Discussion Papers In Economics And Business

Note on the goodness-of-fit measure for GARP;

NP-hardness of minimum cost index

Kohei Shiozawa

Discussion Paper 15-18

Graduate School of Economics and
Osaka School of International Public Policy (OSIPP)
Osaka University, Toyonaka, Osaka 560-0043, JAPAN

Note on the goodness-of-fit measure for GARP; NP-hardness of minimum cost index

Kohei Shiozawa

Discussion Paper 15-18

June 2015

Graduate School of Economics and
Osaka School of International Public Policy (OSIPP)
Osaka University, Toyonaka, Osaka 560-0043, JAPAN

Note on the goodness-of-fit measure for GARP; NP-hardness of minimum cost index

Kohei Shiozawa*

Abstract

The purpose of this paper is to show that the problem of computing minimum cost index (MCI), which is proposed by Dean and Martin (2010, 2015) as a goodness-of-fit measure of GARP, is NP-hard. We show the result by using a reduction from maximum acyclic subgraph problem (MASP) which is a traditional decision problem known to be NP-complete.

Keywords: Revealed Preference; GARP; Goodness of Fit Measure; Minimum Cost Index; Computational Complexity

JEL classification: C60; D11

* Graduate School of Economics, Osaka University, Machikaneyama, Toyonaka, Osaka 560-0043, Japan. E-mail: nge013sk@student.econ.osaka-u.ac.jp

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 $u : R^e ll_+ \to R$ as solutions of utility maximization problem *if and only if* the data is consistent with the generalized axiom of revealed preference (GARP). This is a classical result of Afriat (Afriat, 1962; Diewert, 1973; Varian, 1982). Based on this result, we can test an arbitrary data set $\{(p_k, x_k)\}_{k=1}^n$ whether there is a utility function that rationalize the data or not. This test has a form of a choice between two alternatives and hence, if we have some data with GARP violation, we have no information about the violation severity.

There are several severity-measure proposed in the literature and sometimes these are called *goodness-of-fit* measures. For example, Afriat (1972), Houtman and Maks (1985), Varian (1990), Echenique et al. (2011), and Sumeulders et al. (2013) proposed some indices which are defined using the data and these indices express, in various ways, how severely the data violate GARP. On the other hand, Smeulders et al. (2013, 2014) shows that for many of those indices, computation is not so easy; the problem of computing some indices are NP-hard.[1] They show that the induces of Houtman and Maks (1985), Varian (1990), and Echenique et al. (2011) are NP-hard. On the other hand, they propose an polynomial-time algorithm for Afriat (1972)'s index and two new indexes which have polynomial-time algorithms and behave in a compatible way with Echenique et al. (2011)'s index.

Recently, Dean and Martin (2010, 2015) also proposed a new index called *minimum cost index* (MCI). They suggest that the computation of MCI is not so easy. In this paper, we show that the computation problem of MCI is actually in the class NP-hard and hence. Our argument is very simple using a polynomial-time reduction from the maximum acyclic subgraph problem (MASP), a classical NP-complete problem.

We conclude this paper by proposing a new index. The definition of the new index is similar to MCI and Echenique et al. (2011)'s indices, but it has a polynomial-time exact algorithm and a natural interpretation as a goodness-of-fit measure for GARP in view of the structure of GARP test.

2 Minimum Cost Index; Definition, Validity, and Complexity

The *minimum cost index (MCI)* is a goodness-of-fit measure for GARP which is proposed by Dean and Martin (2010, 2015). MCI expresses the minimum cost of information we must ignore from the data so that the data satisfies an equivalent condition of GARP. We first formalize "an equivalent condition of GARP" using some graph theoretic apparatus.

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

[1] A problem in the class of NP-hard is not so easy in the sense that if we have a polynomial-time algorithm for the problem then every problem in the class of NP can be solved in polynomial-time through the problem. In other wards, the problem is as hard to solve as every NP problems. Moreover, at the present time, no NP problem has a polynomial-time exact algorithm and hence, we conclude that, probably, there is no polynomial-time algorithm for the problem. For more detailed discussions, see, Karp (1972), Garey and Johnson (1979), and Korte and Vygen (2008).

following directed graph: $G := (V, E_{np})$ where

$$V := \{x_1, x_2, \ldots, x_n\} \tag{1}$$
$$E_{np} := \{(x_k, x_{k'}) \,|\, k, k' = 1, \ldots, n,\ k \neq k',\ \text{and}\ \ p_k \cdot (x_{k'} - p_k) \leqq 0\}. \tag{2}$$

We call the graph $G := (V, E_{np})$ the *associated graph.*

As shown in Piaw and Vohra (2003), Fujishige and Yang (2012), and Talla Nobibon et al. (2014), we can translate GARP condition into graph theoretic conditions as stated in the following proposition.[2]

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

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

(ii) *The associated graph* $G_{np} = (V, E_{np})$ *has no strongly connected component (SCC) which has a negative edge.*

(iii) *The associated graph* $G_{np} = (V, E_{np})$ *has no cycle which has a negative edge.*

The first condition is the Cyclical consistency condition defined by Afriat (1967) and equivalent with GARP (Varian, 1987). The second condition is proposed by Talla Nobibon et al. (2014) which is also mentioned in Fujishige and Yang (2012). The third condition is a graph theoretic condition which is related to the definition of MCI as we will see below. Now, we introduce the formal definition of the MCI.

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 the minimum cost index (MCI) as follows:

$$\text{MCI} := \min\{\sum_{(x_k, x_{k'}) \in E'} p_k \cdot (x_k - x_{k'}) \mid E' \subset E_{np} \text{ and } G' = (V, E_{np} \setminus E') \text{ contains no directed cycle}\} \tag{3}$$

where $G_{np} = (V, E_{np})$ is the graph associated with the data $\{(p_k, x_k)\}_{k=1}^n$ defined in definition 1.[3]

While the original definition of MCI we introduced above is based on the idea that "the minimum cost of information we must ignore from the data so that the data satisfies an equivalent condition of GARP", it is defined by a slightly deferent condition from the equivalent condition of GARP (the condition (iii) of Proposition 1). However, the equation

$$\begin{aligned}\text{MCI} &:= \min\{\sum_{(x_k, x_{k'}) \in E'} p_k \cdot (x_k - x_{k'}) \mid E' \subset E_{np} \text{ and } G' = (V, E_{np} \setminus E') \text{ contains no directed cycle}\} \\ &= \min\{\sum_{(x_k, x_{k'}) \in E'} p_k \cdot (x_k - x_{k'}) \mid E' \subset E_{np} \text{ and } G' = (V, E_{np} \setminus E') \text{ contains no directed cycle with a negative edge}\}\end{aligned} \tag{4}$$

holds and we can see the definition of MCI actually grasps the above mentioned idea. Therefore, we

[2] Fujishige and Yang (2012) essntiallly shows this result in a indivisible goods setting (discrete setting).

[3] Dean and Martin (2010, 2015) define the MCI normalizing by the total wealth level of the data. That is, they define MCI dividing the value (3) by the value $\text{sum}_{k=1}^n p_k \dot{x}_k > 0$. Here, we omit this normalization since our goal is to see the computational complexity of MCI and this (omission of) normalization obviously does not change the argument.

can also say, from Proposition 1, that the MCI has some validity as a goodness-of-fit measure. On the other hand, the MCI is not so easy to compute. Actually, as we will see below, the problem to calculate MCI for an arbitrarily given data $\{(p_k, x_k)\}_{k=1}^n$ is NP-hard, hence we say that, probably, there is no polynomial-time algorithm for MCI.

We formally state the problem to compute MCI.

MCI Computation Problem

Instance: A data $\{(p_k, x_k)\}_{k=1}^n$ where $p_k \in R^\ell_{++}$, $x_k \in R^\ell_+$ $(k = 1, \ldots, n)$.
Task: Compute MCI defined by the equation (3).

Theorem 1. MCI computation problem is NP-hard.

Proof. We first note that the following equations hold:

$$\begin{aligned}
&\sum_{(x_k, x_{k'}) \in E_{np}} p_k \cdot (x_k - x_{k'}) - \mathrm{MCI} \\
&= \max\{ \sum_{(x_k, x_{k'}) \in E_{np}} p_k \cdot (x_k - x_{k'}) - \sum_{(x_k, x_{k'}) \in E'} p_k \cdot (x_k - x_{k'}) \mid E' \subset E_{np} \text{ and } G' := (V, E_{np} \setminus E') \text{ contains no cycle} \} \\
&= \max\{ \sum_{(x_k, x_{k'}) \in E'} p_k \cdot (x_k - x_{k'}) \mid E' \subset E_{np} \text{ and } G' := (V, E') \text{ contains no cycle} \}. \qquad (5)
\end{aligned}$$

We can observe the right hand side of this equation is a problem which asks the value of the maximum weight acyclic subgraph. This problem is called the maximum acyclic subgraph problem (MASP) and known to be NP-complete (Garey and Johnson, 1979). Formally, MASP is formalized as follows:

MASP

Instance: A simple directed graph $G = (U, E)$ and an integer $k \geqq 0$.
Question: Is there any subset $A' \subset E$ such that subgraph $G' = (U, A')$ is acyclic and $|A'| \geq k$.

We use a polynomial-time reduction from MASP. Given an arbitrary instance for MASP; a simple directed graph $G = (U, E)$ and an integer $k \geqq 0$, we construct an instance for MCI computation problem; a data $\{(p_k, x_k)\}_{k=1}^n$ where $p_k \in R^n_{++}$, $x_k \in R^n_+$ $(k = 1, \ldots, n)$ and $n := |V|$.[4] First, we define consumptions $x_k \in R^n_+$ $(k = 1, \ldots, n)$ as

$$x_k := (x_{k1}, \ldots, x_{k(k-1)}, x_{kk}, x_{k(k+1)}, \ldots, x_{kn}) := (0, \ldots, 0, 1, 0, \ldots, 0). \qquad (6)$$

[4] The idea for this construction is the same with Smeulders et al. (2014)'s one which they used to show NP-hardness of Houtman-Maks index (Houtman and Maks, 1985) and Varian's index (Varian, 1990).

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

$$p_{kk'} := \begin{cases} 1 & \text{if } (u_k, u_{k'}) \in E \\ 2 & \text{if } k = k' \\ 3 & \text{if } (u_k, u_{k'}) \notin E \end{cases} \tag{7}$$

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

$$p_k \cdot (x_{k'} - x_k) = p_{kk'} - p_{kk} = \begin{cases} -1 < 0 & \text{if } (u_k, u_{k'}) \in E \\ 1 > 0 & \text{if } (u_k, u_{k'}) \notin E \end{cases} \tag{8}$$

for all $k, k' = 1, \ldots, n$ where $k \neq k'$. Therefore, if we construct the associated graph G_{np} for this data $\{(p_k, x_k)\}_{k=1}^n$ then it is evident from equation (8) and the definition of associated graph G_{np} that there is a one-to-one correspondence between the edges in E and edges in E_{np}. Moreover, from equation (8), the equation (5) becomes

$$\begin{aligned} &\sum_{(x_k, x_{k'}) \in E_{np}} p_k \cdot (x_k - x_{k'}) - \text{MCI} \\ &= \max\{ \sum_{(x_k, x_{k'}) \in E'} p_k \cdot (x_k - x_{k'}) \mid E' \subset E_{np} \text{ and } G' := (V, E') \text{ contains no cycle} \} \\ &= \max\{ \sum_{(u_k, u_{k'}) \in E'} 1 \mid E' \subset E \text{ and } G' := (U, E') \text{ contains no cycle} \} \\ &= \max\{ \sum_{(u_k, u_{k'}) \in E'} 1 \mid E' \subset E \text{ and } G' := (U, E') \text{ contains no cycle} \}. \end{aligned} \tag{9}$$

Therefore, the following equation also holds:

$$\begin{aligned} &|E| - \text{MCI} \\ &= \max\{|E'| \mid E' \subset E \text{ and } G' := (U, E') \text{ contains no cycle}\}. \end{aligned} \tag{10}$$

Summing up, we can have a polynomial-time algorithm for MASP using MCI computation problem as a subroutine. **Step 1**: Constitute the instance for MCI computation from the given instance of MASP instance by (6) and (7). **Step 2**: Constitute the associated graph G_{np} for the data constructed in step 1. **Step 3**: Compute the the MCI which is defined by (3). **Step 4**: Compute the LHS of (10) and compare that value with the given integer k. If LHS of (10) $\geqq k$ then the answer for the MASP is yes, and otherwise, the answer is no.

The computational time for these steps can obviously be bounded by polynomial time and hence, MCI computation problem is NP-hard. □

3 Concluding Remark

As we show, the problem of computing MCI is NP-hard and as shown by Smeulders et al. (2013, 2014) many of indices proposed are also NP-hard problems. On the other hand, Smeulders et al. (2014) show that Afriat (1973)'s index has a polynomial-time algorithm and Smeulders et al. (2013) proposed an

index which has a polynomial-time algorithm and behave nicely with money-pump measure of Echenique et al. (2011). (Smeulders et al. (2013) also show that the problem of computing the money-pump measure is also NP-hard.) We propose a new index which has a polynomial-time algorithm and a natural interpretation as a goodness-of-fit measure for GARP.

All indices proposed by Echenique et al. (2011)'s money pump measure, Smeulders et al. (2013)'s measure, and MCI of Dean and Martin (2015), are defined focusing on negative cycles in G_{np}. From proposition 1, we know that each negative cycle in G_{np} creates a violation of GARP. However, also shown in proposition 1, if there is a data with a violation of GARP, the associated graph G_{np} has a negative cycle and these cycles are in a SCC of G_{np}. In other words, all the relevant informations of the data which are related to GARP violations is in SCCs of G_{np}. This observation lead us another natural goodness-of-fit measure for GARP.

$$\text{SCCI} := \frac{\sum_{h=1}^{m}(\sum_{(x_k,x_{k'})\in C_h} p_k \cdot (p_k - p_{k'}))}{\sum_{(x_k,x_{k'})\in E_{np}} p_k \cdot (p_k - p_{k'})}, \tag{11}$$

where $C_h (h = 1, \ldots, m)$ are subgraphs of G_{np} which are induced by SCC of G_{np} and $(x_k, x_{k'}) \in C_h$ expresses that the edge $(x_k, x_{k'})$ in contained the edge of C_h.[5] Note that SCCI $\in [0, 1]$ for any data. This index has a natural interpretation: the ratio of the weight of GARP violation part of the data over the entire weight of revealed preference relation. Moreover, it has a polynomial-time algorithm: **Step 1**: Construct the related graph G_{np}. **Step 2**: Compute all of the SCC by the SCCD algorithm. **Step 3**: Compute the SCCI.

Theorem 1. *The SCCI has an $O(n^2)$ algorithm.*

The computational complexity of this algorithm actually can be bounded by polynomial-time. Hence, we can compute SCCI in a reasonable time for an arbitrary data $\{(p_k, x_k)\}_{k=1}^{n}$.

References

[1] Afriat, S.N., 1967. The construction of utility functions from expenditure data. International Economic Review 8, 1, 67–77.

[2] 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.

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

[4] Dean, M., Martin, D., 2015. Measuring rationality with the minimum cost of revealed preference violations. Review of Economics and Statistics, forthcoming.

[5] Diewert, W.E.,1973. Afriat and revealed preference theory. The Review of Economic Studies 40, 3, 419–425.

[6] Echenique, F., Lee, S., Shum, M., 2011. The money pump as a measure of revealed preference

[5] Note that if the denominator $\sum_{(x_k,x_{k'})\in E_{np}} p_k \cdot (p_k - p_{k'})$ is zero, then the data is rationalizable (Proposition 1). Hence, in such cases, we define SCCI := 0.

violations. Journal of Political Economy 119, 6, 1201–1223.

[7] Fujishige, S., Yang, Z., 2012. On revealed preference and indivisibilities. Modern Economy, 3, 752–758.

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

[9] Houtman M., Maks, J., 1985. Determining all maximal data subsets consistent with revealed preference. Kwantitatieve methoden 19, 89–104.

[10] Karp, R.M.,1972. Reducibility among combinational problems. Complexity of Computer Computations 40, 4, 85–103.

[11] Korte, B., Vygen, J., 2008. Combinational optimization: Theory and algorithms. 4th ed., Springer, Berlin.

[12] Piaw, T.C., Vohra, R.V., 2003. Afriat's theorem and negative cycles. mimeo.

[13] Smeulders, B., Cherchye, L., Spieksma, F.C.R., De Rock, B., 2013. The money pump as a measure of revealed preference violations: A comment. Journal of Political Economy 121, 6, 1248-1258.

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

[15] Talla Nobion, F., Smeulders, B., Spieksma, F.C.R., 2014. A note on testing axioms of revealed preference. Journal of Optimization Theory and Application 1–8.

[16] Varian, H.R., 1982. The nonparametric approach to demand analysis. Econometrica 50, 4, 945–973.

[17] Varian, H.R., 1990. Goodness-of-fit in optimizing models. Journal of Econometrics 46, 1, 125–140.