Testing for Rationality

Mark Dean

Lecture Notes for Fall 2015 PhD Class in Decision Theory - Brown University

Here is an unpleasant fact about 'real life' data. OWC (or GARP) is almost *always* violated. In any actual data set, be it from the laboratory or from the 'real world', individuals will almost certainly fail the relevant axiom. Remember, one mistaken choice, one slip up, and the whole data set will fail OWC. This is problematic, as this is not a very interesting result: if we are going to classify everyone as irrational, then do we throw out all the machinery of economics, possibly due to a very small number of rogue choices? This seems too strong. Therefore, it would be nice to have some measure of how close a particular data set is from satisfying rationality. In this section, we are going to present some tools that try and put a metric on these things. It should be noted that these metrics are somewhat arbitrary, and in general lack a proper statistical grounding. A fruitful area of research might be to put these things on a firmer footing.[1]

1 Definitions

The primitives of the problem are a grand set of *alternatives* Z, a set of *observations* X and a *relation function* $D : X \rightarrow 2^{Z\times Z}$ that characterizes a set of binary relations on Z generated by

[1] See

- Apesteguia, Jose and Ballester, Miguel Angel, (2010), A Measure of Rationality and Welfare, Economics Working Papers, Department of Economics and Business, Universitat Pompeu Fabra, http://econpapers.repec.org/RePEc:upf:upfgen:1220
- Tim Beatty & Ian Crawford, 2010. "How demanding is the revealed preference approach to demand," CeMMAP working papers CWP17/10, Centre for Microdata Methods and Practice, Institute for Fiscal Studies.

For papers that explore this issue.

each observation in X. We call the triple $\{Z, X, D\}$ a *data set*.

As an example, consider the case of a laboratory experiment in which we observe a subject making choices from subsets of Z. Furthermore, assume that we are prepared to say that the chosen object in any set is strictly preferred to all the other available alternatives.[2] In this case we could think of each observation in X as consisting of a tuple (z, A) with $z \in A$ and $A \in 2^Z/\emptyset$., implying that alternative z has been observed as being chosen from the set of alternatives A. The function $D : X \to 2^{Z\times Z}$ would then be defined as the revealed preference relations generated by X:

$$\begin{aligned} \forall \ (z, A) &\in X, \\ D(z, A) &= \{(z, y) \mid y \in A/\{z\}\} \end{aligned}$$

We denote by the binary relation $\succ_x \subset Z \times Z$ the relations generated by the observation $x \in X$, so that $\succ_x = D(x)$. For any $B \subset X$, we define the binary relation $\succ^B$ on Z as

$$z \succ^B w \text{ if, for some } x \in B,\ z \succ_x w.$$

We say that a set of observations $B \subset X$ is *acyclic* if the binary relation $\succ^B$ generated by B contains no cycles.

Notice that, for what follows, we are implicitly treating $\succ$ as a strict preference relation

2 Counting Measures

2.1 The Famulari Measure

The simplest measure of how close a data set is to rationality is a simple counting measure, that works as follows:

1. Calculate the transitive closure of $D(x)$ (let's call it T_x)

[2] Note that this assumption is *not* central to our methodology, which is flexible enough to cope with almost any definition of revealed preference. For example, if we are observing choices from budget sets, we could say that x is revealed preferred to y only if x was chosen when y was available at strictly lower cost (per the Generalized Axiom of Revealed Preference). The precise assumption is defined by the nature of the mapping D.

2. Count up the number of pairs z_1, z_2 in Z such that $z_1 T_X z_2$ and $z_2 \succ^B z_1$ (i.e. z_1 is indirectly revealed preferred to z_2 while z_2 is directly revealed preferred to x_1
3. The proportion of pairs for which this is true is a measure of irrationality.)

Counting measures such as this are easy to calculate[3], and have some nice properties (for example, it is equal to zero if and only the data set is rational). It also works on any arbitrary data set. A major drawback are a lack of robustness to a single bad choice, and a resulting failure to differentiate between cases that we might want to think of as having very different rationality properties. For example, consider the following data set

$$\begin{aligned} C(\{a,b,c,d\}) &= \{a\} \\ C(\{b,c,d\}) &= \{b\} \\ C(\{c,d\}) &= \{c\} \end{aligned}$$

Thus data set is perfectly rational. However, if we add the observation

$$C(\{a,b,c,d\}) = \{d\}^4$$

Then there is indeed a cycle. caused by one bad choice. However, the Famulari measure would score this set as 'maximally' irrational: in the sense that for every pair, we have both a direct revealed preference relation, and the opposite indirect revealed preference relation.

Moreover, if we add the observation

$$C(\{b,c,d,e\}) = \{d\}$$

The Famulari measure does not change, despite the fact that the new data set is clearly 'less' rational that the first

Another problem, which is shared with all the counting measures, is that there is no notion of how 'severe' a particular violation of rationality is: if these cycles are caused by small 'trembles' around an indifference point, then we may not care so much. On the other hand, if they represent 'big' swings in revealed preference, then we may care a lot. This is an issue we will return to below.

[3] The Floyd-Warshall algorithm can calculate the transitive closure of sets relatively easily.

[4] For simplicity, lets assume that we can have multiple observations from the same choice set.

2.1.1 HM-Index

One measure of rationality that does not have the robustness problem described above was proposed by Houtman and Maks [1985]: the size of the largest subset of choice observations that satisfy acyclicality (henceforth the HM index).

Imagine a choice experiment in which subject A exhibits the following behavior:

$$\begin{aligned} C_A(\{x,y\}) &= \{x\} \\ C_A(\{x,y,z\}) &= \{z\} \\ C_A(\{x,z\}) &= \{z\} \\ C_A(\{y,z\}) &= \{y\} \\ C_A(\{x,y,w\}) &= \{w\} \end{aligned}$$

If we assume that choice is synonymous with (strict) revealed preference, then these data are not consistent with acyclicality, as z is revealed preferred to y, while y is revealed preferred to z, which is in turn revealed preferred to x. However, if one were to remove the observation $C_A(\{y,z\}) = \{y\}$, then the resulting system would be consistent with acyclicality.

Now imagine that subject B exhibits the following behavior:

$$\begin{aligned} C_B(\{x,y\}) &= \{x\} \\ C_B(\{x,y,z\}) &= \{z\} \\ C_B(\{x,z\}) &= \{z\} \\ C_B(\{y,z\}) &= \{y\} \\ C_B(\{x,y,w\}) &= \{y\} \end{aligned}$$

This data set is also not consistent with acyclicality. However, in this case one would have to remove *two* observations before subject B's choices are consistent with acyclicality. In this sense, subject B could be described as *less rational* than subject A. This, in essence, is the meaning of the HM index: The HM index of subject A is 4, while for subject B it is 3.

While this measure is not without its problems, (which we will come back to) the HM index can be applied to any set, and .solves the robustness problem we described above. On the downside, the problem is computationally intensive (in fact it is NP-hard), so we are limited to the sets that we can apply it to.[5]

In order to formally define the HM index, we need to define the maximal acyclical set problem (MASP)

Definition 1 *The maximal acyclical set problem (MASP)*[6] *for a data set* $\{Z, X, D\}$ *is the problem of finding the size of a set* $B \subset X$ *such that*

$$\textit{(i) } B \textit{ is acyclic}$$
$$\textit{(ii) if } B' \subset X \textit{ and } \left|B'\right| > |B|\textit{, then } B' \textit{ is not acyclic}$$

In other words, MASP is the problem of finding the size of the largest acyclical subsets of X. Note that the maximal acyclical set may not be unique.

Formally, we define the HM index using the concept of the maximal acyclical set

Definition 2 *The HM index for a data set* $\{Z, X, D\}$ *is a number* M *such that* $M = |A|$*, where* A *is a maximal acyclical set of that data set.*[7]

2.2 Minimal Multiple Rationales

If a data set is cyclical, it cannot be explained perfectly by a single preference order, or rationale. However, it may be that agents have different rationales for different states. For example, the

[5] In Dean and Martin [2011], we introduce an algorithm for finding the size of the largest subset of a choice data set that is consistent with acyclicality. We call this problem the *maximal acyclical set problem*, or MASP. The key to our approach is to take advantage the fact that MASP is equivalent to the *minimum set covering problem* (MSCP), which is well studied in the computer sciences and operations research literature. While the MSCP is also NP-hard, there are a wide variety of methods that are extremely efficient in solving it for practical cases and are included in standard 'solver' software packages (see Caprara, Toth, and Fischetti [2000]). For any choice data set, we can therefore translate the associated MASP into an equivalent MSCP, which can then be solved using one of these software packages. In tests on simulated data, our algorithm can handle data sets about an order of magnitude larger than methods currently used in economics

[6] We assume that X is finite. In this case, we can solve MASP whatever the cardinality of Z. Moreover, acyclicality is enough to guarantee that choices can be rationalized by utility maximization even if Z is uncountable. This is because we can concentrate on the (finite) set of objects $\overline{Z}$ that are chosen in any observation X. If the data is acyclic, we can generate a utility function $u : \overline{Z} \rightarrow \mathbb{R}$ that rationalizes choice between these objects. All remaining alternatives can be assumed to have a utility equal to $\min_{z\in\overline{Z}} u(z) - 1$.

[7] The HM index can be normalized by dividing it by the total number of observations in a data set.

relative ranking of an umbrella and a bicycle may differ depending on whether it is raining or not. If we do not observe these different states, then the resulting choices may appear irrational. This notion was captured by Kalai, Rubinstein, and Spiegler [2002], who introduced the concept of *rationalization by multiple rationales.* A rationale is a preference ordering, and a data set is rationalized by a collection of rationales if all observations are explicable as the maximization of one of the rationales. Thus, choice data that can be rationalized by n rationales can be thought of as being generated by an individual who at any time is in one of n different 'states' and in each state has a different set of preferences. Such an approach has also been applied to the analysis of household-level data, to determine if household choices can be rationalized as preference maximization by one of the members of the household (Deb [2008], Nobibon et al. [2009]) and to determine if households are heterogeneous (Crawford and Pendakur [2008]).

Formally, we define the concept of rationalizing by multiple rationales as follows:

Definition 3 *For a data set $\{Z, X, D\}$, a rationalization by multiple rationales is a set of complete preference relations*[8] *R on Z such that, for every $x \in X$, there exists $\trianglerighteq \in R$ such that $\trianglerighteq$ is an extension of $D(x)$.*

In other words, a rationalization by multiple rationales is a collection of preference relations such that each observation $x \in X$ can be explained by one of these rationales.

Definition 4 *A minimal multiple rationales (MMR) is a rationalization by multiple rationales such that no smaller collection of rationales could rationalize the data.*

The problem of calculating the MMR for a data set is equivalent to the problem of finding a partition of the observation set X such that:

1. The observations in each set of the partition are acyclic.
2. Any partition of X that is composed of fewer divisions has at least one division that is not acyclic.

[8] i.e. complete, transitive and reflexive binary relations.

The size of the minimal number of rationales provides different information on the rationality of a particular data set than does the HM index. To understand the difference, consider the following hypothetical experimental subjects. We observe subject C making the following choices:

$$\begin{aligned} C_C(\{x,y,z\}) &= \{x\} \\ C_C(\{x,y,z,t\}) &= \{x\} \\ C_C(\{x,y,z,v\}) &= \{z\} \\ C_C(\{z,y,x,w\}) &= \{z\} \end{aligned}$$

while subject D chooses as follows:

$$\begin{aligned} C_D(\{x,y,z\}) &= \{x\} \\ C_D(\{x,y,z,t\}) &= \{x\} \\ C_D(\{x,y,z,v\}) &= \{y\} \\ C_D(\{z,y,x,w\}) &= \{z\} \end{aligned}$$

Both C and D have an HM index of 2 because the largest collection of acyclical observations is 2. Thus, according to the HM index, these two data sets are equally 'irrational'. However, the two subjects require different numbers of rationales to explain their behavior. Subject C has a MMR of size 2 – the data can be rationalized by two rationales (one in which x is ranked above all other alternatives and one in which z is ranked above all alternatives). Subject D, however, has a MMR of size 3 – three rationales are needed to rationalize the data (one in which x is ranked above y and z, one with y ranked above x and z and one with z ranked above x and y).

3 Cost Based Measures

3.1 The Afriat Measure

An earlier example at trying to get at the cost of deviations from rationality was developed by Afriat [1972] for the case of choice from budget sets. The measure relies on the concept of being revealed preferred at an efficiency level

Definition 5 *We say that x is revealed preferred to y at efficiency level e if $ep^x x > p^x y$.*

Note that efficiency level 1 is the same as standard revealed preference, while for $e = 0$ the revealed preference relation is empty. Afriat's measure of rationality is the efficiency level e such that the resulting revealed preference relation is acyclic (the previous remark says that this has to be true for some $0 \leq e \leq 1$. This measure has the advantage of being easy to calculate, and taking into account the cost of revealed preference violations. However, like the counting measure, it is not very robust, in the sense that it is very sensitive to a single observation - one crazy choice can send the index to 0.This flaw also means that the index is not good at differentiating between data sets that, intuitively, we might think of as having very different levels of irrationality: If one keeps the 'worst' violation constant, but adds more irrational choices of equal or lower severity, then this will not affect the value of the Afriat measure.

3.2 The Varian Measure

In part to overcome this problems associated with the Afriat measure, Varian proposed an alternative which calculated a vector of efficiency measures, one for each choice. Specifically, for each observation $x \in X$, the Varian approach calculates the maximum efficiency level e_x that removes all preference cycles that involve observation x. The problem, of course, is how to condense this vector into a single number. One way would be to take the minimum of the vector. However, this is then becomes very similar to the Afriat measure. Another approach would be to calculate the vector of e_x's that between them remove all cycles, and are closest to the unit vector in some metric. Such a measure is close to the hybrid measures described below.

4 Hybrid Measures

4.0.1 The HM-e Index

A big problem with the HM index is that it does not contain any notion of the 'severity' of a particular violation of acyclicality. This is most obvious in the case in which the observed choices are over bundles of commodities from different budget sets. Consider the following choice behavior for hypothetical subjects E and F from budget sets in a commodity space that contains two goods (a and b):

- Budget set 1 : income is 10, price of good A is 2, price of good B is 2
  - E buys 1 unit of good A and 4 units of good B
  - F buys 2 units of good A and 3 units of good B
- Budget set 2 : income is 10, price of good A is 3, price of good B is 1
  - E buys 3 unit of good A and 1 unit of good B
  - F buys 3 unit of good A and 1 units of good B

Both of these consumers violate acyclicality, as in both cases the bundle bought in budget set 2 was available in budget set 1, and vice versa. However, the 'cost'[9] of the acyclicality violation for subject E was higher than that for subject F. For E, the bundle chosen from budget set 1 was available at a cost of 7 from budget set 2, while for consumer F, the bundle chosen from set 1 was available at a cost of 9. Thus the 'cost' of the acyclicality violation for E is 3, while for F it is only 1. Yet both subjects would have the same HM index.

More generally, a researcher may have some form of external metric on how different are a pair of objects or how strong is a particular revealed preference relation. A desirable property in a measure of the rationality of a particular data set is that it can take into account differences in the 'strength' of a particular revealed preference relation – punishing cycles that involve only 'strong' relations more than cycles that involve only 'weak' relations. The HM index has no way of incorporating this information into its measure of irrationality.

In the above example, a natural metric for the strength of preference exhibited when x is chosen over y could be the cost difference between x and y measured in terms of some denominator good. In other cases we might have some intuition that two objects are similar (a weak relation that is easy to break), or that they are very different (a strong relation that is hard to break).

The HM-e index attempts to correct this problem with the HM index in the following way.

1. As the basic unit of analysis, we use each revealed preference relation, rather than each observed choice.

[9] Here cost can be thought of as a potential 'money pump' or as 'wasted' income.

2. We allow for the varying 'costs' of removing different revealed preference relations, depending on an external metric for the strength of each preference.

To understand the impact of the first change, consider the case of object y being chosen from a choice set $\{x, y, z, w\}$. The HM index treats this as a single observation, and asks only whether or not this observation needs to be removed in order to guarantee acyclicality. The modified index treats this as three separate pieces of information: that y is preferred to x, that y is preferred to z and that y is preferred to w.

The second change allows for the measure to be modified to account for the strength of the revealed preference relations. Formally, the primitives of the HM-e index are a data set $\{Z, X, D\}$ and a weighting function $w : D(X) \rightarrow \mathbb{R}$. The weighting function carries information on the strength of different revealed preference relations. Note that it do not propose any particular metric – the weighting function forms part of the input to the measure. However, as discussed above, in the case of choice of budget sets, one natural measure would be the cost difference between the chosen and unchosen bundle. Such a weighting function would bridge the gap between rationality measures that only count the *number* of violations of rationality (Famulari's [1995] measure and the HM index) and those that look only at the *cost* of such violations (Afriat [1972] and Varian [1991]).[10]

Given these primitives, the HM-e index is defined as follows:

Definition 6 *For a data set $\{Z, X, D\}$ and a weighting function $w : D(X) \rightarrow \mathbb{R}$, the modified HM index is defined as*

$$W = \min_{B \subset D(X)} \sum_{a \in B} w(a)$$

Definition 7 *such that $D(X)/B$ is acyclic.*

In other words, the modified HM index is the minimum cost way to create an acyclical set of preference relations according to the weighting function w.

In the case of budget sets, one way to generate a weighting matrix is as follows:

$$W(x \succ y) = 1 - e(x, y)$$

[10] For an example of this implementation, see section **??**.

where $e(x,y)$ is defined as

$$e(x,y)p^x x = p^x y$$

4.1 Money-Pump Measure

Another, similar measure has been suggested by Echenique et al [2011], which the call the 'money pump' measure. This measure (which is designed to be applied to choices from budget sets) is constructed in the following way:

1. Identify every preference cycle - i.e. every sequence of choices $x^1, ...x^n$ such that

$$\begin{aligned} p^1x^1 &\geq p^1x^2 \\ p^2x^2 &\geq p^2x^3 \\ &\vdots \\ p^nx^n &\geq p^nx^1 \end{aligned}$$

2. For each cycle, calculate the 'money pump' from that cycle

$$\begin{aligned} &p^1(x^1 - x^2) \\ &+p^2(x^2 - x^3) \\ &\vdots \\ &+p^n(x^n - x^1) \end{aligned}$$

3. Sum the money pump across all cycles.

Again, the interpretation is that this is the amount of money that a person wastes due to preference cycles.

Clearly, this measure is related to the HM-e index. The difference between the two measures can be described from the following data set:

- x_1 chosen when x_2 was \$1 cheaper
- x_2 chosen when x_3 was \$1 cheaper

- x_3 chosen when x_1 was \$10 cheaper

The HM-e index would remove this cycle at a ‘cost’ of \$1 - i.e. the cheapest way of killing the cycle. In contrast, the Money pump measure would remove this cycle at a ’cost’ of \$12 - the total cost of the cycle.

5 Power Measures

One issue with any of the rationality measures described above is that it is hard to interpret what a particular value tells us about the underlying data. For example, consider a data set in which we observe choices from two disjoint choice sets. In this case all our measures will give perfect rationality scores for *any* observed pattern of choice. In other words, such a data set offers no meaningful test of rationality. One way to address this shortcoming is to compare the values of our chosen index to the distribution of values we would see under some alternative ‘null hypothesis’ for behavior. Such a comparison allows one to determine whether observed behavior shows more, less or similar levels of rationality than the null hypothesis.

One popular benchmark is to compare index values to those that one would expect to see under random choice – in each choice set individuals have an equal chance of choosing any object in the choice set.[11] Although random choice is a relatively weak null hypothesis, it is applicable to almost any choice setting. The role of random choice in determining the statistical power of rationality measures is discussed by Bronars [1987] and is applied to Selten’s measure of predictive success by Beatty and Crawford [2010].[12] The latter reference also gives a neat geometric interpretation, and an axiomatic justification for this approach. However, see Dean and Martin [2011] for a discussion of why a uniform distribution over all possible choices may be problematic as a benchmark.

Once we have generated a benchmark, the next question is how to compare the experimental data to this benchmark. For a joint test of all subjects, one can compare the *distribution* of the index scores in the data with the *distribution* of index scores generated under the null hypothesis using some nonparametric measure of the difference between distributions (such as the Kolmogorov-Smirnoff test). In the case of a single observation, one can simply read off the percentile of the

[11] Or, in the case of budget sets, an equal chance of choosing any object on the budget line.

[12] Alternatively, we could generate a distribution of possible index values for a given choice environment using a more plausible error model or decision rule. For example, see Choi et al. [2006] and Andreoni and Harbaugh [2006].

simulated data in which that observation falls. Another intuitive measure is to subtract the average simulated score from an actual score – in the style of Selten's measure of predictive success [1991]. For the HM index, the resulting number would represent the fraction of a data set consistent with rationality over and above what could be explained by random choice.