Economics Bulletin

Volume 35, Issue 4

Note on goodness-of-fit measures for the revealed preference test: The computational complexity of the minimum cost index

Kohei Shiozawa
Graduate School of Economics, Osaka University

Abstract

The purpose of this paper is to show that computing the minimum cost index (MCI) for a given price-amount data set, proposed by Dean and Martin (2010, 2015) as a goodness-of-fit measure for the revealed preference test, is NP-hard. Our proof uses a polynomial reduction from the feedback arc set problem, which is a decision problem known to be NP-complete. Our result refines the NP-hardness result in Dean and Martin (2010), which is presented in a more abstract framework than our economic data setting. Thus the computation of MCI is NP-hard even if we restrict our attention to the revealed preference setting for economic data. We also discuss computational procedures for MCI and provide a way of approximating MCI in polynomial-time using approximation algorithms for the (weighted) feedback arc set problem.

Citation: Kohei Shiozawa, (2015) "Note on goodness-of-fit measures for the revealed preference test: The computational complexity of the minimum cost index", *Economics Bulletin*, Volume 35, Issue 4, pages 2455-2461
Contact: Kohei Shiozawa - nge013sk@student.econ.osaka-u.ac.jp.
Submitted: September 29, 2015. **Published:** November 21, 2015.

1. Introduction

Given a finite data set $\{(p_k, x_k)\}_{k=1}^n$ where $p_k \in R^\ell_{++}$, $x_k \in R^\ell_+$ $(k = 1, \ldots, n)$, the data can be supported by some non-satiated utility function $u : R^\ell_+ \to R$ as solutions of a utility maximization problem if and only if the data are consistent with the *generalized axiom of revealed preference* (GARP). This classical result originates from Afriat (Afriat, 1962; Diewert, 1973; Varian, 1982). Based on this result, we can determine whether or not there is a utility function that rationalizes an arbitrary data set $\{(p_k, x_k)\}_{k=1}^n$. This test is a binary choice between two alternatives and hence, if some of the data violates GARP, then we have no information about the severity of the violation.

To quantify how severely a data set $\{(p_k, x_k)\}_{k=1}^n$ violates GARP, several severity-measures have been proposed as *goodness-of-fit* measures. For example, Afriat (1973), Houtman and Maks (1985), Varian (1990), Echenique et al. (2011), and Smeulders et al. (2013) proposed indices that express, in various ways, how severely the data violate GARP.[1] However, Smeulders et al. (2013, 2014) showed that computing some of those indices is *NP-hard*.[2] More specifically, they showed that the indices in Houtman and Maks (1985), Varian (1990), and Echenique et al. (2011) are NP-hard and proposed a polynomial-time algorithm for the index in Afriat (1973), as well as two new indices that have polynomial-time algorithms and behave similarly to the index in Echenique et al. (2011).

Recently, Dean and Martin (2010, 2015) proposed a new index called the *minimum cost index* (MCI). They computed the MCI for a real economic data set $\{(p_k, x_k)\}_{k=1}^n$ using an NP-hard problem called the *minimum set covering problem* (MSCP). In this paper, we show the converse. We show that an NP-hard problem called the *feedback arc set problem* can be solved using the problem of computing MCI for an economic dataset $\{(p_k, x_k)\}_{k=1}^n$, and thus the MCI computation problem for an economic dataset $\{(p_k, x_k)\}_{k=1}^n$ is in the class of NP-hard problems. Note that Dean and Martin (2010, 2015) and their online appendices show that a problem they call the *maximal acyclical set problem* (MASP) is equivalent to MSCP, and hence is NP-hard. Because MASP contains the computation of MCI in the economic data setting, our result refines their NP-hardness result, in that there is no exact polynomial-time algorithm for the MCI computation problem even if we restrict our attention to the revealed preference setting for a given economic dataset $\{(p_k, x_k)\}_{k=1}^n$.

2. Minimum Cost Index: Definition, Validity, and Complexity Result

MCI is a goodness-of-fit measure for the revealed preference test proposed by Dean and Martin (2010, 2015). MCI expresses the minimum cost of information that must be ignored

[1] Another branch of research in the literature investigates PC-based Monte Carlo procedures for measuring the severity of GARP violations. See Gross (1995), Fleissig and Whitney (2005), and Heufer (2008). For an overview of older goodness-of-fit measures, see Gross (1995). Apesteguia and Ballester (2015) provided a new argument for comparing these alternative measures from the welfare analytic view point.

[2] An *optimization problem* or a *decision problem* in the class of NP-hard (or *NP-complete*) problems is not easy to solve in the sense that, if there is a polynomial-time algorithm for the problem, then every problem in the class of *NP* problems can be solved in polynomial-time. In other words, an NP-hard problem is no easier to solve than any other NP problem. As an example, the well-known *Hamiltonian circuit problem* is in the class of NP problems.) Moreover, no NP problem at present has an exact polynomial-time algorithm, and hence it must also be difficult to find an exact polynomial-time algorithm for an NP-hard problem. For more detailed discussion on the theory of NP, NP-completeness, and NP-hardness, see Karp (1972), Garey and Johnson (1979), and Korte and Vygen (2012).

from the data set so that the remaining information of the data satisfy a condition equivalent to GARP. We first formalize what is meant by a condition equivalent to GARP using some graph theoretic apparatus.

A *directed graph* is a pair $G = (V, E)$ where V is a finite set and E is a subset of the ordered pairs $V \times V$. We call an element $v \in V$ a *vertex* and an element $(v, u) \in E$ an *edge*.[3] A *path* (or a v_{i_0}-v_{i_m} *path*) $v_{i_0} \to v_{i_1} \to \cdots \to v_{i_m}$ is a sequence of vertices and edges $v_{i_0}(v_{i_0}, v_{i_1})v_{i_1}(v_{i_1}, v_{i_2})v_{i_2} \cdots v_{i_{(m-1)}}(v_{i_{(m-1)}}, v_{i_m})v_{i_m}$. A path is called a *cycle* if $v_{i_0} = v_{i_m}$. A graph that contains no cycle is said to be *acyclic*.

Definition 1. For any data $\{(p_k, x_k)\}_{k=1}^n$ where $p_k \in R^\ell_{++}$, $x_k \in R^\ell_+$ $(k = 1, \ldots, n)$, we define the following directed graph: $G_{np} := (V, E_{np})$ where

$$V := \{x_1, x_2, \ldots, x_n\} \tag{1}$$

$$E_{np} := \{(x_k, x_{k'}) \,|\, k, k' = 1, \ldots, n,\ k \neq k',\ \text{and}\ \ p_k \cdot (x_{k'} - x_k) \leqq 0\}. \tag{2}$$

We call the graph $G_{np} := (V, E_{np})$ the *associated graph* or the *graph associated with the data.*

As pointed out in Piaw and Vohra (2003), Fujishige and Yang (2012), and Talla Nobibon et al. (2014), we can translate GARP into the graph theoretic conditions stated in the following proposition.[4]

Proposition 1. *Given the data $\{(p_k, x_k)\}_{k=1}^n$, the following three conditions are equivalent:*

(i) *Cyclical consistency (GARP): $p_{k_0} \cdot (x_{k_1} - x_{k_0}) \leqq 0$, $p_{k_1} \cdot (x_{k_2} - x_{k_1}) \leqq 0$, $\ldots$, $p_{k_m} \cdot (x_{k_{m+1}} - x_{k_m}) \leqq 0$ (where $k_{m+1} = k_0$) implies that $p_{k_i} \cdot (x_{k_{(i+1)}} - x_{k_i}) = 0$ for all $i = 0, \ldots, m$.*

(ii) *The associated graph $G_{np} = (V, E_{np})$ has no cycle that contains an edge $(x_k, x_{k'})$ satisfying $p_k \cdot (x_{k'} - x_k) < 0$.*

The first condition is the *cyclical consistency* condition of Afriat (1967) and is equivalent to GARP (see Varian (1987)). The second condition is a graph theoretic equivalent of the cyclical consistency condition that is related to the definition of MCI as we will see below.

Definition 2. (Dean and Martin (2010, 2015)) For any data $\{(p_k, x_k)\}_{k=1}^n$ where $p_k \in R^\ell_{++}$, $x_k \in R^\ell_+$ $(k = 1, \ldots, n)$, we define MCI as follows:

$$\begin{aligned} \text{MCI} := \min\{ \sum_{(x_k, x_{k'}) \in E'} p_k \cdot (x_k - x_{k'}) \mid E' \subset E_{np} \ \ \text{and} \\ G' := (V, E_{np} \setminus E') \ \ \text{is acyclic}\} \end{aligned} \tag{3}$$

where $G_{np} = (V, E_{np})$ is the graph associated with the data $\{(p_k, x_k)\}_{k=1}^n$.

Note that Dean and Martin (2010, 2015) defined MCI as normalized by the total wealth of the data, that is, by dividing the value (3) by $\sum_{k=1}^n p_k \cdot x_k > 0$. We omit this normalization as our goal is to determine the computational complexity of MCI and (omission of) normalization obviously does not harm the conclusion. Moreover, while the original definition of

[3] In this paper, we only consider *simple* graphs. That is, we assume that there are no *loops* or *parallel edges* in $G = (V, E)$.

[4] For a detailed discussion for this graph theoretic structure, see also Shiozawa (2015).

MCI is based on the *direct revealed preference relation* R^0 of Varian (1987), we adopt its graph theoretic representation G_{np}.

While the definition of MCI is based on the idea of the minimum cost of information that must be ignored so that the remaining data satisfies a condition equivalent to GARP, the condition used in Equation (3) is slightly different from that in Condition (ii) in Proposition 1. However, it is clear that

$$\text{MCI} = \min\{ \sum_{(x_k, x_{k'}) \in E'} p_k \cdot (x_k - x_{k'}) \mid E' \subset E_{np} \text{ and } G' = (V, E_{np} \setminus E') \\ \text{has no cycle containing a negative edge}\}. \tag{4}$$

Thus, we can see that the definition of MCI actually grasps the above-mentioned idea and, from Proposition 1, that MCI also has theoretical validity as a goodness-of-fit measure for GARP violation. Conversely, MCI is not easy to compute in general. Actually, as we will see below, the problem of calculating MCI for an arbitrarily given data set $\{(p_k, x_k)\}_{k=1}^n$ is NP-hard. Hence, there is probably no polynomial-time exact algorithm for computing MCI for a given data set $\{(p_k, x_k)\}_{k=1}^n$. We formally state the problem of computing MCI below:

MCI COMPUTATION
Instance: An integer $\ell \geqq 1$ and a data set $\{(p_k, x_k)\}_{k=1}^n$ where $p_k \in R^\ell_{++}$, $x_k \in R^\ell_+$.
Task: Compute MCI as defined in Equation (3).

Theorem 1. MCI COMPUTATION *problem is NP-hard.*

Proof. Observe that the right-hand side of Equation (3) is a problem that asks the minimum value of edges to remove to get an acyclic subgraph of G_{np}. There is a similar problem called the *feedback arc set problem* (FAS), which is known to be NP-complete (Garey and Johnson, 1979, p.192). FAS is formalized as follows:

FAS
Instance: A directed graph $G = (U, A)$ and an integer $k \leqq |A|$.
Question: Is there any subset $A' \subset A$ such that subgraph $G' = (U, A \setminus A')$ is acyclic and $|A'| \leqq k$.

We use a polynomial reduction from FAS. In other words, we show that any instance of FAS can be reduced to an MCI computation problem of an economic data $\{(p_k, x_k)\}_{k=1}^n$ in polynomial-time. Given an arbitrary instance of FAS, namely a directed graph $G = (U, A)$ and an integer $k \geqq 0$, we construct an instance of the MCI computation problem, that is, data $\{(p_k, x_k)\}_{k=1}^n$, where $p_k \in R^\ell_{++}$, $x_k \in R^\ell_+$ $(k = 1, \ldots, n)$, and $n = \ell := |U|$.[5] First, we define the consumptions $x_k \in R^n_+$ $(k = 1, \ldots, n)$ as

$$x_k := (x_{k1}, \ldots, x_{k(k-1)}, x_{kk}, x_{k(k+1)}, \ldots, x_{kn}) := (0, \ldots, 0, 1, 0, \ldots, 0). \tag{5}$$

Next, we define the prices $p_k \in R^n_{++}$ $(k = 1, \ldots, n)$ as

$$p_{kk'} := \begin{cases} 1 & \text{if } (u_k, u_{k'}) \in A \\ 2 & \text{if } k = k' \\ 3 & \text{if } (u_k, u_{k'}) \notin A \end{cases} \tag{6}$$

[5] A similar reduction technique is used in Smeulders et al. (2014).

where $p_{kk'}$ is the k'-th coordinate of the price $p_k \in R^n_{++}$ and $u_k, u_{k'} \in U$. We have

$$p_k \cdot (x_{k'} - x_k) = p_{kk'} - p_{kk} = \begin{cases} -1 < 0 & \text{if } (u_k, u_{k'}) \in A \\ 1 > 0 & \text{if } (u_k, u_{k'}) \notin A \end{cases} \tag{7}$$

for all $k, k' = 1, \ldots, n$ where $k \neq k'$. Therefore, if we construct the associated graph G_{np} for these data $\{(p_k, x_k)\}_{k=1}^n$, then it is evident from Equation (7) and the definition of the associated graph G_{np} that there is a one-to-one correspondence between the edges in $G = (U, A)$ and the edges in $G_{np} = (V, E_{np})$. Moreover, from Equation (7), the MCI value for the data $\{(p_k, x_k)\}_{k=1}^n$ is

$$\begin{aligned} \text{MCI} &= \min\{ \sum_{(x_k, x_{k'}) \in E'} p_k \cdot (x_k - x_{k'}) \mid E' \subset E_{np} \text{ and } G' := (V, E_{np} \setminus E') \text{ is acyclic} \} \\ &= \min\{ \sum_{(u_k, u_{k'}) \in A'} 1 \mid A' \subset A \text{ and } G' := (U, A \setminus A') \text{ is acyclic} \} \\ &= \min\{ |A'| \mid A' \subset A \text{ and } G' := (U, A \setminus A') \text{ is acyclic} \}. \end{aligned} \tag{8}$$

Therefore, we have an algorithm for solving FAS that uses the MCI computation problem as a subroutine:

Step 1: Construct the instance for MCI computation from the given instance for FAS according to Equations (5) and (6).

Step 2: Construct the associated graph G_{np} for the data constructed in Step 1.

Step 3: Compute the MCI as defined in Equation (3).

Step 4: Compare the MCI value with the given integer k. If the MCI value is strictly greater than k, then the answer to FAS is NO; otherwise the answer is YES.

Assuming that we have a polynomial-time algorithm for the MCI computation problem, the computational time for the above algorithm for FAS is also polynomial. Moreover, because FAS is NP-complete, this implies that every NP problem can be solved in polynomial-time. In summary, if we assume that there is a polynomial-time algorithm for the MCI computation problem, then every NP problem can be solved in polynomial-time, and therefore the MCI computation problem is NP-hard. □

3. Remarks on Computational Procedures for MCI

While there is probably no exact polynomial-time algorithm for the MCI computation problem, this does not mean that we cannot compute the MCI value for a given data set $\{(p_k, x_k)\}_{k=1}^n$. There are many exact algorithms, heuristics, and approximation algorithms for the *minimum feedback arc set problem* (the weighted version of FAS), and hence we can compute or approximate MCI using these algorithms and/or heuristics. For exact algorithms and heuristics, see Eades et al. (1993) and Eades and Lin (1995). For approximation algorithms, see Seymour (1995), Evan et al. (2000), and Demetrescu and Finocchi (2003).

In addition, approximating MCI can be performed using the problem complementary to the minimum feedback arc set problem, that is, the *maximum* feedback arc set problem. Let E_0 be a subset of E_{np} defined as

$$E_0 := \{(x_k, x_{k'}) \in E_{np} \,|\, p_k \cdot (x_{k'} - x_k) = 0\}.$$

We have

$$\begin{aligned}
&\sum_{(x_k,x_{k'})\in E_{np}} p_k \cdot (x_k - x_{k'}) - \text{MCI} \\
&= \max\{ \sum_{(x_k,x_{k'})\in E_{np}} p_k \cdot (x_k - x_{k'}) - \sum_{(x_k,x_{k'})\in E'} p_k \cdot (x_k - x_{k'}) \mid \\
&\qquad\qquad\qquad E' \subset E_{np} \text{ and } G' := (V, E_{np} \setminus E') \text{ is acyclic}\} \\
&= \max\{ \sum_{(x_k,x_{k'})\in E'} p_k \cdot (x_k - x_{k'}) \mid E' \subset E_{np} \text{ and } G' := (V, E') \text{ is acyclic}\} \\
&= \max\{ \sum_{(x_k,x_{k'})\in E'} p_k \cdot (x_k - x_{k'}) \mid E' \subset E_{np} \setminus E_0 \text{ and } G' := (V, E') \text{ is acyclic}\}, \qquad (9)
\end{aligned}$$

and the right-hand side of Equation (9) is thus the optimal value for an instance of the maximum feedback arc set problem defined as $G := (V, E_{np} \setminus E_0)$. We can approximate MCI by approximating the right-hand side of Equation (9) using a polynomial-time approximation algorithm from Hassin and Rubinstein (1994). Note that the known approximation ratio for the maximum feedback arc set problem is better than that for the minimum feedback arc set problem.

References

Afriat, S.N. (1967) "The construction of utility Functions from expenditure data" *International Economic Review* **8** (1), 67–77.

Afriat, S.N. (1973) "On a system of inequalities in demand analysis: An extension of the classical method" *International Economic Review* **14** (2), 460–472.

Apesteguia, J. and M.A. Ballester (2015) "A measure of rationality and welfare" *Journal of Political Economy*, forthcoming.

Dean, M and D. Martin (2010) "How rational are your choice data?" In *Proceedings of the Conference on Revealed Preference and Partial Identification.*

Dean, M. and D. Martin (2015) "Measuring rationality with the minimum cost of revealed preference violations" *Review of Economics and Statistics*, forthcoming.

Demetrescu, C. and I. Finocchi (2003) "Combinatorial algorithms for feedback problems in directed graphs" *Information Processing Letters* **86** (3), 129–136.

Diewert, W.E. (1973) “Afriat and revealed preference theory” *The Review of Economic Studies* **40** (3), 419–425.

Eades, P. and X. Lin (1995) “A new heuristic for the feedback arc set problem” *Australian Journal of Combinatorics* **12** (1), 5–26.

Eades, P., X. Lin, and W.F. Smyth (1993) “A fast and effective heuristic for the feedback arc set problem” *Information Processing Letters* **47** (6), 319–323.

Echenique, F., S. Lee, and M. Shum (2011) “The money pump as a measure of revealed preference violations” *Journal of Political Economy* **119** (6), 1201–1223.

Evan, G., J.S. Naor, S. Rao, and B. Schieber (2000) “Divide-and-conquer approximation algorithms via spreading metrics” *Journal of the ACM* **47** (4), 585–616.

Fleissig, A.R. and G.A. Whitney (2005) “Testing for the significance of violations of Afriat’s inequalities” *Journal of Business & Economic Statistics* **23** (3), 355–362.

Fujishige, S. and Z. Yang (2012) “On revealed preference and indivisibilities” *Modern Economy* **3**, 752–758.

Garey, M.R. and D.S. Johnson (1979) *Computers and intractability: A guide to the theory of NP-completeness*, W. H. Freeman and Company, San Francisco.

Gross, J. (1995) “Testing data for consistency with revealed preference” *The Review of Economics and Statistics* **77** (4), 701–710.

Hassin, R. and S. Rubinstein (1994) “Approximations for the maximum acyclic subgraph problem” *Information Processing Letters* **51** (3), 133–140.

Heufer, J. (2008) “A geometric measure for the violation of utility maximization” Ruhr Economic Paper No. 69.

Houtman, M. and J. Maks (1985) “Determining all maximal data subsets consistent with revealed preference” *Kwantitatieve Methoden* **19**, 89–104.

Karp, R.M. (1972) “Reducibility among combinatorial problems” *Complexity of Computer Computations* **40** (4), 85–103.

Korte, B. and J. Vygen (2012) *Combinatorial optimization: Theory and algorithms*, 5th ed., Springer, Berlin.

Piaw, T.C. and R.V. Vohra (2003) “Afriat’s theorem and negative cycles” Unpublished paper.

Seymour, P.D. (1995) “Packing directed circuits fractionally” *Combinatorica* **15** (2), 281–288.

Shiozawa, K. (2015) “Revealed preference test and shortest path problem; Graph theoretic structure of the rationalizability test” Discussion Papers In Economics And Business No. 15-17-Rev., Osaka University.

Smeulders, B., L. Cherchye, F.C.R. Spieksma, and B. De Rock (2013) “The money pump as a measure of revealed preference violations: A comment” *Journal of Political Economy* **121** (6), 1248-1258.

Smeulders, B., F.C.R. Spieksma, L. Cherchye, and B. De Rock (2014) “Goodness-of-fit measures for revealed preference tests: Complexity results and algorithms” *ACM Transactions on Economics and Computation archive* **2** (1), 3.

Talla Nobion, F., B. Smeulders, and F.C.R. Spieksma (2015) “A note on testing axioms of revealed preference” *Journal of Optimization Theory and Application* **166** (3), 1063–1070.

Varian, H.R. (1982) “The nonparametric approach to demand analysis” *Econometrica* **50** (4), 945–973.

Varian, H.R. (1990) “Goodness-of-fit in optimizing models” *Journal of Econometrics* **46** (1), 125–140.