^c$ (let $r = 1$ and $g(y) = -u(y,y)$), and so $u \sim_{Br} u^c$. This means that a utility function and its canonical form cannot be indistinguished with any data set of covariates and forecasts. We will restrict attention to utility functions in canonical form, i.e. replace $C(D\times D)$ with $\mathcal{C}(D\times D)$. To simplify notation, we will drop the superscript c in referring to the elements of $\mathcal{C}$. Note that for u and v in $\mathcal{C}$, $u \sim_{aff} v$ if and only if $v = r\cdot u$ for some $r > 0$.

2.4. **Useful Notation and Definitions.** At various points, we will use the following.

(1) For $y \in D$, $\delta_y \in \Delta(D)$ denotes point mass on y, i.e. $\delta_y(E) = 1_E(y)$.
(2) For $y \in D$ and $\epsilon > 0$, $B_\epsilon(y) = \{y' \in D : |y'-y| < \epsilon\}$ is the ϵ-ball around y.
(3) For $p \in \Delta(D)$ and $\epsilon > 0$, $B^\rho_\epsilon(p) = \{q \in \Delta(D) : \rho(p,q) < \epsilon\}$ denotes the ϵ-ball around p in the Prokhorov metric.
(4) For $p \in \Delta(D)$, the support of p is the smallest closed set having p-mass 1, i.e. $\cap\{F : F \text{ closed},\ p(F) = 1\}$.
(5) $\mathcal{K}(D)$ denotes the class of compact subsets of D.
(6) For $F \in \mathcal{K}(D)$, $\Delta(F)$ denotes the set of (countably additive Borel) probabilities on F.

(7) $\partial\Delta(F)$ denotes the boundary of $\Delta(F)$ within the set of measures on F with absolute variation of 1.[6]
(8) For $A \subset D$, $[A]^\epsilon = \cup_{y\in A} B_\epsilon(y)$, denotes the ϵ-ball around the set A.
(9) For $A \subset \Delta(D)$, $[A]^\epsilon = \cup_{p\in A} B^\rho_\epsilon(p)$. denotes the Prokhorov ϵ-ball around the set A.
(10) $d_H(A, B) = \inf\{\epsilon \geq 0 : A \subset [B]^\epsilon,\ B \subset [A]^\epsilon\}$ denotes the (Hausdorff) distance between compact sets.

At various points, we will show that some results hold except for a "small" set of "rare" exceptional cases. Our notion of smallness combines Baire's (1899) topological and Anderson and Zame's (2001) measure theoretic definitions.

Definition 4. *If C is a separable, topologically complete, convex subset of a topological vector space, we say that a set $S \subset C$ is* ***totally small*** *if it is both Baire small and relatively shy. A set is* ***totally large*** *if it is the complement of a totally small set.*

A set is small in Baire's sense if it is the countable union of closed sets having no interior, and is large in Baire's sense it its complement is small. If $\{q_n : n \in \mathbb{N}\}$ enumerates the points in $\mathbb{R}^k$ with rational coordinates, then $E(\epsilon) := \cup_n B_{\epsilon/2^n}(q_n)$ is an open dense subset having Lebesgue measure ϵ, and $F(\epsilon)$, the complement of $E(\epsilon)$ is a closed set having no interior. Therefore, $\cap_n E(1/n)$ is a Baire large set having Lebesgue measure 0 and $\cup_n F(1/n)$ is a Baire small set having full Lebesgue measure.

Since there is no Lebesgue measure on infinite dimensional spaces, we use Anderson and Zame's extension of Lebesgue measure 0 to convex subsets of an infinite dimensional metric vector space V. A set S is shy relative to C if for all $c \in C$, all neighborhoods U_c of c, and all $\epsilon > 0$, there exists a compactly supported $\eta \in \Delta(C)$ such that $\eta(U_c \cap [\epsilon C + (1-\epsilon)c]) = 1$ and $(\forall x \in V)[\eta(S' + x) = 0]$. A useful sufficient condition for shyness is **finite shyness**, which involves η being the continuous affine image of the uniform distribution on the unit ball in $\mathbb{R}^k$ for some k.

The class of totally small sets is closed under countable unions, a totally small set has a non-empty interior, and if C is finite dimensional, a totally small set must have Lebesgue measure 0, though this is not sufficient.

[6] If F is finite, this is the set of probabilities assign mass 0 to at least one point in F, more generally, it is the set of probabilities with supports that are a strict subset of F.

2.5. **Preliminary Results.** A compact-valued correspondence Γ is **upper hemicontinuous (uhc)** if for every r and every $\epsilon > 0$ there exists $\delta > 0$ such that $[d(r, r') < \delta] \Rightarrow [\Gamma(r') \subset [\Gamma(r)]^{\epsilon}]$. The Theorem of the Maximum (e.g. Corbae, Stinchcombe, Zeeman (2009), Theorem 4.10.2 (p. 151)) delivers the following.

Lemma 2. *The mapping* $(p, F, u) \mapsto Br\,(p \mid F, u)$ *from* $\Delta(D) \times \mathcal{K}(D) \times \mathcal{C}$ *to* $\mathcal{K}(F)$ *is uhc.*

Expected utility maximization yields a useful property of the sets of beliefs leading to a particular forecast being optimal.

Lemma 3. *For* $F \in \mathcal{K}(D)$ *and* $\hat{y} \in D$*, then* $\{p \in \Delta(F) : \hat{y} \in Br\,(p \mid F, u)\}$ *is closed and convex.*

Proof: Closure comes from Lemma 2. For convexity, suppose that for all $\hat{y}' \in F$,

$$\int u(\hat{y}, y)\, dp(y) \geq \int u(\hat{y}', y)\, dp(y) \text{ and } \int u(\hat{y}, y)\, dq(y) \geq \int u(\hat{y}', y)\, dq(y).$$

By linearity of the integral, $\int u(\hat{y}, y)\, d(\alpha p + (1-\alpha) q)(y) \geq \int u(\hat{y}', y)\, d(\alpha p + (1-\alpha) q)(y)$ for any $\alpha \in [0, 1]$. ∎

Upper hemicontinuous correspondences can 'explode.' In the *nbcc* context, the degree of explosion is limited. The following tells us that if $\hat{y}$ is an optimal forecast at p, then there are q's arbitrarily close to p for which $\hat{y}$ is the unique optimal forecast.

Lemma 4. *If* $F \in \mathcal{K}(D)$*,* u *has nbcc, and* $\hat{y} \in Br\,(p \mid F, u)$*, then for all* $\epsilon > 0$*, there exists* $q \in \Delta(F)$*,* $q \neq p$*,* $\rho(p, q) < \epsilon$ *such that* $Br\,(q \mid F, u) = \{\hat{y}\}$*.*

Proof: Since u has *nbcc*, $Br\,(\delta_{\hat{y}} \mid F, u) = \{\hat{y}\}$, that is, $\int u(\hat{y}, y)\, d\delta_{\hat{y}}(y) > \int u(\hat{y}', y)\, d\delta_{\hat{y}'}(y)$ for all $\hat{y}' \neq \hat{y}$. Since $\hat{y} \in Br\,(p \mid F, u)$, $\int u(\hat{y}, y)\, dp(y) \geq \int u(\hat{y}', y)\, dp(y)$ for all $\hat{y}' \neq \hat{y}$. Therefore, for any $\alpha \in (0, 1)$, $\int u(\hat{y}, y)\, d(\alpha \delta_{\hat{y}} + (1-\alpha) p)(y) > \int u(\hat{y}', y)\, d(\alpha \delta_{\hat{y}} + (1-\alpha) p)(y)$ for all $\hat{y}' \neq \hat{y}$. For α sufficiently close to 0, $\rho(\alpha \delta_{\hat{y}} + (1-\alpha) p, p) < \epsilon$. ∎

3. Nonparametric Identification

We wish to recover an arbitrary continuous expected utility function by observing optimal choices in response to different conditional distributions of the variable to be forecast. As we will show, this requires essentially full variability in the conditional distributions and the sets of available forecasts.

3.1. **Potential Identifiability.** It is only possible to distinguish utility functions if they sometimes lead to different behavior.

Definition 5. *$u, v \in \mathcal{C}(D \times D)$ are **potentially identifiable** if they are not forecast equivalent, that is, if there exists a compact $F \subset D$ and some distribution $p \in \Delta(F)$ for which $Br\,(p \mid F, u) \neq Br\,(p \mid F, v)$.*

Affine equivalence implies that utility functions are not potentially identifiable. The reverse is not true, very different preferences may not be observationally distinguishable.

Example 4 For $D = \{1, 2, 4\}$, the following utility functions depend only on $|y - \hat{y}|$,

$$u(\hat{y}, y) = \begin{bmatrix} 0 & -1 & -3 \\ -1 & 0 & -2 \\ -3 & -2 & 0 \end{bmatrix} \quad \text{and} \quad v(\hat{y}, y) = \begin{bmatrix} 0 & -3 & -4 \\ -3 & 0 & -1 \\ -4 & -1 & 0 \end{bmatrix} \tag{4}$$

where m, n entry in each matrix corresponds to the utility of $(\hat{y}, y) = (m, n)$.

If one deletes row i and column i from both u and v, then the resulting 2×2 matrices are positive scalar multiples. As the scalar multiples differ, u and v are not strictly equivalent, and represent very different preferences over $\Delta(D \times D)$. However,

- for $p = (\frac{1}{2}, \frac{1}{2}, 0)$, $Br\,(p \mid D, u) = Br\,(p \mid D, v) = \{1, 2\}$,
- for $q = (0, \frac{1}{2}, \frac{1}{2})$, $Br\,(q \mid D, u) = Br\,(q \mid D, v) = \{2, 4\}$, and
- for $r = (\frac{1}{2}, 0, \frac{1}{2})$, $Br\,(r \mid D, u) = Br\,(r \mid D, v) = \{1, 2, 4\}$.

By Lemma 3 and *nbcc*, for all $p \in \Delta(D)$, $Br\,(p \mid D, u) = Br\,(p \mid D, v)$. (See Figure 1. NOTE THAT THIS EXAMPLE IS CHANGED FROM PREVIOUS VERSION) ∎

If forecaster preferences are like those in Example 4, then the intersection in the following definition cannot be a singleton set, that is, the forecaster's preferences cannot be pinned down.

Definition 6. *Let $\mathcal{A}(D) \subset \mathcal{K}(D) \times \Delta(D)$ be the set of (F, p) pairs with $p \in \Delta(F)$. A sequence $((F_n, p_n), \hat{y}_n)$ **pins down an affine equivalence class** if there is a unique canonical u consistent with the data, that is, if $\bigcap_n \{v \in \mathcal{C}(D \times D) : \hat{y}_n \in Br\,(p \mid F, v)\} = \{u\}$.*

3.2. **A Totally Large Set of Preferences are Potentially Identifiable.** The utility functions in Example 4 are special in a very delicate fashion — there is a point, specifically $r = (\frac{1}{2}, 0, \frac{1}{2})$, in $\partial\Delta(F)$, the boundary of $\Delta(F)$, at which every point in F is a best response.

Definition 7. *Let $\mathcal{G} = \mathcal{G}(D \times D)$ denote the collection of u in $\mathcal{C}_{nbcc}(D \times D)$ for which there exists a dense $D' \subset D$ such that for any three-point set $F = \{y_1, y_2, y_3\} \subset D'$, there is no $p \in \partial\Delta(F)$ with $Br\,(p \mid F, u) = F$.*

Theorem 1. *If $u \in \mathcal{G}(D \times D)$, then for all $v \in \mathcal{C}_{nbcc}$, $[u \sim_{Br} v] \Leftrightarrow [u \sim_{aff} v]$, and for any dense sequence (F_n, p_n) in $\mathcal{A}(D)$ and any $\hat{y}_n \in Br\,(p_n \mid F_n, u)$, $((F_n, p_n), \hat{y}_n)$ pins down u. Further, $\mathcal{G}(D \times D)$ is a totally large subset of $\mathcal{C}_{nbcc}(D \times D)$,*

The set $\mathcal{G}(D \times D)$ has a rather abstract definition. We turn to sufficient conditions for a utility function to belong to it.

Definition 8. *For convex D, a utility function $u \in \mathcal{C}_{nbcc}(D \times D)$ is* ***almost nowhere piecewise linear*** *if for all $y_1, y_2 \in D$, $y_1 \neq y_2$, for all $\alpha \in (0, 1)$, and for all κ, $\lambda(\{\hat{y} \in D : \alpha u(\hat{y}, y_1) + (1 - \alpha)u(\hat{y}, y_2) = \kappa\}) = 0$ (where λ is Lebesgue measure).*

This asks that no convex combination of any pair $u(\cdot, y_1)$ and $u(\cdot, y_2)$ be flat on a set having positive measure. If $u(\cdot, y_1)$ has a linear interval with positive slope that overlaps with a linear interval of $u(\cdot, y_2)$ having negative slope, the condition fails. More generally, it fails if $u(\cdot, y_1) = b - ru(\cdot, y_2)$ on some interval for some $b \in \mathbb{R}$ and $r > 0$.

Lemma 5. *The following are sufficient for $u \in \mathcal{C}_{nbcc}(D \times D)$ to belong to $\mathcal{G}(D \times D)$:*

(a) *$\#D = 2$;*
(b) *D is convex and for all $y \in D$, $u(\cdot, y)$ is strictly concave; and*
(c) *D is convex and u is almost nowhere piecewise linear.*

The risk averse (strictly concave) case covers generalized mean squared loss utility functions $u(\hat{y}, y) = -h(y)(\hat{y} - y)^2$ where $h(\cdot)$ is continuous and strictly positive, and generalized check functions, $-[|\hat{y} - y|^{\alpha} 1_{(\hat{y}-y)<0} + |\hat{y} - y|^{\beta} 1_{(\hat{y}-y)\geq 0}]$, so long as $\alpha, \beta > 1$. We do not know whether or not piecewise linear functions are potentially identifiable in the case of a convex D.

Proof: If $\#D = 2$, then D has no three point subsets, so every $u \in \mathcal{C}_{nbcc}(D \times D)$ must belong to $\mathcal{G}(D \times D)$.

If each $u(\cdot, y)$ is strictly concave, then it is almost nowhere piecewise linear, so the second sufficient condition follows from the third.[7]

[6] The proof of Theorem 1 is in Appendix A.

[7] For a direct proof of the strictly concave case, note that for any $y_a \neq y_b \in D$ and $\alpha \in (0, 1)$, $V(\hat{y}) := \alpha u(\hat{y}, y_a) + (1 - \alpha)u(\hat{y}, y_b)$ is strictly concave, so that if $V(y_a) = V(y_b)$, then no $y_c \neq y_a, y_b$ is indifferent to y_a and y_b.

Let $\{R_n\}_{n\in\mathbb{N}}$ be an i.i.d. sequence of random variables defined on a probability space $(\Omega, \mathcal{F}, P)$ and having a continuous, strictly positive density with respect to Lebesgue measure on D. We will show that for any u that is nowhere piecewise linear, there exists a probability 1 set of ω such that the dense set $\{R_n(\omega)\}_{n\in\mathbb{N}}$ serves as the D' in Definition 7.

Step 1: Since the R_n have a strictly positive density, there is an $\Omega' \in \mathcal{F}$ with $P(\Omega') = 1$, for all $\omega \in \Omega'$ and $n \neq n'$, $R_n(\omega) \neq R_{n'}(\omega)$, and $\{R_n(\omega)\}_{n\in\mathbb{N}}$ is dense in D.

Step 2: Let n_1, n_2, n_3 be one of the countably many subsets of $\mathbb{N}$ containing three distinct points, and condition on $R_{n_1} = y_1$ and $R_{n_2} = y_2$, $y_1 \neq y_2$. By *nbcc*, the unique probability on $\{y_1, y_2\}$ making y_1 and y_2 indifferent as forecasts is $\alpha\delta_{y_1} + (1-\alpha)\delta_{y_2}$ where $\alpha = \frac{u(y_1,y_2)}{u(y_1,y_2)+u(y_2,y_1)} \in (0,1)$. Letting $\kappa = \alpha u(y_2, y_1) = (1-\alpha)u(y_1, y_2)$, u being nowhere piecewise linear implies that

$$P(\{R_{n_3} : \alpha u(R_{n_3}, y_1) + (1-\alpha)u(R_{n_3}, y_2) = \kappa\}) = 0 \tag{5}$$

because R_{n_3} has a density with respect to Lebesgue measure. Since we conditioned on arbitrary $y_1 \neq y_2$, there is a probability 1 set of ω, call it $\Omega(n_1, n_2, n_3)$, for which there exists no $p \in \Delta(F)$ with $Br\,(p \mid u, F) = F$ where $F = \{R_{n_1}, R_{n_2}, R_{n_3}\}$. Define $\Omega'' = \bigcap \Omega(n_1, n_2, n_3)$ where the intersection is taken over three point subsets of $\mathbb{N}$ so that $P(\Omega' \cap \Omega'') = 1$.

Step 3: For all ω in $\Omega' \cap \Omega''$, and for all three point subsets, $F = \{y_1, y_2, y_3\}$ of the dense set $D' = \{R_n(\omega)\}_{n\in\mathbb{N}}$, there is no $p \in \partial\Delta(F)$ with $Br\,(p \mid F, u) = F$. ∎

3.3. **The Observational Requirements.** The denseness of the sequence (F_n, p_n) in $\mathcal{A}(D)$ of Theorem 1 is sufficient to pin down a utility function in $\mathcal{G}(D \times D)$. It is not necessary, but, as was already seen in the binary case, Example 2, a great deal of variability is required.

3.3.1. *Necessary Variability in the Set of Conditional Distributions.* Evaluating how a forecaster makes tradeoffs between a given $\hat{y}$ and another $\hat{y}'$ requires that $\hat{y}$ at least be in the closure of the set of observed forecasts.[8]

Lemma 6. *For a compact D and $\mathcal{S} \subset \Delta(D)$, $\mathrm{cl}\,(Br\,(\mathcal{S} \mid D, u)) = D$ for all $u \in \mathcal{G}(D \times D)$ iff for all $\hat{y} \in D$, $\delta_{\hat{y}} \in \mathrm{cl}\,(\mathcal{S})$.*

[8]In case of a discrete D, this is the requirement that every $\hat{y}$ be observed as a forecast.

To give some sence of the richness that this entails, in the convex D case, if $\mathcal{S}$ is dominated (say by Lebesgue measure), then $\mathcal{S}$ must be a complete class of distributions.

Proof. The definition of u having *nbcc* directly implies that if $\delta_{\hat{y}} \in \mathcal{S}$ for all $\hat{y} \in D$, then $Br\,(\mathcal{S} \mid D, u) = D$. The upper hemicontinuity of the correspondence $p \mapsto Br\,(p \mid D, u)$ and its single-valuedness at $\delta_{\hat{y}}$ imply that if $\rho(p_n, \delta_{\hat{y}}) \to 0$, then $d_H(Br\,(p_n \mid D, u)\,, \hat{y}) \to 0$.

Now suppose that for all $u \in \mathcal{C}_{nbcc}(D \times D)$, $Br\,(\mathcal{S} \mid D, u) = D$ for some closed $\mathcal{S} \subset \Delta(D)$, and that $\delta_{\hat{y}} \notin \mathcal{S}$. Since $\mathcal{S}$ is closed, for some $\epsilon > 0$, $B^{\rho}_{\epsilon}(\delta_{\hat{y}}) \cap \mathcal{S} = \varnothing$. This means that $p(B_{\epsilon}(\hat{y})) < \epsilon$ for all $p \in \mathcal{S}$. The utility functions $u_n(\hat{y}, y) = -e^{-\eta|\hat{y}-\hat{y}_b|}|\hat{y} - y|^{\beta}$, $\beta > 1$, belong to $\mathcal{G}(D \times D)$ (because they are strictly concave in $\hat{y}$ for each fixed y), and have *nbcc*. For sufficiently large n, $Br\,(p \mid D, u_n)$ does not contain $\hat{y}$ for any p satisfying $p(B_{\epsilon}(\hat{y})) < \epsilon$. ∎

3.3.2. *Differences on Smaller Sets of Probabilities.* The definition of potential identifiability seems to require observation of $Br\,(p \mid F, u)$ for every $p \in \Delta(F)$. A universal separability result is behind needing only a dense set in Theorem 1.

Lemma 7. *For $u \in \mathcal{C}_{nbcc}(D \times D)$ and F a compact subset of D, $Br\,(p \mid F, u) \neq Br\,(p \mid F, v)$ for some $p \in \Delta(F)$ iff there exists a non-empty open $G \subset \Delta(F)$ with $Br\,(p \mid F, u) \cap Br\,(p \mid F, v) = \varnothing$ for all $p \in G$.*

In particular, the difference between correspondences $p \mapsto Br\,(p \mid F, u)$ and $p \mapsto Br\,(p \mid F, v)$ are universally separable — any countable dense set of p will contain an element p' with $Br\,(p' \mid F, u) \cap Br\,(p' \mid F, v) = \varnothing$.
Proof: Suppose that $Br\,(p \mid F, u) \cap Br\,(p \mid F, v) = \varnothing$ for all p in a non-empty open $G \subset \Delta(F)$. Since F is compact and both u and v are continuous, neither $Br\,(p \mid F, u)$ nor $Br\,(p \mid F, v)$ are empty. Therefore $Br\,(p \mid F, u) \cap Br\,(p \mid F, v) = \varnothing$ implies that $Br\,(p \mid F, u) \neq Br\,(p \mid F, v)$ for any $p \in G$.

Now suppose that for some $p \in \Delta(F)$, $Br\,(p \mid F, u) \neq Br\,(p \mid F, v)$. Interchanging u and v if necessary, there exists $\hat{y} \in Br\,(p \mid F, u)$ with $d(\hat{y}, Br\,(p \mid F, v)) = 2\epsilon > 0$. By upper-hemicontinuity, for some $\delta > 0$ and every $q \in B^{\rho}_{\delta}(p)$, $Br\,(q \mid F, v) \subset [Br\,(p \mid F, v)]^{\epsilon}$. Since $u \in \mathcal{C}_{nbcc}$, for every $\eta \in (0, 1)$ and $q_{\eta} = (1-\eta)p + \eta\delta_{\hat{y}}$, $Br\,(q_{\eta} \mid F, u) = \{\hat{y}\}$. Pick η such that $q_{\eta} \in B^{\rho}_{\delta}(p)$. By upper-hemicontinuity again, for some $\delta' > 0$, $B^{\rho}_{\delta'}(q_{\eta}) \subset B^{\rho}_{\delta}(p)$ and for every $p' \in B^{\rho}_{\delta'}(q_{\eta})$, $d_H(\{\hat{y}\}, Br\,(p' \mid F, u)) < \epsilon$. Therefore, for every p' in the ρ-open set $B^{\rho}_{\delta'}(q_{\eta})$, $Br\,(p' \mid F, u)$ and $Br\,(p' \mid F, v)$ are disjoint. ∎

3.3.3. *Variability in the Set of Allowable Forecasts.* In consumer demand theory, preference relations are defined by their behavior on two point sets. Restricting the $F \subset D$ that the forecaster must choose from to have only two points will not work in our context.

Example 5 For $D = \{1, 2, 3\}$ and any $r > 0$, consider the utility function

$$u_r(\hat{y}, y) = \begin{bmatrix} 0 & -1 & -10 \\ -1 & 0 & -r \\ -10 & -r & 0 \end{bmatrix} \tag{6}$$

where, as before, m, n entry in the matrix corresponds to the utility of $(\hat{y}, y) = (m, n)$. For any $r, r' > 0$ and two point $F = \{y_1, y_2\} \subset D$ and any $p \in \Delta(F)$, $Br\,(p \mid F, u_r) = Br\,(p \mid F, u_{r'})$, even though changes in r change when one would make which forecast if $F = D$. ∎

The last example showed that we must have subsets of D with cardinality larger than 2. The next example shows that we must have subsets of the larger sets. It hinges on the existence of a forecast so much better than the others that it swamps the tradeoffs between them.

Example 6 For $D = \{1, 2, 3, 4\}$ and any $r, s > 0$, consider the utility function

$$u_{r,s}(\hat{y}, y) = \begin{bmatrix} 0 & -100 & -100 & -100 \\ -100 & 0 & -r & -100 \\ -100 & -s & 0 & -100 \\ -1 & -1 & -1 & 0 \end{bmatrix} \tag{7}$$

The forecasts $\hat{y} = 1, 2, 3$ are not dominated by $\hat{y} = 4$, but they are 'nearly' dominated. For all but a small set of $q \in \Delta(D)$ near the vertices δ_1, δ_2, and δ_3, $Br\,(q \mid D, u) = \{4\}$. Further, this set of best responses does not vary as r and s vary. Thus, best forecasts given $q \in \Delta(D)$ are not sufficient to recover information about the values of r and s. If the forecaster is restricted to making forecasts in the set $F = \{1, 2, 3\}$, then different values of r and s have a large impact on $p \mapsto Br\,(p \mid F, u_{r,s})$, $p \in \Delta(F)$. ∎

The following parametric class of utility functions on a convex D have the property that for any $p \in \Delta(D)$, their best response in the set D is the same, though no two elements of the class are strictly equivalent and the preferences are different on finite sets.

Example 7 For $\theta > 0$, consider the utility function

$$u(\hat{y}, y; \theta) = \frac{1}{1+\theta}\hat{y}^{-(1+\theta)}y - \frac{1}{2+\theta}\hat{y}^{-(2+\theta)}y^2$$

on $D \times D$ where $D = (0, 1]$.

For $p \in \Delta(D)$, the forecaster's problem is

$$\max_{\hat{y}\in(0,1]} \int u(\hat{y}, y)\, dp(y) = \frac{1}{1+\theta}\hat{y}^{-(1+\theta)}EY - \frac{1}{2+\theta}\hat{y}^{-(2+\theta)}EY^2 \tag{8}$$

where Y is a random variable with distribution p. Setting the first order conditions equal to 0 yields

$$-\hat{y}^{-(2+\theta)}EY + \hat{y}^{-(3+\theta)}EY^2 = 0, \tag{9}$$

which one solves for $\hat{y}^* = EY^2/E^Y$. The second order conditions are

$$(2+\theta)\hat{y}^{-(3+\theta)}EY - (3+\theta)\hat{y}^{-(4+\theta)}EY^2 < 0, \tag{10}$$

satisfied if $\frac{2+\theta}{3+\theta}\hat{y} < EY^2/EY$, i.e. if $\frac{2+\theta}{3+\theta} < 1$. If $p = \delta_r$, then $EY^2/E^Y = r$, so this class of utility functions has *nbcc*.

As can be directly observed, the canonical form of $u(\hat{y}, y; \theta)$ is not a linear multiple of $u(\hat{y}, y; \theta')$ if $\theta \neq \theta'$, so no distinct pair of θ's lead to strictly equivalent preferences. The utility functions differ in their best responses when $F \subset D$, e.g. if $F = \{0.5, 1\}$, a little algebra shows that for $p = \frac{1}{2}\delta_{0.5} + \frac{1}{2}\delta_1$, $\int u(0.5, y; \theta)\, dp(y) > \int u(1, y; \theta)\, dp(y)$ is positive for small θ, and the inequality reverses for large θ. ∎

It is worth bserving how delicate the results in this example are — small changes in the parametric specification of the the utility functions would obviate it.

4. Identification for Parametric Classes

The first part of this section define the parametric classes of utility function that we study and summarizes the identification results. The results concern weak identification, identification by fitting best response curves when all covariates are observed, and the FOC approach to identification when some covariates are not observed by the econometrician.

For simplicity, we give the results for parametrized classes of utility functions in the case $D = \mathbb{R}$, the utility functions $u(\hat{y}, y)$ are of the form $u(y - \hat{y})$, i.e. depend on the error in the forecast only.[9] Thus, we say that $u(\cdot)$ is best response equivalent to $v(\cdot)$ if the associated functions $u^\circ(\hat{y}, y) := u(y - \hat{y})$ and $v^\circ(\hat{y}, y) := v(y - \hat{y})$ are.

[9] At the end of the section, we briefly sketch the analysis and results in other cases.

4.1. **Parametrizations.** We assume throughout that the conditional distributions p have densities satisfying three regularity conditions: for some probability μ having a smooth density with respect to Lebesgue measure, each density $p_{x,x'}$ is (a) smooth, (b) strictly positive, and (c) belongs to $L^2(\mathbb{R}, \mathcal{R}, \mu)$. This implies that each p_x has the same properties.

Further, we assume that all utility functions $u(\cdot)$ also belong to $L^2(\mathbb{R}, \mathcal{R}, \mu)$ so that $\int u(y - \hat{y})\, dp_{x,x'}(y)$ is well-defined. At a fairly large cost in the complexity of the arguments, we could change the "smooth" in most of the following to "locally Lipschitz."

Definition 9. *A* ***parametrization*** *is a smooth function $\theta \mapsto u_\theta$ from an open neighborhood of the compact* ***parameter set*** $\Theta \subset \mathbb{R}^K$ *to $L^2(\mathbb{R}, \mathcal{R}, \mu)$ with norm $\|\cdot\|_2$. For a given Θ, the set of all parametrizations is denoted* $\mathrm{Par}(\Theta) = \mathrm{Par}(\Theta; L^2(\mathbb{R}, \mathcal{R}, \mu))$*, and the set of parametrizations taking values in $T \subset L^2(\mathbb{R}, \mathcal{R}, \mu)$ is denote* $\mathrm{Par}(\Theta; T)$*. The distance between parametrizations $u, v \in \mathrm{Par}(\Theta)$ is the C^1-norm,*

$$d(u, v) := \max_\theta \|u_\theta - v_\theta\|_2 + \sum_{k=1}^{K} \|D_{\theta_k} u(\theta) - D_{\theta_k} v(\theta)\|_2.$$

Because μ has a smooth density, the functions $u(\cdot; \theta)$ from $\mathbb{R}$ to $\mathbb{R}$ need not be smooth for the function $\theta \mapsto u_\theta$ from Θ to $L^2(\mathbb{R}, \mathcal{R}, \mu)$ to be.

4.2. **Identifying Parameters Within a Parametrization.** The first question to ask of a parametrization is whether it is weakly identified, that is, whether or not $\theta \neq \theta'$ implies that $u_\theta \not\sim_{Br} u_{\theta'}$. As seen in Example 7, a parametrization may be weakly identified even if forecasts the unrestricted set D provide no identifying power at all. Theorem 2 a totally large set of parametrizations are weakly identified. Theorem 3 goes further, and shows that a totally large set of parametrizations are identified by the unrestricted forecasts.

When the forecaster uses covariates (X, X') with X' not observed by the econometrician, then, provided that $D_x u(x; \theta)$ exists almost everywhere with respect to Lebesgue measure, one can infer θ from the forecaster first order conditions (FOC),

$$E\left(D_{\hat{y}} u(Y - \hat{y}; \theta) | X, X'\right) = 0 \text{ a.e.} \tag{11}$$

By iterated expectations, $E\left(D_{\hat{y}} u(Y - \hat{y}; \theta) | X\right) = 0$. Provided $\sigma(p_X)$ is sufficiently rich, this locally identifies θ, but global identification happens for a set of parametrizations that is neither totally large nor totally small.

4.3. **Potential Identifiability of Parametrizations.** Here we suppose that $m = 0$, i.e. the econometrician observes all of the covariates used in producing the forecast.

Definition 10. *A parametrization is* ***weakly identified*** *if for all* $\theta \neq \theta' \in \Theta$, $u_\theta \not\sim_{Br} u_{\theta'}$

Typically, parameter sets Θ are connected (even convex). In this case, the set $u_\Theta := \{u_\theta : \theta \in \Theta\}$ is connected. Therefore, if T is not a connected set, the set $\mathrm{Par}(\Theta; T)$ may be very restricted.

Theorem 2. *If* $T \subset \mathcal{G} \cap L^2(\mathbb{R}, \mathcal{R}, \mu)$ *is an infinite dimensional, topologically complete, convex cone, then the set of weakly identified elements of* $\mathrm{Par}(\Theta; T)$ *is totally large.*

Proof: Pick $v_1, \ldots, v_J$ linearly independent points in the algebraic interior of T, $J \geq 2K + 2$, let V be the span of the J points, and set $v = \sum_j v_k$. For each $n \in \mathbb{N}$, let Par_n denote the set of parametrizations with $\mathrm{proj}_V(\frac{1}{n}u_\Theta + (1 - \frac{1}{n})v) \subset V \cap T$. Because v is in the algebraic interior of T, $\mathrm{Par}(\Theta; T) = \cup_n \mathrm{Par}_n$. The proof will be complete once we show that for all $n \in \mathbb{N}$, all but a totally small subset of Par_n is weakly identified.

Because $T \subset \mathcal{G}$, to show weak identifiability of a parametrization, it is sufficient to show that it is an embedding. For any parametrization $\theta \mapsto u_\theta$ in Par_n, the mapping $\theta \mapsto f_{n,\theta} := \mathrm{proj}_V(\frac{1}{n}u_\theta + (1 - \frac{1}{n})v)$ is smooth, as is the mapping $\theta \mapsto g_{n,\theta} := f_{n,\theta}/\|f_{n,\theta}\|_2$. For each θ, $g_{n,\theta}$ belongs to the set of points in $V \cap T$ having norm 1, and $J - 1$ dimensional manifold. If $\theta \mapsto u_\theta$ is not an embedding, then $\theta \mapsto g_{n,\theta}$ cannot be an embedding. However, since $J - 1 \geq 2K + 1$, the open denseness of the set of parametrizations that have $\theta \mapsto g_{n,\theta}$ being an embedding follows from classical results related to Whitney's immersion theorem (e.g. Bröcker and Jänich, 1982, Lemma 7.6 (p. 65)), and the prevalence result from Kaloshin (1997, Lemma 1) because smooth maps are locally Lipschitz. ∎

4.4. **Distance Minizing Identification.** Since the (X_t, Y_{t+1}) are observed, the mapping $p_x \mapsto f(p_x)$ is identified. For any parametrization, $\theta \mapsto u_\theta$, we define $\hat{y}_\theta(p_x) = Br\,(p_x \mid D, u_\theta)$. The distance minimizing approach solves the problem

$$\min_{\theta \in \Theta} \|f - \hat{y}_\theta\| 2. \tag{12}$$

For a fixed parametrization, the set of possible f leading to multiple optima in (12) is $\{f \in L^2(\sigma(p_X)) : \#\,\mathrm{argmin}_{\theta \in \Theta} \|f - \hat{y}_\theta\| \geq 2\}$.

Theorem 3. *If $T \subset \mathcal{G} \cap L^2(\mathbb{R}, \mathcal{R}, \mu)$ is an infinite dimensional, topologically complete, convex cone, and the dimensionality of $L^2(\sigma(p_X)) \geq K$, then the following dual pair of results hold for* $\mathrm{Par}(\Theta; T)$.

A. *There is a totally large set of parametrizations $\theta \mapsto u_\theta$ that are weakly identified, and for which the set*

$$\{f \in L^2(\sigma(p_X)) : \# \operatorname*{argmin}_{\theta \in \Theta} \|f - \hat{y}_\theta\| \geq 2\}$$

is totally small.

B. *For any $f \in L^2(\sigma(p_X))$, the set of $\theta \mapsto u_\theta$ in* $\mathrm{Par}(\Theta; T)$ *for which* $\# \operatorname{argmin}_{\theta \in \Theta} \|f - \hat{y}_\theta\| \geq 2$ *is totally small.*

Proof. A. Since $\theta \mapsto u_\theta$ takes values in $\mathcal{G}$ and is weakly identified, it is an embedding. The mapping $\theta \mapsto \hat{y}_\theta$ is continuous. Since the dimensionality of $L^2(\sigma(p_X))$ is at least K, then for all but a totally small set of parametrizations, $\theta \mapsto \hat{y}_\theta$ is also an embedding. Pick an arbitrary parametrization in this totally large set.

For $\epsilon > 0$, let T_ϵ be the compact set of (θ, θ') with $\|\theta - \theta'\| \geq \epsilon$. By compactness and continuity, the set of $K_\epsilon = \{(\hat{y}_\theta, \hat{y}_{\theta'}) \in L^2 \times L^2 : (\theta, \theta') \in T_\epsilon\}$ is compact. Therefore the set of F_ϵ of $(f, f) \in L^2 \times L^2$ that minimize the distance to K_ϵ is compact because orthogonal projection is continuous. Being compact in L^2 it has no interior and is finitely shy. Finally, the set of f such that there exists $\theta \neq \theta'$ with $d(f, \hat{y}_\theta) = d(f, \hat{y}_{\theta'}) \geq d(f, \hat{y}_\Theta)$ is a subset of $\cup_\epsilon F_\epsilon$ where the union is taken over the countable set of strictly positive rational ϵ.

B. Fix an arbitrary $f \in L^2(\sigma(p_X))$. It is immediate that the set of parametrizations $\theta \mapsto u_\theta$ such that $\# \operatorname{argmin}_{\theta \in \Theta} \|f - \hat{y}_\theta\| \geq 2\}$ is closed and has no interior. Since the dimensionality of $L^2(\sigma(p_X))$ is at least K, there are K different $p_{x_k} \in \Delta(Y)$ with disjoint open neighborhoods having positive probability. A TINY LAST STEP IS NEEDED HERE. ∎

Part A. of the previous immediately yields the following.

Corollary 3.1. *Under the conditions of Theorem* 3*, for a totally large set of parametrizations in* $\mathrm{Par}(\Theta; T)$*, if $f = \hat{y}_{\theta^\circ}$, then θ° is the unique solution to* $\min_{\theta \in \Theta} \|f - \hat{y}_\theta\|_2$.

In summary, generic parametrizations produce a unique best fitting forecast function whether or not the parametrization bears any resemblance to the forecaster's utility function, but if the forecaster's utility function belongs to the parametrization, then, generically, the utility function is recovered.

4.5. **FOC Local Identification.** By the law of iterated expectations and equation (11), we are interested in finding a value of θ for which

$$E\left(D_{\hat{y}}u(Y-\hat{y};\theta)|X\right)=0. \tag{13}$$

Provided that $u(\cdot;\theta)$ is differentiable enough and $\sigma(p_X)$ is rich enough, the solution to (13) is locally unique, i.e. locally identified. To handle the differentiability issue, we suppose that the utility functions take values in the Sobolev space $W^{1,2}(\mathbb{R},\mathcal{R},\mu)$, which is the set of elements of $L^2(\mathbb{R},\mathcal{R},\mu)$ with weak derivatives of order 1.

Theorem 4. *If $T\subset\mathcal{G}\cap W^{1,2}(\mathbb{R},\mathcal{R},\mu)$ is an infinite dimensional, topologically complete, convex cone, and the dimensionality of $L^2(\sigma(p_X))\geq K$, then for a totally large subset of $\mathrm{Par}(\Theta;T)$, the solutions to $E\left(D_{\hat{y}}u(Y-\hat{y};\theta)|X\right)=0$ are locally unique.*

Proof. Pick K linearly independent elements, ψ_k, $k=1,\ldots,K$, of $L^2(\sigma(p_X))$, and consider the K smooth equations

$$h_k(\theta)=\int D_{\hat{y}}u(Y-\hat{y};\theta)\psi_k(p_X)\,dP, \tag{14}$$

$k=1,\ldots,K$. Since the equations are depend smoothly on the parametrization, Kaloshin (1997, Lemma 1) delivers the shyness, and textbook standard results (e.g. Bröcker and Jänich, 1982, Lemma 7.6 (p. 65)), deliver the Baire smallness. ∎

More than local uniqueness cannot be guaranteed without addition of extra structure to the equations in (14). In the one dimensional case $K=1$, Elliot et. al. add monotonicity in θ of the expected value of the derivative. Even in this case, the set of parametrizations for which the expected value of the derivative crosses 0 more than once has non-empty interior. More generally, see Gale and Nikaido (1965) for sufficient conditions for uniqueness in (14).

Appendix: Proofs

For shyness and Baire smallness to be useful, we must show that the space $\mathcal{C}_{nbcc}(D \times D)$ is separable and topologically complete.

Lemma 8. *$\mathcal{C}_{nbcc}(D \times D)$ is separable and topologically complete.*

Proof. If D is finite, separability and topological completeness aretrivial, though the proof for general strongly σ-compact D can also be applied. The outline of the proof is as follows:

(a) define a metric on $\mathcal{C}_{nbcc}(D \times D)$ for compact D,
(b) show that it is topologically equivalent to the sup norm metric d_∞ for compact D,
(c) show that $\mathcal{C}_{nbcc}(D \times D)$ is complete in the metric for compact D, and
(d) show that this carries over to strongly σ-compact D.

(a) Defining a metric for $\mathcal{C}_{nbcc}(D \times D)$ for compact D. For $m \in \mathbb{N}$ and $u \in \mathcal{C}_{nbcc}(D \times D)$, let $r_m(u) = \min_{|\hat{y}-y| \geq 1/m} |u(\hat{y}, y)|$, and for $u, v \in \mathcal{C}_{nbcc}(D \times D)$, let $f_m(u, v) = \min\left\{1, \left|\frac{1}{r_m(u)} - \frac{1}{r_m(v)}\right|\right\}$. Define

$$d(u, v) = d_\infty(u, v) + \sum_m \frac{1}{2^m} f_m(u, v).$$

Remarks: (*i*) $u(\hat{y}, y) = 0$ iff $\hat{y} = y$ implies $r_m(u) > 0$, so d is well defined. (*ii*) It is straightforward to verify that d is a metric. (*iii*) $|r_m(u) - r_m(v)| \leq d_\infty(u, v)$. To see this, note that the set $\{(\hat{y}, y) \in D \times D : |\hat{y} - y| \geq 1/m\} \subset \mathbb{R}^2$ is compact; hence, we can choose from it $(\hat{y}_0, y_0)$ such that $|v(\hat{y}_0, y_0)| = r_m(v)$. Further, we can write

$$\begin{aligned} r_m(u) \leq |u(\hat{y}_0, y_0)| &\leq \Big| |u(\hat{y}_0, y_0)| - |v(\hat{y}_0, y_0)| \Big| + |v(\hat{y}_0, y_0)| \\ &\leq |u(\hat{y}_0, y_0) - v(\hat{y}_0, y_0)| + |v(\hat{y}_0, y_0)| \\ &\leq d_\infty(u, v) + r_m(v). \end{aligned}$$

Reversing the roles of u and v gives (*iii*).

(b) Topological equivalence for compact D. Since $d(u, v) \geq d_\infty(u, v)$, if $d(u_n, u) \to 0$, $d_\infty(u_n, u) \to 0$. Suppose that $d_\infty(u_n, u) \to 0$ and pick $\epsilon > 0$. We must show that there exists $N \in \mathbb{N}$ such that for all $n \geq N$, $d(u_n, u) < \epsilon$. Pick N_1 such that for all $n \geq N_1$, $d_\infty(u_n, u) < \epsilon/3$. Pick M such that $\sum_{m>M} \frac{1}{2^m} < \epsilon/3$. Finally, using property (*iii*) above, pick N_2 such that for all $n \geq N_2$ and for all $m \leq M$, $f_m(u_n, u) < \epsilon/(3M)$. For all $n \geq \max\{N_1, N_2\}$,

$$d(u_n, u) = d_\infty(u_n, u) + \sum_{m \leq M} \frac{1}{2^m} f_m(u_n, u) + \sum_{m>M} \frac{1}{2^m} f_m(u_n, u) < \frac{\epsilon}{3} + \frac{\epsilon}{3} + \frac{\epsilon}{3}.$$

(c) d-completeness for compact D. Let u_n be a d-Cauchy sequence in $\mathcal{C}_{nbcc}(D \times D)$, hence a d_∞-Cauchy sequence in $C(D \times D)$. Since C

is d_∞-complete, there exists a $u \in C$ such that $d_\infty(u_n, u) \to 0$. All that is left to show is that $u \in \mathcal{C}_{nbcc}$. Since $u_n(\hat{y}, y) \leq 0$ for all $(\hat{y}, y)$, and $d_\infty(u_n, u) \to 0$ implies $u_n(\hat{y}, y) \to u(\hat{y}, y)$ pointwise, it follows that $u(\hat{y}, y) \leq 0$ for all $(\hat{y}, y)$. Suppose that $u \notin \mathcal{C}_{nbcc}$, i.e. $u(\hat{y}_0, y_0) = 0$ for some $\hat{y}_0 \neq y_0$. Let m_0 be the smallest value of m with $|\hat{y}_0 - y_0| \geq 1/m$. As $u_n(\hat{y}_0, y_0) \to 0$, it follows that $r_m(u_n) \to_n 0$ for all $m \geq m_0$; in fact, $\sup_{m \geq m_0} r_m(u_n) = r_{m_0}(u_n) \to_n 0$. Therefore, for any fixed integer $n \in \mathbb{N}$ there exists $K \in \mathbb{N}$ so that $f_m(u_n, u_{n+k}) = 1$ for all $k \geq K$ and $m \geq m_0$. Hence, for all k large enough,

$$d(u_n, u_{n+k}) \geq \textstyle\sum_{m \geq m_0} \frac{1}{2^m} f_m(u_n, u_{n+k}) \geq \frac{1}{2^{m_0}},$$

contradicting u_n being a d-Cauchy sequence.
(d) Since D is countable union of open sets with compact closure, it can be expressed as a countable increasing union of open sets with compact closure, D_n. For each of these compact D_n, form the metric as above, labeled D_n. In the metric $\rho := \sum_n \frac{1}{2^n} \min\{d_n, 1\}$, $\mathcal{C}_{nbcc}(D \times D)$ is topologically complete. ∎

Lemma 9. *$\mathcal{G}(D \times D)$ is totally large.*

Proof. We proceed from finite D to compact D to strongly σ-compact D.

We first show that for finite D, $\mathcal{C}_{nbcc} \setminus \mathcal{G}$ is negligible. Given $\#D = M$, we must show that the closure of $\mathcal{C}_{nbcc} \setminus \mathcal{G}$ has Lebesgue measure 0 as a subset of (the negative orthant of) $\mathbb{R}^{M^2 - M}$. If $M = 2$ then $\mathcal{C}_{nbcc} \setminus \mathcal{G}$ is empty. Let $M \geq 3$, and pick an arbitrary three point subset $F = \{y_1, y_2, y_3\}$ from D. Restricted to $F \times F$, any $u \in \mathcal{C}_{nbcc}$ can be represented by 6 negative numbers, a through f, ordered clockwise as

$\hat{y} \downarrow$			
y_1	0	a	b
y_2	f	0	c
y_3	e	d	0
$y \rightarrow$	y_1	y_2	y_3

According to Definition 7, if u fails to be in $\mathcal{G}$, then there must be a $y_i \in F$ such that for all $p \in \Delta(F)$ with $p(y_i) = 0$ and $Br\,(p \mid F, u) = F$. Suppose, for the sake of concreteness, that $y_i = y_2$, so that $p = (\alpha, 0, (1 - \alpha))$ for some $\alpha \in (0, 1)$. Note that $Br\,(p \mid F, u) = F$ iff

$b(1-\alpha) = f\alpha + c(1-\alpha) = e\alpha$, equivalently, iff

$$\begin{bmatrix} 0 & (1-\alpha) & (\alpha-1) & 0 & 0 & -\alpha \\ 0 & 0 & (1-\alpha) & 0 & -\alpha & \alpha \end{bmatrix} \begin{bmatrix} a \\ b \\ c \\ d \\ e \\ f \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \tag{15}$$

For each $\alpha \in [0,1]$, let S_α be the set of $(a,b,c,d,e,f) \in \mathbb{R}^6$ satisfying (15). Since the 2×6 matrix is of full row rank for all α, each S_α is a 4-dimensional linear subspace of $\mathbb{R}^6$. Since α smoothly parametrizes the S_α, $S := \cup_\alpha S_\alpha$ is a closed manifold of dimension at most 5. Since $\mathcal{C}_{nbcc}$ is convex and has non-empty interior in $\mathbb{R}^6$, $S \cap \mathcal{C}_{nbcc}$ is negligible. This argument, repeated two more times, covers the cases where $y_i = y_1$ and $y_i = y_3$. Finally, the result follows as there are only finitely many subsets of D that are of size 3, and a finite union of negligible sets is negligible.

We now show that for infinite compact D, $\mathcal{C} \setminus \mathcal{G}$ is totally small. Let D' be an arbitrary, countable dense subset of D (which exists because D is compact). There is a countable class of three point sets, $F \subset D'$. For each F, the previous step showed that there are at most 3 closed 5-dimensional manifolds for which u fails the conditions in Definition 7. Such a set is necessarily totally small. By Lemma 8, $\mathcal{C}_{nbcc}$ is topologically complete. This in turn implies that the countable union of totally small sets is totally small.[10]

Finally, for strongly σ-compact D, apply the previous result to each D_n, where D_n is an increasing sequence of open sets with compact closure and $D = \cup_n D_n$. ∎

C. Proof of Theorem 1. By Lemmas 2, 3, 4, 8, and 9, all that is left to show is that if $u \in \mathcal{G}(D \times D)$, then for all $v \in \mathcal{C}_{nbcc}$, $[u \sim_{Br} v] \Leftrightarrow [u \sim_{aff} v]$.

Outline of the proof: (**I**) Show result for $\#D = 3$; (**II**) use induction to show result for $\#D = M < \infty$; (**III**) use continuity and denseness of D' in D to show result when D is a general compact set.

Additional notation: For $F \subset D$ and $u \in \mathcal{C}$, we will denote the restriction of u onto $F \times F$ by $u_{|F}$. It follows immediately from the definition of best response equivalence that $u \sim_{Br} v$ implies $u_{|F} \sim_{Br} v_{|F}$.

[10] See Anderson and Zame (2001) for the relevant facts and definitions about shy subsets of convex sets.

Part (**I**): $\#D = 3$

To keep the notation simple, let $D = \{1, 2, 3\}$. Each $u \in \mathcal{C}_{nbcc}(D \times D)$ can be represented as 6 negative numbers, a through f, ordered clockwise as

$\hat{y} \downarrow$			
1	0	a	b
2	f	0	c
3	e	d	0
$y \rightarrow$	1	2	3

Further, each $u \in \mathcal{C}_{nbcc}(D \times D)$ can be classified according to how it "behaves" for $p \in \Delta(D)$ with exactly one zero component. There are three mutually exclusive cases:

Case 1:: For all $i \in D$, if $p(i) = 0$, then $i \notin Br\,(p \mid D, u)$. In this case one might say that the utility function "ignores irrelevant alternatives". This can be violated in two ways:

Case 2:: There is an $i \in D$ such that for all α in some nonempty open interval $(r, s) \subset (0, 1)$, if $p(i) = 0$, $p(j) = \alpha$ and $p(k) = 1 - \alpha$, then $Br\,(p \mid D, u) = \{i\}$ and, further, $Br\,(p \mid D, u) = \{i, j\}$ when $\alpha = s$ and $Br\,(p \mid D, u) = \{i, k\}$ when $\alpha = r$.

Case 3:: There is an $i \in D$ such that if $p(i) = 0$ then $Br\,(p \mid D, u) = \{1, 2, 3\}$.

The third case is ruled out in the definition of the subset $\mathcal{G}$; as we assume $u \in \mathcal{G}$, we need to consider the first two cases only.

Discussion of Case 1: Using the definition of Case 1, the *nbcc* property, and the continuity of expected utility in the components of the probability measure, it is easy to show that for each u falling under this case there exist distributions satisfying

$p = (p_1, p_2, 0)$ with $Br\,(p \mid D, u) = \{1, 2\}$,
$q = (q_1, 0, q_2)$ with $Br\,(q \mid D, u) = \{1, 3\}$,
$r = (r_1, r_2, r_3)$ with $Br\,(r \mid D, u) = \{1, 2, 3\}$,
$s = (0, s_1, s_2)$ with $Br\,(s \mid D, u) = \{2, 3\}$,

where p_1, p_2, q_1, q_2, etc., are all strictly positive. We will show that the indifference conditions implicit in these best response sets determine u up to a multiplicative constant. Therefore, if $v \in \mathcal{G} \subset \mathcal{C}_{nbcc}$ is another utility function with $u \sim_{Br} v$ then we must have $u \sim_{aff} v$.

Let us normalize a to -1. Combining this normalization with the five equalities that come from the indifference conditions for p, q, r and s (there are two conditions associated with r), we obtain the following six linear equations in the six unknowns, a, b, c, d, e and f:

(16)

$$\begin{bmatrix} -1 & 0 & 0 & 0 & 0 & 0 \\ p_2 & 0 & 0 & 0 & 0 & -p_1 \\ 0 & q_2 & 0 & 0 & -q_1 & 0 \\ r_2 & r_3 & 0 & -r_2 & -r_1 & 0 \\ 0 & 0 & r_3 & -r_2 & -r_1 & r_1 \\ 0 & 0 & s_2 & -s_1 & 0 & 0 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \\ d \\ e \\ f \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \quad \begin{array}{l} \text{normalization} \\ Br\,(p \mid D, u) = \{1, 2\} \\ Br\,(q \mid D, u) = \{1, 3\} \\ Br\,(r \mid D, u) = \{1, 2, 3\} \\ Br\,(r \mid D, u) = \{1, 2, 3\} \\ Br\,(s \mid D, u) = \{2, 3\} \end{array}$$

We will show that the determinant of the 6×6 coefficient matrix is non-zero, meaning that there is exactly one *normalized* $u \in \mathcal{C}_{nbcc}$ with the best response sets determined by p, q, r and s. To do this, we first expand into co-factors along the top row, which has only one non-zero entry, -1. In the remaining 5×5 matrix, we again expand into co-factors along the top row, which has only one non-zero entry, $-p_1$. Thus, we arrive at needing to show that

$$\det \begin{bmatrix} q_2 & 0 & 0 & -q_1 \\ r_3 & 0 & -r_2 & -r_1 \\ 0 & r_3 & -r_2 & -r_1 \\ 0 & s_2 & -s_1 & 0 \end{bmatrix} = q_2 \det \begin{bmatrix} 0 & -r_2 & -r_1 \\ r_3 & -r_2 & -r_1 \\ s_2 & -s_1 & 0 \end{bmatrix} - r_3 \det \begin{bmatrix} 0 & 0 & -q_1 \\ r_3 & -r_2 & -r_1 \\ s_2 & -s_1 & 0 \end{bmatrix} \neq 0.$$

After expanding the 3×3 matrices, this is $q_2 r_1 s_1 r_3 + r_3 q_1 r_2 s_2 - q_1 s_1 r_3^2$. Since $r_3 > 0$, we take it out as a common factor so that we need to show $q_2 r_1 s_1 + q_1 r_2 s_2 - q_1 s_1 r_3 \neq 0$. In this last expression, replace q_2 with $(1 - q_1)$, s_2 with $(1 - s_1)$ and rearrange, arriving at needing to show $r_1 s_1 (1 - q_1) + r_2 q_1 (1 - s_1) + r_3 q_1 s_1 \neq 0$. Since each term in this sum is strictly positive, this is indeed the case.

Discussion of Case 2: The strategy of proof is the same as in Case 1. For concreteness, suppose that in the definition of Case 2, $i = 2$. There are five relevant probability distributions:

$p = (p_1, p_2, 0)$ with $Br\,(p \mid D, u) = \{1, 2\}$,
$q = (q_1, 0, q_2)$ with $Br\,(q \mid D, u) = \{1, 2\}$,
$t = (t_1, 0, t_2)$ with $Br\,(t \mid D, u) = \{2\}$ and $Br\,(t \mid F, u) = F$ where $F = \{1, 3\}$,
$r = (r_1, 0, r_2)$ with $Br\,(r \mid D, u) = \{2, 3\}$, and
$s = (0, s_1, s_2)$ with $Br\,(s \mid D, u) = \{2, 3\}$,

where p_1, p_2, q_1, q_2, etc., are all strictly positive.

Again, we normalize a to -1. Combining this normalization with the five equalities that come from the indifference conditions for p, q, r, s and t, we have the following six linear equations in the six unknowns,

a, b, c, d, e and f:

(17)

$$\begin{bmatrix} -1 & 0 & 0 & 0 & 0 & 0 \\ p_2 & 0 & 0 & 0 & 0 & -p_1 \\ 0 & q_2 & -q_2 & 0 & 0 & -q_1 \\ 0 & t_2 & 0 & 0 & -t_1 & 0 \\ 0 & 0 & r_2 & 0 & -r_1 & r_1 \\ 0 & 0 & s_2 & -s_1 & 0 & 0 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \\ d \\ e \\ f \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \quad \begin{array}{l} \text{normalization} \\ Br\,(p \mid D, u) = \{1, 2\} \\ Br\,(q \mid D, u) = \{1, 2\} \\ Br\,(t \mid F, u) = F,\ F = \{1, 3\} \\ Br\,(r \mid D, u) = \{2, 3\} \\ Br\,(s \mid D, u) = \{2, 3\} \end{array}$$

Again, we will show that the determinant of the 6×6 coefficient matrix is non-zero. We first expand into co-factors along the top row, which has only one non-zero entry, -1. In the remaining 5×5 matrix, we expand into co-factors along the third column, which has only one non-zero entry, $-s_1$. In the remaining 4×4 matrix, we expand along the top row, which has only one non-zero entry, $-p_1$. The remaining 3×3 matrix is

$$\begin{bmatrix} q_2 & -q_2 & 0 \\ t_2 & 0 & -t_1 \\ 0 & r_2 & -r_1 \end{bmatrix} \text{ which has determinant } q_2 \left[\begin{vmatrix} 0 & -t_1 \\ r_2 & -r_1 \end{vmatrix} + \begin{vmatrix} t_2 & -t_1 \\ 0 & -r_1 \end{vmatrix} \right] = q_2 \left[t_1 r_2 - t_2 r_1 \right].$$

Combining all of this, the determinant of the 6×6 matrix is $\kappa[t_1 r_2 - t_2 r_1]$, where $\kappa = -s_1 p_2 q_2 < 0$. The term $[t_1 r_2 - t_2 r_1]$ can be re-written as $[t_1(1 - r_1) - (1 - t_1) r_1] = [t_1 - r_1]$. Since $t \neq r$, we know that $t_1 \neq r_1$.

Part (**II**): Induction on $\#D$

The inductive hypothesis is that Theorem 1 holds for $\#D \leq M$, where M is some fixed integer greater than or equal to three. The inductive step is to show that the theorem also holds for $\#D = M + 1$.

Let $D \subset \mathbb{R}$ with $\#D = M + 1$, and suppose that for $u, v \in \mathcal{G} \subset \mathcal{C}_{nbcc}(D \times D)$, $u \sim_{Br} v$. To keep the notation simple, put $D = \{1, 2, \ldots, M, M + 1\}$. Let $F_1 = \{1, \ldots, M\}$ and $F_2 = \{2, \ldots, M + 1\}$. As $u_{|F_1} \sim_{Br} v_{|F_1}$, by the inductive hypothesis there exists $r_1 > 0$ such that $u_{|F_1} = r_1 \cdot v_{|F_1}$. Also, as $u_{|F_2} \sim_{Br} v_{|F_2}$, by the inductive hypothesis there exists unique $r_2 > 0$ such that $u_{|F_2} = r_2 \cdot v_{|F_2}$. By considering the common part of $F_1 \times F_1$ and $F_2 \times F_2$, which includes points at which u and v are both non-zero, we must have $r_1 = r_2 := r$. Therefore, $u(\hat{y}, y) = r \cdot v(\hat{y}, y)$ must hold for all $(\hat{y}, y) \in D \times D$ except possibly at the points $(1, M + 1)$, $(M + 1, 1)$; see Table 1. Replacing F_2 with $F_3 = \{1, 3, 4, \ldots, M, M + 1\} \subset D$ and repeating the arguments above shows that $u(\hat{y}, y) = r \cdot v(\hat{y}, y)$ holds even at these points.

Part (**III**): General compact D

Table 1. $D \times D$

	$(1,1)$	$(1,2)$	...	$(1,M)$	$(1,M+1)$	
$F_1 \times F_1$	$(2,1)$	$(2,2)$	...	$(2,M)$	$(2,M+1)$	
	$\vdots$	$\vdots$		$\vdots$	$\vdots$	
	$(M,1)$	$(M,2)$	...	(M,M)	$(M,M+1)$	$F_2 \times F_2$
	$(M+1,1)$	$(M+1,2)$	...	$(M+1,M)$	$(M+1,M+1)$	

Suppose that for $u, v \in \mathcal{C}_{nbcc}(D \times D)$, $u \sim_{Br} v$. The previous two steps have shown that there exists a unique $r > 0$ such that for all finite $F \subset D'$, $u_{|F} = r \cdot v_{|F}$. This implies that $u_{|D'} = r \cdot v_{|D'}$. Since $D' \times D'$ is dense in $D \times D$ and u and v are continuous, $u = r \cdot v$.

The extension to strongly σ-compact D is immediate.