18th World IMACS / MODSIM Congress, Cairns, Australia 13-17 July 2009
http://mssanz.org.au/modsim09

How Consistent are your Choice Data?

Dean, M. [1] and D. Martin [1]

[1] *Department of Economics, New York University*
Email: daniel.martin@nyu.edu

Abstract: Economists have a long tradition of specifying their models in terms of axioms, or restrictions on data which must be obeyed for it to be consistent with a particular model. One common axiom is that a binary relation on some set Z must be acyclic.

Acyclic, antisymmetric relations have the important property that they can be extended to complete linear orders. As such, we can interpret such relations as being an incomplete observation of some consistent ordering on Z. The most famous use of this result is Strong Axiom of Revealed Preference (SARP), which shows that a set of choice data can be considered consistent with the maximization of some underlying preference ordering if and only if the generated revealed preference relation is acyclic. Caplin and Dean [2008] provide an example in which the acyclic property is central to characterising a model of search and choice.

One problem with the axiomatic method of characterizing a model is that it provides only a very stark measure of whether a data set is consistent with a particular model: data either does or does not violate the stated axioms. There is no concept of whether a data set is `close' to satisfying an axiom set. Recognising this problem, several authors have proposed measures of how `far away' a data set is from satisfying a set of axioms (see Afriat [1972], Varian [1991], and Houtman and Maks [1985]).

The Houtman and Maks measure is based on finding the largest subset of observations which satisfy the axiomatic system. While this measure is not without its problems (see Choi et al. [2006] for a discussion), it has the advantage of being applicable to wide variety of data sets and axiomatic systems. In contrast, the Afriat and Varian measures are only applicable to data obtained by observing choices derived from different budget sets.

One possible reason that the Houtman and Maks measure has not been widely adopted is that it can be extremely computationally intensive (see Choi et al. [2007] and Fisman et al. [2007] for examples in which computational constraints have been binding). The innovation in this paper is to show that the problem of finding the maximal acyclic subset can be reduced to a well-studied problem within the computer sciences and operations research: the Minimum Set Covering Problem (MSCP). While MSCP is NP-hard in the strong sense, there are a wide variety of algorithms built to solve this problem (see Caprara, Toth, and Fischetti [2000]), which can be used to find the maximal acyclic set quickly and exactly for reasonably-sized data sets. This paper describes some of these algorithms, and a companion website (www.danielmartin.com) provides code which adapts them to the Houtman-Maks measure. Furthermore, we demonstrate that with this approach, the measure can be calculated in under a second for cases that were previously insoluble. This result opens up the possibility that the Houtman-Maks measure can be developed into a more formal statistical test of axiomatic consistency.

Keywords: *Revealed Preference, Choice Data, Axiomatic Consistency*

1. INTRODUCTION

Economists have a long tradition of specifying their models in terms of axioms, or restrictions on data which must be obeyed for it to be consistent with a particular model. One common axiom is that a binary relation on some set Z must be acyclic.

Acyclic, antisymmetric relations have the important property that they can be extended to complete linear orders. As such, we can interpret such relations as being an incomplete observation of some consistent ordering on Z. The most famous use of this result is Strong Axiom of Revealed Preference (SARP), which shows that a set of choice data can be considered consistent with the maximization of some underlying preference ordering if and only if the generated revealed preference relation is acyclic. Caplin and Dean [2008] provide an example in which the acyclic property is central to characterising a model of search and choice.

One problem with the axiomatic method of characterizing a model is that it provides only a very stark measure of whether a data set is consistent with a particular model: data either does or does not violate the stated axioms. There is no concept of whether a data set is `close' to satisfying an axiom set. Recognising this problem, several authors have proposed measures of how `far away' a data set is from satisfying a set of axioms (see Afriat [1972], Varian [1991], and Houtman and Maks [1985]).

The Houtman and Maks measure is based on finding the largest subset of observations which satisfy the axiomatic system. While this measure is not without its problems (see Choi et al. [2006] for a discussion), it has the advantage of being applicable to wide variety of data sets and axiomatic systems. In contrast, the Afriat and Varian measures are only applicable to data obtained by observing choices derived from different budget sets.

One possible reason that the Houtman and Maks measure has not been widely adopted is that it can be extremely computationally intensive (see Choi et al. [2007] and Fisman et al. [2007] for examples in which computational constraints have been binding). The innovation in this paper is to show that the problem of finding the maximal acyclic subset can be reduced to a well-studied problem within the computer sciences and operations research: the Minimum Set Covering Problem (MSCP). While MSCP is NP-hard in the strong sense, there are a wide variety of algorithms built to solve this problem (see Caprara, Toth, and Fischetti [2000]), which can be used to find the maximal acyclic set quickly and exactly for reasonably-sized data sets. This paper describes some of these algorithms, and a companion website (www.danielmartin.com) provides code which adapts them to the Houtman-Maks measure. Furthermore, we demonstrate that with this approach, the measure can be calculated in under a second for cases that were previously insoluble. This result opens up the possibility that the Houtman-Maks measure can be developed into a more formal statistical test of axiomatic consistency.

2. METHOD

Using this result to calculate the HM Index requires two algorithmic components. First, to fully specify the mapping F on C, we need an algorithm that identifies the set of all cycles C and the observations $x \in X$ that break each cycle $c \in C$. One option is Johnson's Algorithm, a computationally efficient graph theory algorithm (see Johnson [1975]). A graph G is composed of nodes N and edges E, and preferences can be represented as a directed graph by creating a node for object (E=Z) and placing a directed edge between nodes when one object is preferred to another (e.g., $e_1=(z_1,z_2)$ if $z_1 \succ z_2$). Johnson's Algorithm is based on `depth-first' search, a standard approach to finding cycles, which looks at the objects preferred to an initial object, then looks for the objects that are preferred to the first of those preferred objects and so on until a cycle is found or the process terminates. At that point, the algorithm goes back one level and proceeds from the second preferred object until all possibilities are exhausted.

To gain efficiency, Johnson adds a blocking function to prevent redundant searching on the tree, which gives it a computation time upper bound of $O((n+e)(c+1))$, where n is the number of nodes, e is the number of edges and c is the number of cycles. We add additional efficiency by modifying Johnson's Algorithm to only look at those cycles without subcycles.

Second, we need an algorithm to solve the Minimum Set Covering Problem (MCSP), which is NP-hard. However, MSCP has been studied exhaustively because it can be applied to many real-world situations, such as train scheduling and city planning. As a result, algorithms have been developed to solve or approximate solutions to MSCP quickly for larger and larger data sets. Branch and bound algorithms find an exact solution by iteratively `relaxing' the integer programming problem so that linear programming techniques can be used to create bounds on the problem. These algorithms have been integrated into standard Integer

Programming (IP) software packages, including commercial solvers (e.g., CPLEX) and non-commercial solvers (e.g., SCIP and MINTO), which outperform stand-alone algorithms and tend to work quickly for most data sets (see Caprara, Toth, and Fischetti [2000]). In the following benchmark comparison, we use the bintprog command in the MATLAB Optimization Toolbox to solve MSCP.

3. BENCHMARK

To benchmark our approach, we use the data and analytical computer program from Choi et al [2006]. In this paper, 93 subjects allocate tokens between account x and account y using a novel graphical interface. Over 50 rounds, subjects face randomly selected budget lines, with a constant, normalized wealth level. The state of the world is uncertain, and in one state of the world, each x token pays \$.50 and each y token pays nothing, and in the other state of the world, the reverse is true. We say a bundle (x_1,y_1) is revealed preferred to a bundle (x_2,y_2) if both are available in the budget set and (x_1,y_1) is selected instead of (x_2,y_2).

Choi et al [2006] report that finding the HM Index with their approach is infeasible for subjects with data that is `far away' from acyclicality. To see why calculating the HM Index can be difficult in such cases, imagine a data set that has 50 observations, where the maximal acyclic subset contains just 41 observations. Before you can find the maximal acyclic subset, you must first determine that all subsets of size 49 contain a cycle, then all subsets of 48 observations and so on until you have checked all possible subsets of 42 observations, which involves over 655,000,000 checks.

To calculate the HM Index more quickly, the authors first partition their choice data using the strongly connected components of the corresponding graph, and then look for the largest acyclic subset of each component. In addition to reducing computation time, partitioning sets allows the authors to calculate a lower-bound on the size of the maximal acyclic subset for each subject.

Even though most subjects were close to being perfect rational, Choi et al [2006] were unable to produce an exact HM Index score for 5% (6 of 93) of subjects, who are listed in Table 1. Our algorithmic approach was able to solve even the most difficult case in under a quarter of a second. Additionally, the benchmark lower bound was often significantly smaller than the size of the largest acyclic set. The program files for our approach are available at www.danielmartin.com.

	Benchmark Approach		New Approach	
ID	HM Index (Lower Bound)	Run Time (Lower Bound)	HM Index (Exact)	Run Time (Exact)
211	34	> 30 seconds	35	0.1156 secs
324	31	> 30 seconds	43	0.0475 secs
325	34	> 30 seconds	41	0.0709 secs
406	30	> 30 seconds	40	0.0451 secs
504	33	> 30 seconds	45	0.0401 secs
608	30	> 30 seconds	40	0.0451 secs

Table 1. Benchmarking results

The computational difficulty in calculating the HM Index also meant that Choi et al. [2006] were unable to benchmark their results against random choices. Bronars [1987] discusses the role that random choice can play in determining the power of a measure. To apply such an idea to the HM Index, one must be able to calculate the measure for random choices, which can be very far from rationality. The improved efficiency of our approach allows for such benchmarking. Figure 1 shows the distribution of HM Index scores for the subjects' choice data (reproduced from Choi et al. [2006]) and for hypothetical subjects selecting baskets of goods at random from the budget line. This simulation was run for 25,000 subjects making uniformly distributed choices on randomly generated budget lines in keeping with the experimental design.

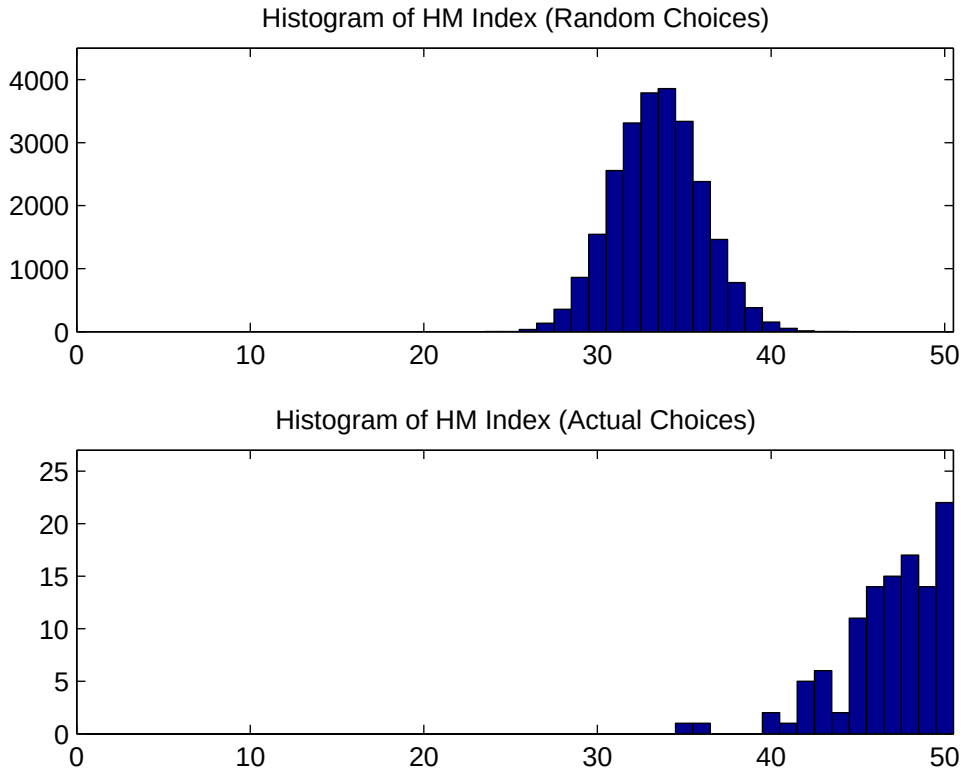

Figure 1. Comparison with random choice data

4. DISCUSSION AND CONCLUSIONS

This paper provides a new tool to help answer the question: `How close is a set of choice data to rationality?' While not perfect, the HM Index is a flexible and powerful way to answer this question, but has largely been abandoned on the basis of its computational difficulty. By showing that the problem of finding the HM Index can be reduced to the Minimum Set Covering Problem, we have removed these constraints for many data sets. We can now solve problems in fractions of a second that were previously insoluble in any reasonable length of time. By doing so, we have opened the door to more sophisticated use of the HM Index, including simulations to benchmark models of choice.

ACKNOWLEDGMENTS

We are grateful to Shachar Kariv for providing the data and computer program used in Choi et al [2006]; Debraj Ray, Michael Richter, Hiroki Nishimura and participants in the NYU NRET seminar for helpful comments; Jesse Perla for programming suggestions; and Adam Sachs for research assistance.

REFERENCES

Afriat, S. (1972), Efficiency Estimates of Production Functions. *International Economic Review*, 8, 568-598.

Bronars, S. (1987), The Power of Nonparametric Tests of Preference Maximization. *Econometrica*, 55, 693-698.

Caprara, A., Toth, P., and Fischetti, M. (2000), Algorithms for the Set Covering Problem. *Annals of Operations Research*, 98, 352-371.

Caplin, A., and Dean, M. (2008), Information Search and the Choice Process, mimeo.

Choi, S., Gale, D., Fisman, R., and Kariv, S. (2006), Substantive and Procedural Rationality in Decisions under Uncertainty, mimeo.

Choi, S., Gale, D., Fisman, R., and Kariv, S. (2007), Consistency and Heterogeneity of Individual Behavior under Uncertainty. *American Economic Review*, 97(5), 1921-1938.

Fisman, R., Kariv, S., and Markovits, D. (2007), Individual Preferences for Giving. *American Economic Review*, 97(2), 153-158.

Garey, M.R. and Johnson, D.S. (1979), Computers and Intractability: A Guide to the Theory of NP-Completeness. New York: W.H. Freeman.

Houtman, M., and Maks, J.A.H. (1985), Determining all Maximal Data Subsets Consistent with Revealed Preference. *Kwantitatieve Methoden*, 19, 89-104.

Johnson, D. (1975) Finding All the Elementary Circuits of a Directed Graph. *SIAM Journal of Computing*, 4, 77-84.

Varian, H. (1991) Goodness-of-Fit for Revealed Preference Tests. *University of Michigan CREST Working Paper # 13*.