# CORE DETERMINING CLASS: CONSTRUCTION, APPROXIMATION AND INFERENCE

By Ye Luo and Hai Wang

*University of Florida, Singapore Management University*

The relations between unobserved events and observed outcomes in partially identified models can be characterized by a bipartite graph. We propose an algorithm that explores the structure of the graph to construct the "exact Core Determining Class", i.e., the set of irredudant inequalities. We prove that if the measure on the observed outcomes are non-degenerate, the Core Determining Class does not depend on the probability measure of the outcomes but only on the structure of the graph. For more general linear inequalities selection problem with noisy outcome observations, we investigate a sparse assumption on the entire set of inequalities, i.e., only a few inequalities are truly binding. We show that the sparse assumption is equivalent to certain sparse conditions on the dual problems. We propose a statistical procedure similar to the Dantzig Selector to select the truly informative constraints. We analyze the properties of the procedure and show that the feasible set defined by the selected inequalities is a nearly sharp estimator of the true feasible set. Under our sparse assumption, we prove that such a procedure can significantly reduce the number of inequalities without losing too much information. We apply the procedure to the Core Determining Class problem and obtain stronger results by taking advantage of the structure of the bipartite graph.

We design Monte-Carlo experiments to demonstrate the good performance of our selection procedure, while the traditional CHT inference is difficult to apply in practice.

*Keywords*: Core Determining Class, Sparse Model, Linear Programming, Inequality Selection

[*]

## 1. Introduction.

Suppose we observe the outcomes but not the events that might imply the outcomes. In many situations the relations between events and outcomes are indeterministic, i.e., a single event may lead to different outcomes, and an outcome may have several events that may lead

[*] Date: The present version is of Dec 15, 2016. We thank Alexandre Belloni for providing inspiring ideas on using Farkas' Lemma to select inequalities. We thank Amedeo Odoni for improving our paper. We also thank Cynthia Barnhart, Victor Chernozhukov, Alfred Galichon, Jerry Hausman, Anna Mikusheva, Whitney Newey, and participants in MIT Econometrics seminar for many helpful discussions. Correspondence address: yeluo@ufl.edu, haiwang@smu.edu.sg.

to it. Such relations can be characterized by a bipartite graph $G = (\mathcal{U}, \mathcal{Y}, \varphi)$, where $\mathcal{U}$ is a set of unobservable events, $\mathcal{Y}$ is a set of observed outcomes, and $\varphi$ is a correspondence mapping from $\mathcal{U}$ to $\mathcal{Y}$ such that $\varphi(u) \subset \mathcal{Y}$ is the set of all possible outcomes that could be led by event $u \in \mathcal{U}$. One application is, in game theory, to infer individual player's private information given the observations of players' strategies when there exist multiple equilibria; another application is to infer demand/customer characteristics given the purchase histories and sales data.

Given such a bipartetite graph and an observed measure on $\mathcal{Y}$, the key interest of our paper is to estimate bounds on the probability measure on $\mathcal{U}$, as the feasible set of probability measure on $\mathcal{U}$ is defined by a set of linear inequality constraints. In practice, the number of inequality constraints could grow with $|\mathcal{U}|$, even exponentially. Such many inequalities may lead to two problems for performing inference on the measure on $\mathcal{U}$: (1) asymptotics in inference procedures such as those described in Chernozhukov, Hong and Tamer (2007) (later CHT) may fail; (2) those inference procedures are computationally intractable when $|\mathcal{U}|$ is large.

An inequality selection procedure may dramatically reduce the number of inequalities defining the feasible set of probability measure on $\mathcal{U}$. Such a procedure allows us to perform valid inference with much less computational cost.

(1) We propose a method to select the set of irredundant inequalities for the bipartite graph when data noise is not taken into consideration. Such set is referred as a Core Determining Class described in Galichon and Henry (2011). More specificially, we show that our selected inequalities form the "smallest" Core Determining Class. We prove that the inequalities selected are only dependent on the structure of the graph but independent from the probability measure observed on $\mathcal{Y}$ under certain mild conditions.

(2) For a general linear inequalities selection problem under noise, we propose a selection procedure similar to the Dantzig-selector described in Candes and Tao (2007). We prove that the selection procedure has good statistical properties under some sparse assumptions.

(3) We apply the selection procedure to construct the set of irredundant inequalities for the bipartite graph with data noise. We prove that the selection procedure has better statistical properties compared to that applied to the general problem due to the structure of the graph.

(4) We demonstrate the good performance of our selection procedure through several sets of Monte-Carlo experiments: first, the inference based on the selection procedure has desired size; second, it has strong power against local alternatives; third, it is relatively computationally efficient.

The closest researches to our topic are Galichon and Henry (2006, 2011) and Chesher and Rosen (2012). Galichon and Henry (2011) proposes the Core Determining Class problem, i.e., finding the minimum set of inequalities to describe the feasible region of probability measure on $\mathcal{U}$. Chesher and Rosen (2012) provides an inequality selection algorithm, but may still contain some redundant inequalities in the selected set. Andrews and Soares(2013) proposes moment inequality selection procedure using criterions such as BIC.

There are many studies on performing inference of sets. CHT (2007) proposes a general inference procedure with moment inequality constraints. Romano and Shaikh (2010) provides improvements for CHT (2007). Beresteanu, Molchanov and Molinari (2011) uses random set theory to perform inference with convex inequality restrictions. Andrews and Shi (2013) construct inference based on conditional moment inequalities. For related empirical studies, see Tamer and Manski (2002), Bajari, Benkard and Levin (2004), Bajari, Hong and Ryan (2010) and etc..

There is also a wide literature on detection and elimination of redundant constraints when data noise is not taken into consideration. For example, Telgen (1983) develops two methods to identify redundant constraints and implicit equalities. Caron, McDonald and Ponic (1989) presents a degenerate extreme point strategy which classifies linear constraints as either redundant or necessary. Paulraj, Chellappan and Natesan (2010) proposes a heuristic approach using an intercept matrix to identify redundant constraints.

We organize the paper as follows: Section 2 introduces the model and basic assumptions through out the entire paper. Section 3 studies the Core Determining Class from the structure of the bipartite graph and provides a method to construct the exact Core Determining Class when data noise is not taken into consideration. Section 4 proposes a general linear inequalities selection procedure under noisy data with the definition of sparse assumptions. Section 5 discusses the additional technical assumptions and states main theorems of the statistical properties of the selection procedure, with application to the Core Determining Class. Section 6 implements our selection procedure in a large bipartite graph through Monte-Carlo experiments and illustrates its good performance. Section 7 concludes the paper.

**2. Core Determining Class.** Given a bipartite graph $G = (\mathcal{U}, \mathcal{Y}, \varphi)$, suppose $\mathcal{U}$ is a set of vertices representing events, and $\mathcal{Y}$ is a set of vertices representing outcomes. Suppose an event $u \in \mathcal{U}$ leads to a set of possible outcomes $\varphi(u)$, where $\varphi(u)$ is a subset of $\mathcal{Y}$. For any set $A \subset \mathcal{U}$, define $\varphi(A) := \cup_{u\in A}\varphi(u)$. Therefore, $\varphi : 2^{\mathcal{U}} \mapsto 2^{\mathcal{Y}}$ is a correspondence mapping between $\mathcal{U}$ and $\mathcal{Y}$. The inverse of $\varphi$, denoted as $\varphi^{-1}$ is defined as $\varphi^{-1} : 2^{\mathcal{Y}} \mapsto$

$2^{\mathcal{U}}$, $\varphi^{-1}(B) = \{u \in \mathcal{U} | \varphi(u) \cap B \neq \emptyset\}$, $\forall B \subset \mathcal{Y}$.

Let $v$ be the probability measure on $\mathcal{U}$. Let $\mu_{n,0}$ be the true measure on $\mathcal{Y}$ which could change with the model. Let $\widehat{\mu}_n$ be the measure observed in a sample set of outcomes $\mathcal{Y}$. Denote $d_u = |\mathcal{U}|$ and $d_y = |\mathcal{Y}|$. For a graph $G = (\mathcal{U}, \mathcal{Y}, \varphi)$, say $G$ is connected if $\forall A_1, A_2 \subset \mathcal{U}$, $A_1 \cap A_2 = \emptyset$ and $A_1 \cup A_2 = \mathcal{U}$, it holds that $\varphi(A_1) \cap \varphi(A_2) \neq \emptyset$.

Assumption: C.1 (Non-Degeneracy of $G$, $\mu_{n,0}$ and $\widehat{\mu}_n$). *(1) Assume $G$ is connected. We say $G$ is non-degenerate if $G$ is connected.*

*(2) For the probability measure $\mu = \mu_{n,0}$ or $\widehat{\mu}_n$, assume that for any $y \in \mathcal{Y}$, $\mu(y) > 0$. We say that $\mu$ is non-degenerate if $\mu(y) > 0$ for any $y \in \mathcal{Y}$.*

We assume that Assumption **C.1** holds through out the paper.

The parameter of interest in this paper is the $d_1 \times 1$ vector $v$, which is the probability measure which generates the events $u \in \mathcal{U}$. In general we are unable to obtain a point estimation of $v$ unless additional information is provided. Instead, we can obtain inequality bounds on $v$ given the bipartite graph $G = (\mathcal{U}, \mathcal{Y}, \varphi)$ and the measure $\mu$ on $\mathcal{Y}$. More specifically, for any set of events $A \subset \mathcal{U}$, the outcome should fall into the set $\varphi(A)$. Thus, for any $A \subset \mathcal{U}$, we can obtain the inequality $v(A) := \sum_{u \in A} v(u) \leqslant \mu(\varphi(A)) := \sum_{y \in \varphi(A)} \mu(y)$.

The Artstein's theorem stated in Artstein (1983) presents that all information of $v$ in the biparte graph model $G = (\mathcal{U}, \mathcal{Y}, \varphi)$ is characterized by the set of constraints described below:

Lemma 1 (Artstein's Theorem). *The following set of inequalities/equalities contains sharp information on $v$:*

1. *For any $A \subset \mathcal{U}$,*
$$v(A) := \sum_{u \in A} v(u) \leqslant \mu(\varphi(A)),$$
*where $\mu(\varphi(A)) := \sum_{y \in \varphi(A)} \mu(y)$;*
2. $\sum_{u \in \mathcal{U}} v(u) = 1$.

Our model, denoted as $\mathcal{P}_G$, is presented below:

Definition 2.1 (Problem $\mathcal{P}_G$). *Find the set of all feasible probability measure $v$ on $\mathcal{U}$ such that:*

*(1) For any $A \subset \mathcal{U}$, $v(A) \leqslant \mu(\varphi(A))$;*

*(2) $\sum_{u \in \mathcal{U}} v(u) = 1$.*

Comment 2.1. *The non-degeneracy assumption prevents the problem $\mathcal{P}_G$ from decomposition, i.e., we can not decompose graph $G$ into $G_1$ and $G_2$ and proceed with problem $\mathcal{P}_{G_1}$ and $\mathcal{P}_{G_2}$. Otherwise the problem can be simplified by looking at $G_1$ and $G_2$ separately.*

In general, the set of inequality constraints stated in Definition 2.1 may contain redundant inequalities. Define the minimum model $T_0$ of $\mathcal{P}_G$ as the set of linear constraints stated in (1) such that $T_0$ together with the equality (2) has the minimum number of constraints which generate the same set of feasible measure as $\mathcal{P}_G$. In other words, $T_0$ consists of all irredundant constraints in $\mathcal{P}_G$. If the number of irredundant constraints in $T_0$ is much less than $2^{d_1} - 1$ stated in Definition 2.1, then it is more accurate and computational efficient to conduct inference on the Core Determining Class using $T_0$. Galichon and Henry (2011) proposes the concept "Core Determining Class" as follows.

Definition 2.2 (Core Determining Class problem). *The Core Determining Class problem is the problem of finding all binding constraints in model $\mathcal{P}_G$. The Core Determining Class is any collection of subsets of $\mathcal{U}$ that contains the sharp information. The exact Core Determining Class is defined as the set of subsets of $\mathcal{U}$ which corresponds to the irredundant inequalities in $T_0$.*[1]

Comment 2.2. *In many cases there may exist a parametric model for $v$, denoted as $v_i = F_i(\theta)$. The function $F_i$ can be non-linear. The inference on $\theta$ can be generally difficult if the number of inequalities about $v$ is large. Therefore, we can find the truly binding inequalities about $v$, we would perform estimation and inference on $\theta$ much faster.*

We provide an example on the model $\mathcal{P}_G$.

Example 1 (Two players entry game). *Suppose there are two firms in a market. The cost for firm 1 and firm 2 is $c+r_1$ and $c+r_2$ respectively, where $c$ is a constant, $r_1$ and $r_2$ are random shocks which are observable only by the corresponding firm.*

*The two firms face a total demand $D = a_1 - a_2 p$. If they are both in the market, they will play a Cournot Nash equilibrium. If there is only one*

[1] The definition of Core-Determining Class in Galichon and Henry (2006) is slightly different from ours. Galichon and Henry (2006) defines Core-Determining Class as any set that contains all the binding inequalities. In this paper, we refer "exact Core-Determining Class" as the set of binding inequalities, i.e., the smallest set (in cardinality) which characterizes the identified set of parameter of interest.

player, then this player will reach a monopolist's equilibrium. If the costs are too large for both players that even a monopolist is unprofitable, then there will be no player in the market. Therefore, there are 4 possible equilibria: $(0,0)$, $(1,0)$, $(0,1)$, and $(1,1)$:

(1) if $\frac{a_1}{a_2} - c \geqslant 2/3r_1 - 1/3r_2$ and $\frac{a_1}{a_2} - c \geqslant 2/3r_2 - 1/3r_1$, then the equilibrium is $(1,1)$;

(2) if $\frac{a_1}{a_2} - c < 2/3r_1 - 1/3r_2$ and $\frac{a_1}{a_2} - c \geqslant 2/3r_2 - 1/3r_1$, then the equilibrium is $(0,1)$;

(3) if $\frac{a_1}{a_2} - c \geqslant 2/3r_1 - 1/3r_2$ and $\frac{a_1}{a_2} - c < 2/3r_2 - 1/3r_1$, then the equilibrium is $(1,0)$;

else if $\frac{a_1}{a_2} - c < 2/3r_1 - 1/3r_2$ and $\frac{a_1}{a_2} - c < 2/3r_2 - 1/3r_1$:

(4) if $c + r_1 \leqslant \frac{a_1}{a_2}$ and $c + r_2 \leqslant \frac{a_1}{a_2}$, then there are two equilibria:$(1,0)$ and $(0,1)$;

(5) if $c + r_1 \leqslant \frac{a_1}{a_2}$ and $c + r_2 > \frac{a_1}{a_2}$, then the equilibrium is: $(1,0)$;

(6) if $c + r_1 > \frac{a_1}{a_2}$ and $c + r_2 \leqslant \frac{a_1}{a_2}$, then the equilibrium is: $(0,1)$;

(7) if $c + r_1 > \frac{a_1}{a_2}$ and $c + r_2 > \frac{a_1}{a_2}$, then the equilibrium is: $(0,0)$.

Let $\mathcal{U} = \{u_1, u_2, u_3, u_4, u_7\}$, where $u_i$ is the event representing case $(i)$, with the exceptions that $u_2$ represents (2) and (6), and $u_3$ represents (3) and (5). Let $Y := \{y_1, y_2, y_3, y_4\}$, where $y_1 = (1,1)$, $y_2 = (0,1)$, $y_3 = (1,0)$, and $y_4 = (0,0)$. So $d_1 = |\mathcal{U}| = 5$ and $d_2 = |\mathcal{Y}| = 4$. The correspondence mapping $\varphi$ between $\mathcal{U}$ and $\mathcal{Y}$ is:

$\varphi(u_1) = \{y_1\}$, $\varphi(u_2) = \{y_2\}$, $\varphi(u_3) = \{y_3\}$, $\varphi(u_4) = \{y_2, y_3\}$, and $\varphi(u_7) = \{y_4\}$.

The correspondence mapping for Example 1 is illustrated in Figure 1.

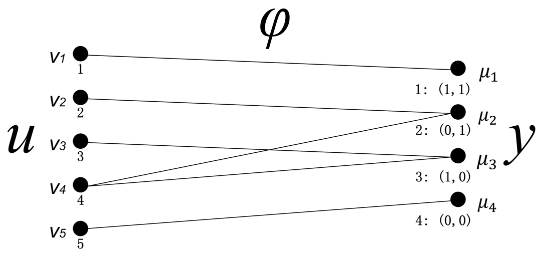

FIG 1. Correspondence Mapping for Example 1

Given the probability measure $\mu$ on $\mathcal{Y}$, the bounds of the probability measure $v$ on $\mathcal{U}$ is given by the inequalities stated in the Artstein's theorem. According to the Artstein's theorem statement (1), there are $2^5 - 2 = 30$ inequalities. In fact, it is obvious that the Core-Determining Class in this *example consist of only* 5 *sets (inequalities):* $\{u_1\}$, $\{u_2\}$, $\{u_3\}$, $\{u_2, u_3, u_4\}$ *and* $\{u_5\}$.

We observe $\widehat{\mu}_n$ on $\mathcal{Y}$ instead of the true measure $\mu_{n,0}$. Due to uncertainty of the data, we would like to solve a relaxed problem $\mathcal{P}'_G$, whose solution set covers the solution set of the true model $\mathcal{P}_G$ with probability approaching 1 as the data sample size $n$ approaching infinity. This relaxed problem $\mathcal{P}'_G$ provides conservative inference for model $\mathcal{P}_G$.

Definition 2.3 (Problem $\mathcal{P}'_G$). *For a small* $\lambda$*, find the set of all feasible probability measure* $v$ *on* $\mathcal{U}$ *such that:*
*(1) For any* $A \subset \mathcal{U}$, $v(A) := \sum_{u \in A} v(u) \leqslant \widehat{\mu}_n(\varphi(A)) + \lambda$;
*(2)* $\sum_{u \in \mathcal{U}} v(u) = 1$.

Ideally $\lambda$ should converge to 0 when $n \to \infty$. The dimensionality of the problem, $|\mathcal{U}|$, and the number of inequalities in $\mathcal{P}'_G$, should affect the tuning parameter $\lambda$. In fact, $\lambda$ should be chosen properly such that: (1) the feasible set of $v$ found in model $\mathcal{P}'_G$ covers the feasible set of $v$ found in model $\mathcal{P}_G$ with probability approaching 1, so $\mathcal{P}'_G$ provides inference on $\mathcal{P}_G$; and (2) $\lambda$ is not be too large to exaggerate the feasible set of $v$ found in model $\mathcal{P}'_G$. We will discuss the choice of $\lambda$ in Section 5.

According to the Artstein's theorem, model $\mathcal{P}_G$ contains $2^{d_1} - 2$ inequalities. It is a very large number when $d_1$ is large and even grows with $n$ in some contexts. The numerous inequalities lead to both computational difficulties and undesirable statistical properties. In fact, some or even most of the inequalities stated in the Artstein's theorem may be redundant. Galichon and Henry (2011) analyzes the monotonic structure of the graph $G$ and claims that there are at most $2d_1 - 2$ sets in the Core Determining Class under a special structure. Chesher and Rosen (2012) provides an algorithm which could get rid of some, but not necessarily all redundant inequalities. In Section 3, we fully characterize the Core Determining Class by the exploring the combinatorial structure of the bipartite graph $G$. We prove that the Core Determining Class only rely on the structure of $G$ under the non-degeneracy assumption of $\mu$. The results are novel compared to existing studies. We also propose a fast algorithm in Section 3 to compute the exact Core Determining Class when data noise is not taken into consideration.

In addition, besides those redundant inequalities, many of the binding inequalities could be "nearly" redundant, meaning that although they are informative in $\mathcal{P}_G$ under the empirical measure $\widehat{\mu}$, they could be "implied" by other inequalities in $\mathcal{P}'_G$ with a small relaxation of $\lambda$. Therefore, it may be possible to use a smaller number of inequalities, i.e., a "small" model, to approximate the full one, with the approximation error controlled by $\lambda$. Such a small model may enjoy better statistical properties compared to the full model, i.e., it will be less sensitive to modeling errors. We propose an general inequality selection procedure similar to the Dantzig Selector in Section 4.

## 3. Exact Core Determining Class.

In this section, we present our discovery of the combinatorial structure of the Core Determining Class, along with a fast algorithm to generate the Core Determining Class. In Galichon and Henry (2011), whether an inequality $v(A) \leqslant \mu(\varphi(A))$ is in the Core Determining Class is examined by numerical computation using the probability measure $\mu$.

Given the correspondence mapping $\varphi$ of the bipartite graph $G = (\mathcal{U}, \mathcal{Y}, \varphi)$, we can identify the redundant inequalities without any observations of the outcomes in $\mathcal{Y}$. For example, for $A_1 \in \mathcal{U}$ and $A_2 \in \mathcal{U}$, if $A_1 \cap A_2 = \emptyset$ and $\varphi(A_1) \cap \varphi(A_2) = \emptyset$, then the two inequalities, $v(A_1) \leqslant \mu(\varphi(A_1))$ and $v(A_2) \leqslant \mu(\varphi(A_2))$ can generate the inequality $v(A_1 \cup A_2) = v(A_1) + v(A_2) \leqslant \mu(\varphi(A_1)) + \mu(\varphi(A_2)) = \mu(\varphi(A_1) \cup \varphi(A_2)) = \mu(\varphi(A_1 \cup A_2))$, which is exactly the inequality corresponding to $A = A_1 \cup A_2$. In another word, the inequality $v(A) \leqslant \mu(A)$ is redundant given $v(A_1) \leqslant \mu(\varphi(A_1))$ and $v(A_2) \leqslant \mu(\varphi(A_2))$. Also, if $u \notin A$ satisfies $\varphi(\{u\}) \subset \varphi(A)$, then the inequality $v(\{u\} \cup A) \leqslant \mu(\varphi(\{u\} \cup A))$ will imply a redundant inequality $v(A) \leqslant \mu(\varphi(A))$.

Without loss of generality, we can assume that the graph $G$ is connected. Otherwise we can split the graphs into connected branches and our theory applies for every connected branch.

In this section, we propose a combinatorial method to generate the exact Core Determining Class. We prove that, in theory, if the probability measure $\mu$ is non-degenerate, our method excludes all redundant inequalities in the model $\mathcal{P}_G$ regardless the values of $\mu$. That is to say, the Core Determining Class can be exactly constructed with the method and the Core Determining Class is independent from $\mu$.

Definition 3.1 (Set $\mathcal{S}_u$). *$\mathcal{S}_u \subset 2^{\mathcal{U}}$ is the collection of all non-empty subsets $A \subset \mathcal{U}$ and $A \neq \mathcal{U}$, such that*

$$v^M(A) > \mu(\varphi(A)),$$

*where $v^M(A) := \max\{v(A) | v(A') \leqslant \mu(\varphi(A')), \forall A' \subset \mathcal{U}, A' \neq A\}$.*

Set $\mathcal{S}_u$ is defined with probability measure $\mu$. The inequality generated by any $A \in \mathcal{S}_u$ is informative: it is irredundant given other inequalities described in statement (1) of the Artstein's theorem. Essentially, $\mathcal{S}_u$ identifies the irreducible inequalities for Model $\mathcal{P}_G$ when the critical equality $\sum_{u\in\mathcal{U}} v(u) = 1$ is not taken into consideration.

Definition 3.2 (Set $\mathcal{S}'_u$). *$\mathcal{S}'_u \subset 2^{\mathcal{U}}$ is the collection of all non-empty subsets $A \subset \mathcal{U}$ and $A \neq \mathcal{U}$, such that:*

*(1)A is self-connected, i.e., $\forall A_1, A_2 \subset A$ such that $A_1, A_2 \neq \emptyset$ and $A_1 \cup A_2 = A$, it holds that $\varphi(A_1) \cap \varphi(A_2) \neq \emptyset$;*

*(2)There exists no $u \in \mathcal{U}$, such that $u \notin A$ and $\varphi(u) \subset \varphi(A)$.*

Lemma 2. *If $\mu$ is non-degenerate, the collection of subsets defined in Definition 3.1 and Definition 3.2 are identical. $\mathcal{S}_u = \mathcal{S}'_u$.*

By definition $\mathcal{S}_u$ describes the irredundant inequalities in $\mathcal{P}_G$ if the equality $\sum_{u\in\mathcal{U}} v(u) = 1$ is not taken into consideration. Lemma 2 shows that the irredundant inequalities can be described via rules (1) and (2) stated in Definition 3.2. Theorem 5 of Chesher and Rosen (2012) selects a subset of inequalities which is equivalent to those corresponded to $S'_u$, which is a Core Determining class, but not neccessarily the smallest. Lemma 2 shows that with an additional property (2) in Definition 3.2, we can find all binding inequalities without considering the equality: $\sum_{u\in\mathcal{U}} v(u) = 1$.

To find the minimum set of irredudant inequalities, i.e., the exact Core-Determining Class, we consider the same bipartite graph $G$, but correspondence $\varphi^{-1}$ mapping from $2^{\mathcal{Y}} \mapsto 2^{\mathcal{U}}$. For any non-degenerate probability measure $\widetilde{v}$ on $\mathcal{U}$, we define $\mathcal{S}_y$ and $\mathcal{S}'_y$ as the following:

Definition 3.3 (Set $\mathcal{S}_y$). *Given a non-degenerate probability measure $\widetilde{v}$ on $\mathcal{U}$, $\mathcal{S}_y \subset 2^{\mathcal{Y}}$ is the collection of all subsets $B \subset \mathcal{Y}$ and $B \neq \mathcal{Y}$, such that*

$$\mu^M(B) > \widetilde{v}(\varphi^{-1}(B)),$$

*where $\mu^M(B) := \max\{\widetilde{\mu}(B) | \widetilde{\mu}(B') \leqslant \widetilde{v}(\varphi^{-1}(B')), \forall B' \subset \mathcal{Y}, B' \neq B\}$, where $\widetilde{\mu}$ is a probability measure on $\mathcal{Y}$.*

Definition 3.4 (Set $\mathcal{S}'_y$). *$\mathcal{S}'_y \subset 2^{\mathcal{Y}}$ is the collection of all subsets $B \subset \mathcal{Y}$ and $B \neq \mathcal{Y}$, such that:*

*(1) B is self-connected, i.e., $\forall B_1, B_2 \subset B$, such that $B_1, B_2 \neq \emptyset$ and $B_1 \cup B_2 = B$, it holds that $\varphi^{-1}(B_1) \cap \varphi^{-1}(B_2) \neq \emptyset$;*

*(2) There exists no $y \in \mathcal{Y}$, such that $y \notin B$ and $\varphi^{-1}(y) \subset \varphi^{-1}(B)$.*

The Lemma below presents result similar to Lemma 2.

Lemma 3. *The collection of subsets defined in Definition 3.3 and 3.4 are identical, i.e., $\mathcal{S}_y = \mathcal{S}'_y$.*

Definition 3.5 (Set $\mathcal{S}_y^{-1}$). *Set $\mathcal{S}_y^{-1}$ is the collection of $A \subset \mathcal{U}$ and $A \neq \mathcal{U}$ such that there exists $B \subset \mathcal{S}_y'$ that $A = \varphi^{-1}(B)^c$.*

Below we give a numerical definition of the exact Core Determining Class using linear programming:

Definition 3.6 (Set $\mathcal{S}^*$). *The Exact Core Determining Class $\mathcal{S}^*$ is the collection of all subsets $A \subset \mathcal{U}$ and $A \neq \mathcal{U}$, such that*

$$v^{M^*}(A) > \mu(\varphi(A)),$$

*where $v^{M^*}(A) := \max\{v(A)|v(A') \leqslant \mu(\varphi(A')), \forall A' \subset \mathcal{U}, A' \neq A; v(\mathcal{U}) = 1\}$.*

The theorem below characterizes the exact Core Determining Class $\mathcal{S}^*$:

Theorem 1. *The exact Core Determining Class is characterized by the following equation:*

$$\mathcal{S}^* = \mathcal{S}_u \cap \mathcal{S}_y^{-1}$$

Notice that both $\mathcal{S}_u$ and $\mathcal{S}_y^{-1}$ are defined via combinatorial rules, the Core Determining Class is independent from $\mu$ if $\mu$ is non-degenerate.

In Example 2, we show that $S_u$ or $S_y^{-1}$ may not able to substantially reduce the number of inequalities, while $S_u \cap S_y^{-1}$ can be a very small set in cardinality.

Example 2. *Consider set $\mathcal{U} = \{u_1, ..., u_5\}$ and set $\mathcal{Y} = \{y_1, ..., y_4\}$. $\varphi$ is the correspondence mapping between $\mathcal{U}$ and $\mathcal{Y}$ such that: $\varphi(u_j) = \{y_j\}$ for all $1 \leqslant j \leqslant 4$ and $\varphi(u_5) = \{y_1, y_2, y_3, y_4\}$.*

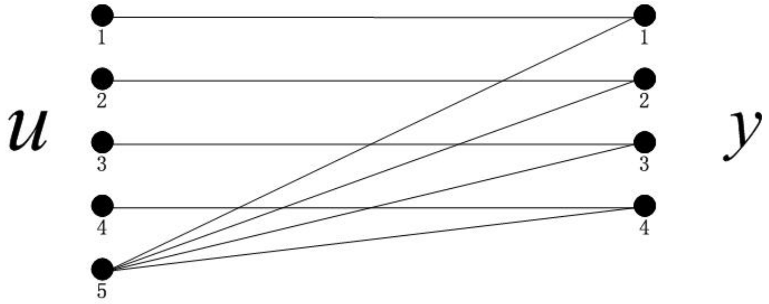

Fig 2. *Correspondence Mapping of Example 2*

*In this example, $\mathcal{S}_y^{-1} = \{\mathcal{U}_1 | \mathcal{U}_1 \subset \{u_1, u_2, u_3, u_4\}, \mathcal{U}_1 \neq\}$, which consists of $2^4 - 2$ subsets, while $\mathcal{S}_u = \{u_j | 1 \leqslant j \leqslant 4\}$. The Core Determining Class*

*$\mathcal{S}^*$ constructed in our approach is $\{u_j | 1 \leqslant j \leqslant 4\}$. It is obvious that this is the minimum class of subsets that is carrying the full information for $\mathcal{P}_G$.*

We utilize the combinatorial structure revealed in Definition 3.2 and Definition 3.4 to construct $\mathcal{S}'_u$ and $\mathcal{S}'_y$: algorithm 1 computes $\mathcal{S}'_u$ and a similar algorithm computes $\mathcal{S}'_y$ and $\mathcal{S}_y^{-1}$. Then we obtain the Core Determining Class $\mathcal{S}^* = \mathcal{S}_u \cap \mathcal{S}_y^{-1}$.

The complexity of the algorithm is $o(2^{\max(d_u,d_y)} \cdot d_u^2 \cdot d_y^2)$, where

$d_u$ is defined as

$$d_u := \max_A |A|$$

$$s.t. A \subset \mathcal{U}$$

$$\varphi(A) = \mathcal{Y}$$

$$\varphi(A/u) \subsetneq \mathcal{Y}, \forall u \in A$$

$d_y$ is defined as

$$d_y := \max_B |B|$$

$$s.t. B \subset \mathcal{Y}$$

$$\varphi^{-1}(B) = \mathcal{U}$$

$$\varphi^{-1}(B/y) \subsetneq \mathcal{U}, \forall y \in B$$

Under the assumption of non-degenerate $G$ and $\mu$, in a bipartite graph with practical application, $d_u$ and $d_y$ is much smaller than $d_u$ and $d_y$ respectively, so the algorithm is fast in practice.

```
input  : Bipartite graph $G = (\mathcal{U}, \mathcal{Y}, \varphi)$
output: Set $\mathcal{S}'_u$
Initiation: $\mathcal{S}'_u = \{\emptyset\}$
for $i \leftarrow 0$ to $|\mathcal{U}| - 1$ do
    Identify additional $A' \in \mathcal{S}'_u$ as union of $u \in \mathcal{U}$ and $A \in \mathcal{S}'_u$ with $|A| = i$,
    foreach $A \in \mathcal{S}'_u$ with $|A| = i$ do
        foreach $u \notin A$ that $\varphi(u) \cap \varphi(A) \neq \emptyset$ do
            $A' \leftarrow A \cup \{u\}$,
            if $\varphi(A') < 1$ then
                foreach $u' \notin A'$ do
                    if $\varphi(u') \subset \varphi(A')$ then $A' \leftarrow A' \cup u'$;
                end
                if $A' \notin \mathcal{S}'_u$ then $\mathcal{S}'_u \leftarrow \mathcal{S}'_u \cup \{A'\}$;
            end
        end
    end
end
Termination: $\mathcal{S}'_u \leftarrow \mathcal{S}'_u - \{\emptyset\}$
```

**Algorithm 1:** Generate set $\mathcal{S}'_u$

**4. A general selection procedure and sparse assumption.** $\mathcal{P}_G$ defines a feasible set of $v$ given observation $\widehat{\mu}$, i.e., $\widehat{Q} := \{v|v(A) \leqslant \widehat{\mu}_n(\varphi(A)), \forall A \subset \mathcal{U};\ \sum_{u \in \mathcal{U}} v(u) = 1, v \geqslant 0\}$. In Section 3 we explore the structure of the bipartite graph $G$ to obtain the set of irredundant inequalities that defines $\widehat{Q}$. In this section, we propose a general procedure for selecting linear inequalities under noise. This procedure chooses a subset of linear inequalities which defines a region approximating $\widehat{Q}$ as $n \to \infty$. It can possibly delete some inequalities which are indeed binding but "close" to redundant, so to further reduce the number of inequalities selected. The procedure can be applied to linear inequality selection, including the Core Determining Class problem allowing mixed strategy as defined in Galichon and Henry (2011).

4.1. *General Selection Procedure.* Problem $\mathcal{P}$ can be characterized as computing the feasible region of a collection of linear inequality constraints. In general, we consider the following setting:

$$Q := \{v|Mv \leqslant b, v \geqslant 0\},$$

where $M$ is a $m \times d_u$ matrix, $v$ is $d_u \times 1$ vector, and $b$ is a $m \times 1$ vector. For each $j = 1, 2, ..., m$, we can define $M_k$ as the $j^{th}$ row of matrix $M$. Then, $Mv \leqslant b$ includes $m$ inequalities: $M_j v \leqslant b_j$, $j = 1, 2, ..., m$. Our key interest is to estimate $Q$ and select the irredudant inequalities.

In some applications the number of inequalities, $m$, is too large to effectively conduct any known estimation and inference procedure, both theoretically and computationally. For example, there are $m = 2^{d_u}$ inequalities in the Core Determining Class problem without any inequalities selection procedures.[2] There are two reasons to not to use the entire set of inequalities: first, there could be many redundant inequalities which are not informative; second, the compuational cost of using all inequalities can be large.

Notice that the random noise of $b$, which comes from $\widehat{\mu}_n - \mu_{n,0}$, does not affect the exact Core-Determing Class as proved in Section 3. Such a result is due to the bipartite graph structure that is forming all the inequalities. This is no longer true for arbitrary set of linear inequalities.

For any subset $\mathcal{I}$ of $\{1, 2, ..., m\}$, denote $M_{\mathcal{I}}$ as the matrix comprised of the rows indexed by $\mathcal{I}$ in matrix $M$. Similarly, denote $b_{\mathcal{I}}$ as the subvector of $b$ indexed by $\mathcal{I}$.

By the Farkas Lemma, for a general matrix $M$ and a vector $b$, if the set of constraints indexed by $\mathcal{I}$ can imply all other constraints, i.e., the set $Q_{\mathcal{I}} := \{v | M_{\mathcal{I}} v \leqslant b_{\mathcal{I}}, v \geqslant 0\}$ equals $Q := \{v | Mv \leqslant b, v \geqslant 0\}$, then there must exist a non-negative $m \times m$ matrix $\Pi$ such that:

$$(1)\ \Pi M \geqslant M,$$

$$(2)\ \Pi b \leqslant b,$$

$$(3)\ \Pi_{jk} = 0,$$

for any $1 \leqslant j \leqslant m$ and $k \notin \mathcal{I}$.

For any $j, k \in \{1, 2, ..., m\}$, denote $\Pi_{j*}$ as the $j^{th}$ row of $\Pi$, $\Pi_{*k}$ as the $k^{th}$ column of $\Pi$, and denote $\Pi_{jk}$ as the $(j, k)^{th}$ entry of the matrix $\Pi$.

The coefficient matrix $\Pi$ described above can serve as a signal of the importance of each inequality: $\Pi_{jk}$ indicated the contribution of $k^{th}$ inequality in reconstructing $j^{th}$ inequality.

Instead of observing $(M, b)$, in practice we assume that we observe $(M, \widehat{b})$ where $\widehat{b}$ is an estimator of $b$, e.g., the frequency of outcome observed from data. We define $\widehat{Q} := \{v | Mv \leqslant \widehat{b}\}$ and $\widehat{Q}_{\mathcal{I}} := \{v | M_{\mathcal{I}} v \leqslant \widehat{b}_{\mathcal{I}}\}$, where $\mathcal{I}$ is any subset of $\{1, 2, ..., m\}$.

If all the coefficients associated with the $k^{th}$ inequality, i.e., all entries in the $m \times 1$ vector $\Pi^k$ are zeros, then this inequality is redundant given the rest of the inequalities. We propose the following Dantzig-Selector type of selection procedure based on Farkas Lemma:

Procedure $\widehat{\mathcal{R}}_0$ :

$$\min_{\Pi} \sum_{k=1}^{m} g(\Pi_{*k})$$

[2] We could view $v(\mathcal{U}) = 1$ as two inequalities: $v(\mathcal{U}) \leqslant 1$ and $v(\mathcal{U}) \geqslant 1$.

subject to:

$$(1)\Pi M \geqslant M, \Pi \geqslant 0,$$
$$(2)\Pi(\widehat{b} - \Lambda_n) + 2diag(\Pi)\Lambda_n \leqslant \widehat{b} + \Lambda_n,$$

where the observed $\widehat{b}$ is a $m \times 1$ vector, $\Lambda_n = (\lambda_{n,1}, \lambda_{n,2}, ..., \lambda_{n,m})'$ in which $\lambda_{n,i}$ is a relaxing parameter measuring the maximum violation allowed for the $i^{th}$ inequality, and $diag(\Pi)$ is the diagonal matrix $diag(\Pi_{11}, \Pi_{22}, ..., \Pi_{mm})$.[3]

Comment 4.1. The Dantzig Selector allows us to identify the parameter pointwisely. In our case, since the bounds are inequalities, therefore an estimated set is produced by the procedure.

We choose the objective function $g(\cdot)$ such that it measures the overall importance of the $i^{th}$ inequality. One choice is $g_0(\Pi_{*k}) = sign(\Sigma_{1 \leqslant j \leqslant m}\Pi_{jk})$. With the function $g_0(\cdot)$, the selection procedure $\widehat{\mathcal{R}}$ is a binary integer programming problem. Define $\widehat{\mathcal{I}}_{L0} := \{k : g_0(\Pi_{*k}) > 0, k = 1, 2, ..., m\}$ as the set of inequalities selected. We call $\widehat{\mathcal{I}}_{L0}$ as the "$L-0$ selector", and $\widehat{\mathcal{R}}$ with $g = g_0$ as "$L-0$" selection procedure.

The $L-0$ selector is difficult to compute when $m$ is large. Below we propose a computational tractable function for the choice of the objective function $g(\cdot)$:

$$g_\infty(\Pi_{*k}) := \max_{1 \leqslant j \leqslant m} \Pi_{jk}, \tag{4.1}$$

where $\Pi_{jk}$ is the $(j,k)^{th}$ entry of $\Pi$.

With $g(\cdot) = g_\infty(\cdot)$, we define the procedure $\widehat{\mathcal{R}}_{L_1}$ as the following:

Procedure $\widehat{\mathcal{R}}_{L_1}$ :

$$\min_{\Pi} \sum_{k=1}^{m} \max_{1 \leqslant j \leqslant m} \Pi_{jk}$$

subject to:

$$(1)\ \Pi M \geqslant M, \Pi \geqslant 0,$$
$$(2)\Pi(\widehat{b} - \Lambda_n) + 2diag(\Pi)\Lambda_n \leqslant \widehat{b} + \Lambda_n.$$[4]

[3] For the $j^{th}$ inequality which is believed to be important, we could set the corresponding $\lambda_{n,j}$ to 0.

[4] The formulation of the problem $\widehat{\mathcal{R}}$ is similar to the Dantzig Selector described in Candes and Tao (2005). The main difference is that the Dantzig Selector has two-sided constraints, which shrink the feasible solution to a point, while our problem has one-sided constraints, which consign the feasible solution to a convex set. The benefit of this formulation is that it turns an integer programming problem (minimize $L-0$ norm) into a linear programming problem (minimize $L-1$ norm).

Denote the solution of the Procedure $\widehat{\mathcal{R}}_{L_1}$ as $\widehat{\Pi}^{L1}$. Define $\widehat{\mathcal{I}}_{L_1} := \{k : g_\infty(\widehat{\Pi}^{L1}_{*k}) > 0\}$ as the set of inequalities being selected. For simplicity, we denote $\widehat{\mathcal{I}} := \widehat{\mathcal{I}}_{L_1}$. In next subsection, we study the property of $\widehat{\mathcal{I}}$ under certain sparsity assumptions.

4.2. *Selection Property of $\widehat{\mathcal{R}}_{L_1}$ and Inference on $Q$.* Sparse assumptions play the essential role in the analysis of some $L-1$ penalization procedures, such as LASSO and the Dantzig Selector.

By Farkas Lemma, there exists a $m \times d$ matrix, denoted as $\widetilde{\Pi}$, such that:

$$(a)\widetilde{\Pi}M \geqslant M, \widetilde{\Pi} \geqslant 0,$$

$$(b)\widetilde{\Pi}b \leqslant b,$$

$$(c)\widetilde{\Pi}_{*k} = 0 \text{ if } k \notin \mathcal{I}_0,$$

$$(d)\widetilde{\Pi}_{jj} \leqslant 1, \text{ for any } j \in \{1, 2, ..., m\}.$$

For any $m \times m$ matrix $\Pi$ and a function $g(\cdot)$, denote $g(\Pi) := (g(\Pi_{*1}), ..., g(\Pi_{*m}))'$, which is a $m \times 1$ vector. Denote $s_0 := |\mathcal{I}_0|$ as the number of truly informative inequalities.

We consider the following sparsity assumption.

For any $1 \leqslant j \leqslant m$, define separation of inequality $j$ as:

$$c_j := \max_{v \in Q_j} M_j v - b_j,$$

where

$$Q_j := \{v | M_i v \leqslant b_i, \forall i \neq j; v \geqslant 0\}$$

$c_j$ measure the maximal separation of the $j^{th}$ inequality for all points in $Q_j$. If $c_j > 0$, the $j^{th}$ inequality is irredundant, otherwise the $j^{th}$ inequality is redundant. Let $T_0$ be the set of indices $j$ with $c_j > 0$ to denote the set of irredundant inequalities. Since $c_j$ characterizes the information carried by the $j^{th}$ inequality, we define a sparse assumption using $c_j$.

Assumption: C.2 (Exact Sparse). *Recall that $T_0$ is the subset of $\{1, 2, ..., m\}$ denoting all irredundant inequalities. Let $\widetilde{\Pi}$ be a solution of the following problem:*

*Problem $\mathcal{R}$ :*

$$\min_{\Pi} \sum_{k \in T_0} g_\infty(\Pi_{*k})$$

*subject to:*

$$(a)\Pi M \geqslant M, \Pi \geqslant 0,$$

$$(b)\Pi b \leqslant b,$$

$$(c)\Pi_{*k} = 0 \textit{ if } k \notin \mathcal{I}_0.$$

*The exact sparse assumption is defined below:*

*There exists absolute positive constants $K_L$ and $K_\infty$, and a constant $c_{g,n}$ which may depend on $n$, such that:*

*(1) $s_0 := |\mathcal{I}_0| = o(n \wedge m)$, which may increase with $n$.*

*(2) The sum of coefficients needed to construct each inequality is bounded: $\max_{1 \leqslant j \leqslant m} ||\widetilde{\Pi}_{j*}||_{L^1} \leqslant K_L$;*

*(3) $\max_{1 \leqslant j \leqslant m} g_\infty(\widetilde{\Pi}_{*j}) \leqslant K_\infty$ where $K_\infty$ is a constant;*

*(4) $\min_{j \in \mathcal{I}_0} c_j \geqslant c_{g,n}$.*

Comment 4.2. Assumption 2 (1) restricts the number of informative inequalities to be small. Assumption 2 (2) allows us to reconstruct any inequalities using the inequalities in $\mathcal{I}_0$ with bounded coefficients.

Under exact sparse condition, we are able to derive the following Lemma that describes the relationship between the $L-1$ selector compared to the set $\mathcal{I}_0$ when there is no noise.

Assumption: C.3 (Dominance of $\lambda$). *Suppose we have data $b^1, b^2, ..., b^n$ with dimension being $m \times 1$ such that $b = \mathbb{E}[b^i]$, $1 \leqslant i \leqslant n$. In practice we observe $\widehat{b} := \mathbb{E}_n[b^i]$ as an estimator of $b$. Suppose that with probability at least $1 - \alpha$,*

*(1) $(1+\eta)|\widehat{b}_j - b_j| \leqslant \lambda_{n,j}$, for any $j \in \{1, 2, ..., m\}$ and some small constant $\eta > 0$;*

*(2) $\lambda_{n,j} \leqslant C\sqrt{\log(m)/n}$ for some constant $C > 0$, $j = 1, 2, ..., m$.*

In the Assumption C.3, we require that the choice of relaxation parameter $\lambda_{n,j}$ should dominate the maximal discrepancy between $\widehat{b}_j$ and $b_j$ for all $j \in \{1, 2, ..., m\}$. In additional, $\max_{1 \leqslant j \leqslant m} \lambda_{n,j}$ should be converging to 0 as sample size increases to guarantee consistency. Below we discuss how to pick the relaxation parameter $\Lambda_n$ in practice.

Denote $\widehat{\sigma}_j^2 := \mathbb{E}_n[(b_j^i)^2] - [\mathbb{E}_n b_j^i]^2$ for $j = 1, 2, ..., m$. Define

$$r_n := \max_{1 \leqslant j \leqslant m} |\widehat{b}_j - b_j| / \widehat{\sigma}_j.$$

Ideally, for any pre-specified small $\alpha > 0$, we would like to pick $\lambda_{n,j} := r_n(1-\alpha)\widehat{\sigma}_j$, where $r_n(1-\alpha)$ is the $1-\alpha$ quantile of random variable $r_n$.

Chernozhukov, Chetverikov and Kato (2013) (CCK later) shows that the distribution of $\sqrt{n} r_n$ can be well approximated by the distribution of the maxima of a Gaussian vector when $\frac{(\log m)^7}{n} \to 0$ along with other mild regularity conditions. The calculation can be performed via Gaussian Multiplier bootstrap. A weaker bound (but still relatively sharp in many cases) of the $(1-\alpha)$ quantile of $r_n$ could be obtained using modest deviation theory of self-normalized vectors described in De La Puna (2009), which in theory requires $\frac{(\log m)^{(2+\delta)}}{n} \to 0$ where $\delta > 0$ is a constant.

Assumption: C.4 (Regularity Conditions). *(1) The data $b^i$ is i.i.d..*[5] *(2) There exists an absolute constant $C > 0$ such that*

$$\max_{1 \leqslant i \leqslant n, 1 \leqslant j \leqslant m} |b_j^i| \leqslant C.$$

*(3) There exists absolute positive constant $c_1$ such that*

$$\min_{1 \leqslant j \leqslant m} \mathbb{E}[(b_j^i)^2] \geqslant c_1.$$

Given Assumption C.4, we are able to obtain two practical ways of chosing $\Lambda_n$ whose validity is supported by results in CCK (2013) and De La Puna (2009).

Lemma 4 (Choosing $\Lambda$ using Multiplier Bootstrap (Theorem 3.1 of CCK(2013))). *Let $r_n^G := \max_{1 \leqslant j \leqslant m} \frac{\sum_{1 \leqslant i \leqslant n} b_j^i e_{ij}}{n}$, where $e_{ij}$ are independent standard normal random variables. Suppose Assumption C.3 holds and $\frac{\log(m \vee n)^7}{n} \to 0$, then the $1-\alpha$ quantile of $\sqrt{n} r_n^G$ is a consistent estimator of the $1-\alpha$ quantile of $\sqrt{n} r_n$. The $\lambda_{n,j} := \widehat{\sigma}_j r_n$.*

Lemma 5 (Choosing $\Lambda$ using Modest Deviation Theory of Self-Normalized Vectors). *Denote $\widehat{\sigma}_j^2 := \mathbb{E}_n[(b_j^i)^2] - [\mathbb{E}_n b_j^i]^2$ for $j = 1, 2, ..., m$. Let $\lambda_{n,m} := \frac{C\widehat{\sigma}_j \Phi^{-1}(1-\frac{\alpha}{2m})}{\sqrt{n}} \precsim \sqrt{\log(m)/n}$ for some constant $C > 1$. Suppose Assumption C.4 holds and $\frac{(\log m)^{2+\delta}}{n} \to 0$ for some $\delta > 0$, then as $n \to \infty$, with probability at least $1-\alpha$, for all $j = 1, 2, ..., m$,*

$$|\widehat{b}_j - b_j| \leqslant \lambda_{n,j}.$$

For any $m \times m$ matrix $M$ and $m \times 1$ vector $b$, and $Q := \{v | Mv \leqslant b\}$, define $M \oplus \Lambda := \{v | Mv \leqslant b + \Lambda\}$ for any $m \times 1$ vector $\Lambda$.

[5] The i.i.d. assumption can be extended to the i.n.i.d. assumption as Lemma 5 and Lemma 6 both allow i.n.i.d data with small modifications in the statement.

Lemma 6. *If Assumption C.3-C.4 hold. Assume that $\frac{s_0^2 \log(m)}{n} \to 0$, then with probability at least $1-\alpha$, the solution to the problem $\widehat{\mathcal{R}}_{L_1}$, $\widehat{\Pi}_{L_1}$ must exist. In addition, the corresponding $\widehat{Q}_{\widehat{\mathcal{I}}}$ satisfies:*

*(1) $Q \subset \widehat{Q}_{\widehat{\mathcal{I}}} \oplus \Lambda_n$.*

*(2) $\widehat{Q}_{\widehat{\mathcal{I}}} \subset Q \oplus (K_\infty \max_{1 \leqslant j \leqslant m} \lambda_{n,j} + 2\Lambda_n)$,*

Results of Lemma 6 does not require any sparsity assumptions. With the sparsity assumption C.2, we prove the following main theorem.

Theorem 2 (Recovery of Informative Inequalities under Exact Sparse Assumption). *Suppose Assumptions C.2-C.4 hold. Recall that $c_j$ is the maximal separation of the $j^{th}$ inequality and $c_{g,n} \leqslant c_j$ for all $j \in T_0$. Let $0 < \eta < 1$ be an absolute constant. Assume that $n, s_0$ and $c_{g,n}$ obeys the following condition:*

$$\frac{K_L s_0 \max_{1 \leqslant j \leqslant m} \lambda_{n,j}}{c_{g,n}} \to 0.$$

*Then, with probability $\geqslant 1-\alpha$, the set $\widehat{\mathcal{I}}_\eta := \{j | g(\widehat{\Pi}_{*j}) \geqslant \eta\}$ has the following properties:*

*(1) There exists an absolute constant $C_T$ such that $||\widehat{\mathcal{I}}_\eta||_0 \leqslant \frac{C_T s_0}{\eta}$;*

*(2) $\widehat{\mathcal{I}}_\eta \supset \mathcal{I}_0$;*

*(3) $\widehat{Q}_{\widehat{\mathcal{I}}_\eta} \subset Q \oplus \Lambda_n$;*

*(4) $Q \subset \widehat{Q}_{\widehat{\mathcal{I}}_\eta} \oplus \Lambda_n$.*

When the data is i.i.d., under weak conditions such as those in Lemma 5, $\max_{1 \leqslant j \leqslant m} \lambda_{n,j} \precsim_p \sqrt{\log(m) n}$. So the growth condition in Theorem 2 can be rewritten as $\frac{K_L^2 s_0^2 \log(m)}{n c_{g,n}} \to 0$.

## 5. Properties of the Selection procedure $\widehat{\mathcal{R}}$ with Application in the Core Determining Class Problem.

5.1. *Application in Core Determining Class problem.* To find the Core Determining Class given a bipartite graph $G = (\mathcal{U}, \mathcal{Y}, \varphi)$, we can use the method proposed in Section 3 to eliminate all the redundant inequalities and find exact solution when data noise is not taken into consideration. We can also use the $L_1$ selector proposed in Section 5.1 to find an approximate solution to the Core Determining Class problem. In addition, we may consider a hybrid method: first, we find the exact solution according to the method described in Section 3, and then apply the selection procedure presented in Section 5.1 using the inequalities selected from the previous step to further reduce the number of inequalities. The hybrid method may speed up the selection procedure significantly. In this subsection, we discuss the general selection procedure first, and then briefly discuss the hybrid method.

In the Core Determining Class problem, the equality $v(\mathcal{U})=1$ is never redundant when the graph $G$ is non-degenerate. Therefore, we let the $(m-1)^{th}$ and $m^{th}$ inequalities be $v(\mathcal{U})\geqslant 1$ and $v(\mathcal{U})\leqslant 1$ among the total $m$ inequalities. Since there is no reason to drop the last two inequalities, we define problems $\mathcal{R}^C$ and $\widehat{\mathcal{R}}^C$ as special modifications of the procedures introduced in the previous subsection:

Problem $\mathcal{R}^C$ :

$$\min_{\Pi}\sum_{k=1}^{m-2}\max_{1\leqslant j\leqslant m-2}\Pi_{jk},$$

subject to:

$$(1)\Pi M\geqslant M,\Pi\geqslant 0,$$

$$(2)\Pi b\leqslant b,$$

and Problem $\widehat{\mathcal{R}}^C$ :

$$\min_{\Pi}\sum_{k=1}^{m-2}\max_{1\leqslant j\leqslant m}\Pi_{jk},$$

subject to:

$$(1)\Pi M\geqslant M,\Pi\geqslant 0,$$

$$(2)\Pi(\widehat{b}-\Lambda_n)+2diag(\Pi)\Lambda_n\leqslant\widehat{b}+\Lambda_n,$$

where $\Lambda_n:=(\lambda_{n,1},...,\lambda_{n,m-2},0,0)$ with $\lambda_{n,m}$ left to be chosen according to procedures proposed by Lemmas 4 or 5.

Let $\Pi_0$ be the solution to $\mathcal{R}^C$ and $\widehat{\Pi}$ be the solution to $\widehat{\mathcal{R}}^C$. First, we prove a result specific to the Core Determining Class.

Lemma 7 (Perfect Recovery of the Minimum Model $T_0$). *If $\widehat{\mu}_n$ is non-degenerate and $\lambda_{n,j}=0$, $j=1,2,...,m-2$, then:*

*(1) The $L^0$ norm of $\widehat{g}$, $||\widehat{g}||_0$, satisfies $||\widehat{g}||_0=s_0$;*

*(2) $\max_{1\leqslant j\leqslant m-2}||\widehat{\Pi}_{j*}||_1\leqslant d_u$;*

*(3) $\max_{1\leqslant j\leqslant m-2}\widehat{g}(\widehat{\Pi}_{*j})\leqslant 1$;*

*(4) The set of non-zero entries of $\widehat{g}:=(g_\infty(\widehat{\Pi}_{*1}),...,g_\infty(\widehat{\Pi}_{*(m-2)}))$ satisfies:*

$$\widehat{\mathcal{I}}:=\{k|\widehat{g}_k\neq 0\}=\mathcal{I}_0.$$

For the Core-Determining Class problem, we can combine the methods proposed in Section 3 and Section 4: we can use the method in Section 3 to select the binding inequalities first, then followed by the method proposed in Section 4 to further reduce the number of inequalities. We can such a procedure as the "hybrid method".

(1) When $s_0$ is small, the hybrid method performs similarly to the combinatorial method only.

(2) When $s_0$ is large, there may be significant gains from the hybrid method in terms of computational speed compared to the selection procedure only, and significant inequality reduction compared to the combinatorial method only.

We illustrate these points in the Monte-Carlo experiments in the next section.

**6. Monte-Carlo Experiments.** Consider a simple setting in which many marginal firms are facing a volatile market. Let $u$ be a random variable representing the cost of a firm. Let $\theta \in \{H, L\}$ be the private information of the firm which we do not observe. Let $y$ be the action of the firm based on the information $\theta$ and cost $u$. Assuming the objective of firm is to maximize profit $\pi(y, u, \theta)$, they might adopt different actions when facing $\theta = H$ or $\theta = L$.

Suppose action $y$ is the price set by the firm. We consider a simple case of decision making problem by the firms. Given observations of a sequence of decisions, we are interested in learning the distribution of the costs of these firms.

Assume that the profit function is

$$\pi(y, u, H) = (y - u)(C - y),$$

$$\pi(y, u, L) = (y - u)(C/2 - y),$$

where $C$ is a constant.

If the firm consider any price $y^* \in \{y | \pi(y, u, \theta) \geqslant \max_y \pi(y, u, \theta) - w, a_1 \leqslant y \leqslant a_2\}$, where $w$ is a constant for robust price control and $a_1, a_2$ are bounds on $y$, then $\varphi(u) := \{y | \pi(y, u, \theta) \geqslant \max_y \pi(y, u, \theta) - w, \theta \in \{H, L\}, a_1 \leqslant y \leqslant a_2\}$ is the correspondence mapping from the set of cost (event) $\mathcal{U}$ to the set of price (outcome) $\mathcal{Y}$.

We can only observe $\widehat{y}$, the empirical measures on price $\mathcal{Y}$. The objective is to find an approximate feasible set of probability measure on cost $\mathcal{U}$. Assume that $u$ is i.i.d. across observations.

Example 3 (Monte-Carlo Experiment 1). *Set $\mathcal{U}$, $\mathcal{Y}$, $C$ and $w$ as follows:*

*$C = 4$, $\mathcal{U} = [0, 3]$, $\mathcal{Y} = [1, 3.5]$, $w = 0.01$.*

*So $\varphi(u) = [(1.9 + u/2), (2.1 + u/2) \wedge 3.5] \cup [(0.9 + u/2 \vee 1), (1.1 + u/2)]$.*

*To estimate the probability measure on cost $\mathcal{U}$, we discretize the continuous set of cost (event) $\mathcal{U}$ and price (outcome) $\mathcal{Y}$. Let $d_u = 15$ and $d_y = 25$ be the number of discretized segments of cost and price, respectively. Then $u_i = ((i-1)/5, i/5)$ and $y_j = ((j-1)/10+1, j/10+1)$ for $i = 1, 2, ..., d_u$ and $j = 1, 2, ..., d_2$. The correspondence mapping $\varphi_d$ from the discretized set $\mathcal{U}_d = \{u_i | i = 1, 2, ..., 15\}$ to the discretized set $\mathcal{Y}_d = \{y_j | j = 1, 2, ..., 25\}$ is generated by:*

$$\varphi_d(u_i) = \{y_j | y_j \cap \varphi(u_i) \neq \emptyset\}$$

*Therefore, $\varphi(u_1) = \{y_1, y_2, y_{10}, y_{11}, y_{12}\}$, $\varphi(u_{15}) = \{y_{14}, y_{15}, y_{16}, y_{24}, y_{25}\}$. For any $2 \leqslant i \leqslant 14$, $\varphi(u_i) = \{y_{i-1}, y_i, y_{i+1}, y_{i+9}, y_{i+10}, y_{i+11}\}$. Figure 3 illustrates the correspondence mapping for Example 3.*

Suppose $\mu$, the true probability measure on $\mathcal{Y}$, follows the formula $\mu(j) \propto \max(1, |j - 13|^{1.5})$ for $1 \leqslant j \leqslant 25$. Suppose the sample size $n$ (the number of observed $y$) is 2000 and 500 and the sample $y$ is randomly drawn according to measure $\mu$. Let $\widehat{\mu}_n$ be the empirical measure (observed frequency) of $y$. The penality $\Lambda_n$ is chosen according to Lemma 4.

Problem $\widehat{\mathcal{R}}$ is implemented to further select the inequalities. The results of a set of Monte-Carlo experiments with 100 repetitions are presented in Table 1. For each instance, we apply a cut-off value $\eta$ to the optimal $L^1$ coefficient $g(\Pi_{*j})$: select an inequality if the corresponding $g(\Pi_{*j}) \geqslant \eta$ and discard it otherwise. We present the average, maximum and minimum number of selected inequalities with cut-off value $\eta = 0, 0.1$ and $0.2$.

A critical concept concerning the selection performance is “coverage”: in one instance, the feasible set (of the probability measure) on $\mathcal{U}$ corresponding to the true $\mu$ is subset of the feasible set (of the probability measure) on $\mathcal{U}$ defined by the selected inequalities with empirical measure $\widehat{\mu}_n$. We present the “frequency of coverage” corresponding to different cut-off value $\eta$. As $\eta$ increases, the procedure selects fewer inequalities, so the approximate feasible set will become larger, which is more likely to “cover” the true feasible set and produce a larger “frequency of coverage”. In the numerical experiments, the “frequency of coverage” is greater than 95% when the cut-off value $\eta = 0$ (essentially the case of no cut-off). It agrees with the parameter selection $\alpha = 0.05$ (type 1 error) in the formula of the penalty term $\lambda$ described in Lemma 6.

We compare the inequalities selection of the integer programming $L^0$ procedure with the linear programming $L^1$ procedure. Figure 4 illustrates the comparisons with respect to the magnitude of the $L^1$ selector coefficient

| | | |
|---|---|---|
| Number of experiments ($M$) | 100 | |
| Number of events × number of outcomes ($d_1 \times d_2$) | $15 \times 25$ | |
| Number of inequalities in true model ($m$, after the selection algorithm in Section 3) | 471 | |
| Conservative bound of acceptance rate ($1-\alpha$) | 0.95 | |
| Sample size ($n$) | 500 | 2000 |
| Average $\lambda$ | 0.0710 | 0.0355 |
| Frequency of Coverage ($\eta = 0$) | 97% | 99% |
| Avg. number of inequalities selected ($\eta = 0$) | 184.66 | 187.42 |
| Max. number of inequalities selected ($\eta = 0$) | 241 | 234 |
| Min. number of inequalities selected ($\eta = 0$) | 145 | 92 |
| Frequency of Coverage ($\eta = 0.1$) | 99% | 100% |
| Avg. number of inequalities selected ($\eta = 0.1$) | 32.59 | 86.02 |
| Max. number of inequalities selected ($\eta = 0.1$) | 43 | 145 |
| Min. number of inequalities selected ($\eta = 0.1$) | 27 | 27 |
| Frequency of Coverage ($\eta = 0.2$) | 99% | 100% |
| Avg. number of inequalities selected ($\eta = 0.2$) | 26.73 | 56.69 |
| Max. number of inequalities selected ($\eta = 0.2$) | 28 | 108 |
| Min. number of inequalities selected ($\eta = 0.2$) | 24 | 27 |
| Running time (sec/instance) | 87 | 146 |

Table 1

*Results of Monte-Carlo Experiments on Example 3*

$g(\Pi_{*j})$ in the optimal solution of problem $\widehat{\mathcal{R}}$. Figure 5 illustrates the comparisons with respect to the separation of each inequality, which is

$$c(A) := \max\{v(A) - \mu(\varphi(A)) | v(A') \leqslant \mu(\varphi(A')), \forall A' \subset \mathcal{S}_u^*, A' \neq A\}$$

Table 2 presents the detailed selection results. It can be seen that the $L^0$ selector (the model to select minimum number of inequalities) is recovered by the $L^1$ selector to a large extent, while the $L^1$ selector enjoys extremely high computational advantage. Generally, inequalities selected by the $L^0$ selector have comparatively large $L^1$ coefficients $g(\Pi_{*j})$, which makes it easy to be selected by the $L^1$ selector under a reasonable cut-off value $\eta$. In addition, the $L^1$ selector is able to successfully differentiate inequalities with close separation values but opposite $L^0$ coefficients.

We project our $L^1$ estimator compared to $L^0$ and the true feasible set onto $v_1, v_2, v_3$, a three-dimension subspace. Figure 6 compares the performance of $L^0$ and $L^1$ selectors: the $L^1$ selector with $\eta = 0.1$ is slightly more conservative than the $L^0$ selector. Figure 7 shows that the $L^1$ selector covers the true feasible set by a small margin.

We also demonstrate through a smaller example the sharpness of the relaxing parameter $\lambda$. If (1) the empirical measure $\widehat{\mu}_n$ is largely mis-specified from the true measure $\mu$, and (2) the true feasible set (of the probability

| | |
|---|---|
| Number of inequalities selected in $L^0$ | 79 |
| Number of inequalities selected in $L^1$ | 211 |
| Number of inequalities that $L^0$ model selected in $L^1$, $\eta = 0$ | 79 |
| Number of inequalities that $L^0$ model selected in $L^1$, $\eta = 0.05$ | 78 |
| Number of inequalities that $L^0$ model selected in $L^1$, $\eta = 0.10$ | 78 |
| Number of inequalities that $L^0$ model selected in $L^1$, $\eta = 0.15$ | 77 |
| Number of inequalities that $L^0$ model selected in $L^1$, $\eta = 0.20$ | 72 |
| Running time of $L^0$ model (min) | 2195 |
| Running time of $L^1$ model (min) | 1.45 |

TABLE 2
*Comparisons of $L^0$ and $L^1$*

| $M = 10000, n = 200$ | $\gamma = 1$ | $\gamma = 2$ | $\gamma = 3$ | $\gamma = 4$ | $\gamma = 5$ |
|---|---|---|---|---|---|
| Type 1 error, $\eta = 0$ | 3.64% | 3.73% | 3.84% | 3.85% | 4.21% |
| Type 1 error, $\eta = 0.1$ | 3.56% | 3.55% | 3.71% | 3.77% | 4.07% |
| Type 2 error, $\eta = 0$ | 29.65% | 11.18% | 6.98% | 5.86% | 4.58% |
| Type 2 error, $\eta = 0.1$ | 29.88% | 11.35% | 7.08% | 6.04% | 4.73% |

TABLE 3
*Type 1 and Type 2 Errors*

measure) on $\mathcal{U}$ is still a subset of the approximate feasible set (of the probability measure) on $\mathcal{U}$ obtained in the selection procedure, then a type 2 error occurs. The $\lambda$ implied by $\alpha$ limits the magnitude of the type 1 error, and a sharp $\lambda$ will also limit the occurrences of type 2 errors at the same time. In another set of Monte-Carlo experiments below, we examine the type 2 error in the case that the empirical measure $\widehat{\mu}_n$ is locally mis-specified.

EXAMPLE 4 (Monte-Carlo Experiment 2). *Figure 8 is the correspondence mapping for an example with size of* $7 \times 7$*.*

*Assuming the true probability measure $\mu$ on $\mathcal{Y}$ is* $(0.1, 0.25, 0.2, 0.1, 0.1, 0.2, 0.05)$*, we perturb $\mu_{n,0}$ with $\gamma(a_1, a_2, ..., a_7)/\sqrt{n}$, where $a_i$ is randomly and uniformly drawn from $\{-1, 1\}$, $1 \leqslant i \leqslant 7$. So the empirical measure $\widehat{\mu} \propto \mu + \gamma(a_1, a_2, ..., a_7)/\sqrt{n}$ in the case of mis-specified perturbation.*

We run 10000 instances for each setting of perturbation $\gamma$ and cut-off value $\eta$. Table 3 presents the type 1 and type 2 error for each setting. The results show that, while the type 1 error is less than 0.05 as designed, the type 2 error is also relatively small, which means the approximate feasible set of probability measure on $\mathcal{U}$ does not over exaggerate the true feasible set.

**7. Conclusion.** In this paper we consider estimating the probability measure on the unobservable events given observations on the outcomes.

We try to select the set of minimum number of inequalities, which is called the Core Determining Class, to describe the feasible set of target probability measure. We propose a procedure to construct the exact Core Determining Class when data noise are not taken into consideration. We prove that, if there is no degeneracy, the Core Determining Class only depends on the structure of the bipartite Graph, not the probability measure $\mu$ on the outcomes.

For a general problem of linear inequalities selection under noise, we propose a selection procedure similar to the Dantzig selector. A formulation is proposed to identify the importance of each inequality in a feasible set defined by many inequalities constraints. We describe the exact sparse assumptions and approximate sparse assumptions, which are are similar to the traditional sparse assumptions in a linear regression environment. We prove that the selection procedure has good statistical properties under the sparse assumptions.

We apply the selection procedure to the Core Determining Class problem and develop a hybrid selection method combined with a combinatorial algorithm. We prove that the hybrid selection procedure has better statistical properties due to the structure of the graph.

We demonstrate the good performance of our selection procedure through several set of Monte-Carlo experiments. First, the inference based on the selection procedure has desired size; second, it has power against local alternatives; third, it is relatively computationally efficient.

**8. Reference.**

Ackerberg, D., Lanier Benkard, C., Berry, S., Pakes, A. (2007). Econometric tools for analyzing market outcomes. Handbook of econometrics, 6, 4171-4276.

Andrews, D. W., Shi, X. (2013). Inference based on conditional moment inequalities. Econometrica, 81(2), 609-666.

Andrews, D. W., Soares, G. (2010). Inference for parameters defined by moment inequalities using generalized moment selection. Econometrica, 78(1), 119-157.

Artstein, Z. (1983). Distributions of random sets and random selections. Israel Journal of Mathematics, 46(4), 313-324.

Bajari, P., Benkard, C. L., Levin, J. (2007). Estimating dynamic models of imperfect competition. Econometrica, 75(5), 1331-1370.

Bajari, P., Hong, H., Ryan, S. P. (2010). Identification and estimation of a discrete game of complete information. Econometrica, 78(5), 1529-1568.

Belloni, A., Freund, R. M. (2008). On the symmetry function of a convex set. Mathematical Programming, 111(1-2), 57-93.

Beresteanu, A., Molchanov, I., Molinari, F. (2011). Sharp identification regions in models with convex moment predictions. Econometrica, 79(6), 1785-1821.

Candes, E., Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. The Annals of Statistics, 2313-2351.

Caron, R. J., McDonald, J. F., Ponic, C. M. (1989). A degenerate extreme point strategy for the classification of linear constraints as redundant or necessary. Journal of Optimization Theory and Applications, 62(2), 225-237.

Chernozhukov, V., D., Chetverikov and K., Kato(2014). Central Limit Theory and multiplier bootstrap when $p$ is much larger than $n$. The Annals of Statistics.

Chesher, A., Rosen, A. (2012). Simultaneous equations models for discrete outcomes: coherence, completeness, and identification, cemmap working paper CWP21/12.

Eaves, B. C., Freund, R. M. (1982). Optimal scaling of balls and polyhedra. Mathematical Programming, 23(1), 138-147.

Galichon, A., Henry, M. (2006). Inference in incomplete models. Available at SSRN 886907.

Galichon, A., Henry, M. (2011). Set identification in models with multiple equilibria. The Review of Economic Studies, 78(4), 1264-1298.

Jovanovic, B. (1989). Observable implications of models with multiple equilibria. Econometrica: Journal of the Econometric Society, 1431-1437.

Manski, C. F., Tamer, E. (2002). Inference on regressions with interval data on a regressor or outcome. Econometrica, 70(2), 519-546.

Paulraj, S., Chellappan, C., Natesan, T. R. (2006). A heuristic approach for identification of redundant constraints in linear programming models.International Journal of Computer Mathematics, 83(8-9), 675-683.

Puna, V., T, Lai and Qi, Shao(2009). Self-Normalized Processes: Limit Theory and Statistical Applications. Springer.

Romano, J. P., Shaikh, A. M. (2010). Inference for the identified set in partially identified econometric models. Econometrica, 78(1), 169-211.

Telgen, J. (1983). Identifying redundant constraints and implicit equalities in systems of linear constraints. Management Science, 29(10), 1209-1222.

## 9. Appendix A: Proofs in Section 3.

**Proof of Lemma 2**. For any $A \notin \mathcal{S}_u^{'}$, one of the following two statements must be true:

(1) $\exists A_1, A_2 \subset A, A_1, A_2 \neq \emptyset, A_1 \cup A_2 = A$, such that $\varphi(A_1) \cap \varphi(A_2) = \emptyset$;

(2)$\exists u \in \mathcal{U}$, such that $u \notin A$, and $\varphi(u) \subset \varphi(A)$.

[What is VM?] If (1) is true, then $v^M(A) = v^M(A_1 \cup A_2) = v(A_1) + v(A_2) \leqslant \mu(\varphi(A_1)) + \mu(\varphi(A_2)) = \mu(\varphi(A_1) \cup \varphi(A_2)) = \mu(\varphi(A_1 \cup A_2)) = \mu(\varphi(A))$, so $A \notin \mathcal{S}_u$.

If (2) is true, $v^M(A) \leqslant v(A \cup \{u\}) \leqslant \mu(\varphi(A \cup \{u\})) = \mu(\varphi(A))$, so $A \notin \mathcal{S}_u$.

Therefore, by definition 3.2,

$$\mathcal{S}_u \subset \mathcal{S}_u'. \tag{9.1}$$

Consider an arbitrary set $A \subset \mathcal{U}$ such that $A \notin S_u$. It is suffice to prove that $A \notin S_u'$ to show $\mathcal{S}_u' \subset \mathcal{S}_u$.

Denote $\mathcal{S}_u := \{A_i | 1 \leqslant i \leqslant r := |\mathcal{S}_u|\}$. For every set $A \subset \mathcal{U}$, we can consider a vector $V := w(A) \in \{0,1\}^{d_u}$ such that $V_i = 1$ if and only if $u_i \in A$. By definition, there exists a $1 \times d_u$ vector $\pi \geqslant 0$, such that (1) $\sum_{i=1}^r \pi_i w(A)_i \geqslant w(A)$, (2) $\sum_{i=1}^r \pi_i \mu(\varphi(A_i)) \leqslant \mu(\varphi(A))$, where $r := |S_u|$. Without loss of generality, assume $\pi_i > 0$, $i = 1, 2, ..., r$, otherwise we would simply omit the $A_i$ which corresponds to $\pi_i = 0$ in the statements above. Such an assumption does not affect our analysis below.

Since $\sum \pi_i w(A)_i \geqslant w(A)$, so

$\sum_{i=1}^r \pi_i 1(A_i \cap \varphi^{-1}(y) \neq \emptyset) \geqslant 1(A \cap \varphi^{-1}(y) \neq \emptyset)$, for any $y \in \mathcal{Y}$.

By Galichon and Henry (2011), $\mu \circ \varphi$ is sub-modular on $2^{\mathcal{U}}$. Therefore,

$\sum \pi_i \mu(\varphi(A_i)) = \sum_{y \in \mathcal{Y}} \sum_{i=1}^r \pi_i \mu(y) 1(A_i \cap \varphi^{-1}(y) \neq \emptyset) \geqslant \sum_{y \in \mathcal{Y}} \mu(y) 1(A \cap \varphi^{-1}(y)) = \mu(\varphi(A))$.

But we know that $\sum \pi_i \mu(\varphi(A_i)) \leqslant \mu(\varphi(A))$, by assumption. Hence the inequality above holds as an equality, i.e., for any $y \in \mathcal{Y}$, $\sum_{i=1}^r \pi_i 1(A_i \cap \varphi^{-1}(y) \neq) = 1(A \cap \varphi^{-1}(y))$. Also, the inequality $\sum_{i=1}^r \pi_i w(A)_i \geqslant w(A)$ must hold as an equality, i.e.,

$$\sum_{i=1}^r \pi_i w(A)_i = w(A). \tag{9.2}$$

Given the above results, we claim that for any $y \in \mathcal{Y}$, either $\varphi^{-1}(y) \cap A \subset A_i$ or $\varphi^{-1}(y) \cap A \cap A_i = \emptyset$ for all $i$. We prove this argument by contradiction.

Assuming that there exists a $y \in \mathcal{Y}$ and $1 \leqslant i \leqslant r$ such that $\varphi^{-1}(y) \cap A \cap A_i \neq$, and $\varphi^{-1}(y) \cap A \subsetneq A_i$. Therefore, there exists $u \neq u'$ such that $u, u' \in \varphi^{-1}(y)$, $u \in A \cap A_i'$, $u' \in A$ but $u' \notin A_i$. Thus,

$\sum_{i=1}^{r} \pi_i A_i 1(A_i \cap \varphi^{-1}(y) \neq) = \pi_i + \sum_{j\neq i} \pi_j A_j 1(A_j \cap \varphi^{-1}(y) \geqslant \pi_i + \sum_{j\neq i} \pi_j 1(u' \in A_j) = \pi_i + \sum_{j=1}^{r} \pi_j 1(u' \in A_j) \geqslant \pi_i + 1 > 1 = 1(A \cap \varphi^{-1}(y))$, contradiction!

Thus, for any $y \in \mathcal{Y}$,

$$\text{either } \varphi^{-1}(y) \cap A \subset A_i \text{ or } \varphi^{-1}(y) \cap A \cap A_i = \emptyset, \tag{9.3}$$

for all $i$.

We divide our next step into two cases:

Case 1: If $A$ is self-connected, by definition of $A_i$, for any $i$, either $A_i \cap A = \emptyset$ or $A_i \cap A = A$.

By the equation (9.2), there exists no $y \in \varphi(A_i)$ such that $y \notin \varphi(A)$ for any $i$ with $\pi_i > 0$. By connectivity of $A$, $\varphi(A_i) = \varphi(A)$ for any $i$ such that $\pi_i > 0$.

Since $A_i \neq A$ for any $i = 1, 2..., r$, but there exists $i$ such that $\varphi(A_i) = \varphi(A)$. Since for any $u \in \mathcal{U}$, $\varphi(u) \subset \varphi(A_i)$ implies $u \in A_i$, there must exist $u_0 \in \mathcal{U}$ such that $\varphi(u_0) \subset \varphi(A)$, but $u_0 \notin A$, i.e., $A \notin \mathcal{S}'_u$.

Case 2: If $A$ is not self-connected, by definition $A \notin \mathcal{S}'_u$.

Therefore, in both cases, $A \notin \mathcal{S}'_u$, which implies that $\mathcal{S}_u \supset \mathcal{S}'_u$. Combining with (9.1), $\mathcal{S}_u = \mathcal{S}'_u$.

□

**Proof of Theorem 1**. Without loss of generality, we can assume that the graph $G$ is connected, because otherwise the problem can be decomposed to each connected branch of $G$.

By definition, $\mathcal{S}^*$ is the set of inequalities that are binding together with $v(\mathcal{U}) = 1$. Therefore, it is easy to see that $\mathcal{S}^* \subset \mathcal{S}_u$, $\mathcal{S}^* \subset \mathcal{S}_y^{-1}$, and $\mathcal{S}^* \subset \mathcal{S}_u \cap \mathcal{S}_y^{-1}$.

Denote $w(A) \in \{0,1\}^{d_u}$ as the indicator vector of $A$.

Suppose there exists a set $A$ such that $A \in \mathcal{S}_u \cap \mathcal{S}_y^{-1}$, but $A \notin \mathcal{S}^*$. By definition of $\mathcal{S}^*$, there exists $\pi_i \geqslant 0$ and $A_i \in \mathcal{S}^*$, $1 \leqslant i \leqslant r$, with $r = |\mathcal{S}^*|$, and $\pi_0 \geqslant 0$, such that:

(1)$\sum_{1\leqslant i} \pi_i w(A_i) - \pi_0 \geqslant w(A)$.

(2)$\sum_{1\leqslant i} \pi_i \mu(\varphi(A_i)) - \pi_0 \geqslant \mu(\varphi(A))$.

Without loss of generality, we can assume that $\pi_i > 0$ for all $i = 1, 2, ..., r$.

By the similar argument of Lemma 2, all inequalities in (1) and (2) must hold as equalities. Again, for any $y \in \mathcal{Y}$, either $\varphi^{-1}(y) \cap A$ is a subset of $A_i$, or its intersection with $A_i$ is empty. Since $A \in \mathcal{S}_u$ is connected, so for any $A_i$,

$$\text{either } A_i \supset A \text{ or } A_i \cap A = \emptyset. \tag{9.4}$$

Since $A \in \mathcal{S}_u \cap \mathcal{S}_y^{-1}$ ,there exists $B$ such that $\varphi^{-1}(B) = A^c$. Then, it is easy to see that $\varphi^{-1}(\varphi(A)^c) = A^c$. Denote $B := \varphi(A)^c$. Since the graph $G$ is connected, so it must be that $\varphi(u) \cap \varphi(A) \neq \emptyset$, for some $u \in A^c$.

If $\pi_0 = 0$, then by similar arguments of Lemma 2, it is easy to see that $A$ equals to some $A_i$ with $A_i \in \mathcal{S}^*$, contradiction!

Else we assume that $\pi_0 > 0$, then there must exist a set $A_{i_0}$ such that $u \in A_{i_0}$. Since $\varphi(u) \cap \varphi(A) \neq \emptyset$, by statement (9.4), $A_{i_0} \supset A$.

By the same argument, for any $y \in B$, either $\varphi^{-1}(b) \subset A_i$ or $\varphi^{-1}(b) \cap A_i = \emptyset$, for any $i$. By definition, the set $B$ is self-connected. Therefore for any $A_i$, $B \subset A_i$ or $B \cap A_i = \emptyset$. This statement also holds for $A_{i_0}$. Since $A_{i_0} \supset A$ and $A_{i_0} \neq A$, then $A_{i_0} \cap B \neq$. So $A_{i_0} \supset B$. Consequently, $A_{i_0} \supset B \cup A = \mathcal{U}$, which contradicts to the definition of $S^*$.

Therefore, $\mathcal{S}^* = \mathcal{S}_u \cap \mathcal{S}_y^{-1}$. □

## 10. Appendix B: Proofs in Section 5.

**Proof of Lemma 6**. It is obvious that $Q \subset Q_{\widehat{\mathcal{I}}} := \{v | M_{\widehat{\mathcal{I}}} v \leqslant b_{\widehat{\mathcal{I}}}, v \geqslant 0\}$.

By definition of $\widehat{\mathcal{I}}$ and $\widehat{\Pi}^{L_1}$, we know that for any $j \in \widehat{\mathcal{I}}$, $M_j v \leqslant \widehat{b}_j \leqslant b_j + \lambda_{n,j}$, for all $j$ with probability $\geqslant 1 - \alpha$. Therefore, $Q_{\widehat{\mathcal{I}}}$ must be a subset of $\widehat{Q}_{\widehat{\mathcal{I}}} \oplus \Lambda_{n\widehat{\mathcal{I}}}$, so $Q \subset \widehat{Q}_{\widehat{\mathcal{I}}} \oplus \Lambda_{n\widehat{\mathcal{I}}}$.

For any $v \in \widehat{Q}_{\widehat{\mathcal{I}}}$, we know that $M_{\widehat{\mathcal{I}}} v \leqslant \widehat{b}_{\widehat{\mathcal{I}}}$.

For any $j \notin \widehat{\mathcal{I}}$, $\Pi_{jj} = 0$. by definition, $\widehat{\Pi}^{L_1}_{j*} M \geqslant M_j$, and $\widehat{\Pi}_{j*}(\widehat{b} - \Lambda_n) \leqslant \widehat{b}_j + \lambda_{n,j}$. So $M_j v - b_j \leqslant \Pi_{*j} M v - b_j \leqslant \Pi^{*j} \widehat{b} - b_j \leqslant \widehat{\Pi}_{j*} \Lambda_n + 2\lambda_{n,j}$.

By optimality of $\widehat{\Pi}^{L_1}$, since we know that $\widetilde{\Pi}$ is a solution to $\widehat{\mathcal{R}}_{L_1}$, so $\sum_{j \in \widehat{\mathcal{I}}} g_\infty(\widehat{\Pi}^{L_1}_{*j}) \leqslant \sum_{j \in \mathcal{I}} g_\infty(\widetilde{\Pi}_{*j}) := K_\infty$. It follows that $\widehat{\Pi}_{j*} \Lambda_n \leqslant K_\infty \max_{1 \leqslant j \leqslant m} \lambda_{n,j}$.

Hence, $\widehat{Q}_{\widehat{\mathcal{I}}} \subset Q \oplus (K_\infty \max_{1 \leqslant j \leqslant m} \lambda_{n,j} + 2\Lambda_n)$. □

**Proof of Theorem 2**. First, it is obvious that $\widetilde{\Pi}$ is a feasible solution to the problem $\widehat{\mathcal{R}}^{L_1}$ with probability at least $1 - \alpha$. Now let's assume that $\widetilde{\Pi}$ is a feasible solution of $\widehat{\mathcal{R}}^{L_1}$.

Let $\widehat{\Pi}$ be the solution to the problem $\widehat{\mathcal{R}}^{L_1}$. So

$$||g(\widehat{\Pi})||_1 \leqslant ||g(\Pi^*)||_1 \leqslant K_\infty.$$

So $\widehat{\mathcal{I}}_\eta \leqslant \frac{K_\infty}{\eta}$.

On the other hand, the exact sparsity assumption C.2 implies that $K_\infty \leqslant s_0 K_L$, therefore $\widehat{\mathcal{I}}_\eta \leqslant \frac{s_0 K_L}{\eta}$..

For any $j \in \mathcal{I}_0$, let $v_j$ be the point such that the maximal separation of the $j^{th}$ inequality is reached while other inequalities hold.

Therefore, by construction of $\widehat{\mathcal{R}}^{L_1}$

$$\widehat{\Pi}(Mv_j - \widehat{b}) \geqslant Mv_j - \widehat{b}_j - (\widehat{\Pi}\Lambda_n)_j + 2\Pi_{jj}\lambda_{n,j} - \lambda_{n,j}. \tag{10.1}$$

By assumption, we have $M_j v_j \geqslant b_j + c_{g,n}$, and $M_{j'} v_j - b_{j'} \leqslant 0$ for all $j' \neq j$.

Plugging the above inequalities in (10.1), we get: $(1 - \widehat{\Pi}_{jj})(c_{g,n} - 2\lambda_{n,j}) \leqslant (\widehat{\Pi}\Lambda_n)_j \leqslant \lambda_S := \max_{1 \leqslant j \leqslant m} \lambda_{n,j} s_0 K_L$.

By the growth condition on $s_0$ and $n$, we know that $1 - \widehat{\Pi}_{jj} \to 0$ for all $j \in \mathcal{I}_0$. Hence, $j \in \widehat{\mathcal{I}}_\eta$ for $n$ large enough.

Now we know that $\mathcal{I}_0 \subset \widehat{\mathcal{I}}_\eta$, so $\widehat{Q}_{\widehat{\mathcal{I}}_\eta} \subset \widehat{Q}_{I_0} \subset Q \oplus \Lambda_{n\mathcal{I}_0}$. Similarly, $Q \subset \widehat{Q} \oplus \Lambda_n \subset \widehat{Q}_{\widehat{\mathcal{I}}_\eta} \oplus \Lambda_{n\widehat{\mathcal{I}}_\eta}$. □

**Proof of Lemma 7**. Let $\Pi$ be a feasible solution of the following problem:

$$\min_{\Pi} \sum_{k=1}^{m} \max_{1 \leqslant j \leqslant j} (\Pi_{jk}),$$

subject to:

$$(1) \Pi M \geqslant M, \Pi \geqslant 0,$$

$$(2) \Pi b \leqslant b.$$

$$\Pi_{ij} = 0, \text{ if j } \notin T_0.$$

Any feasible solution of this above problem is that $\Pi_{ii} = 1$, for all $i \in T_0$, and $\Pi_{ij} = 0$, for all $i \neq j$. Hence, the optimal value of the objective function is $s_0$.

In our case, except for the $p^{th}$ row of $M$, every row satisfies: $M_i \in \{0,1\}^d$. Again, for the problem $\mathcal{R}$, any optimal solution must satisfy $\Pi_{ii} = 1$, for any $i \in T_0$. Therefore the value of the objective function is at least $s_0$.

Meanwhile, for any $i \notin T_0$, by definition, there exists $\alpha_j \geqslant 0$, for any $j \neq i$, $j \in T_0$

and $\alpha_p \geqslant 0$ such that:

$\sum_{j \in T_0} \alpha_j M_j - \alpha_p(1, 1, ..., 1) \geqslant M_i$,

and $\sum_{j \in T_0} \alpha_j b_j - \alpha_p \leqslant b_i$.

Without loss of generality, we could assume that $\alpha_1 \geqslant \alpha_2 \geqslant ... \geqslant \alpha_r > 0 = \alpha_{r+1} = ... = \alpha_{p-1}$. Next we prove that there must be a feasible vector of $\alpha_i$ such that $\alpha_1 \leqslant 1$. Then we could conclude that the minimum value of the objective function in problem $\mathcal{R}$ is $s_0$, and the optimal solution exactly recovers the true model. Denote the set $A$ correspond to $M_j$, and $A_i$ correspond to $M_i$. Without loss of generality, assume that

By Galichon and Henry (2011), $\mu(\varphi(A))$ is a sub-modular.

$b_j = \mu(\phi(A))$, therefore $\sum_{1\leqslant i\leqslant r}\alpha_i b_j - \alpha_p = \sum_{1\leqslant i\leqslant r}\alpha_i \mu(\varphi(A_i)) - \alpha_p \mu(\varphi(\mathcal{U})) \geqslant \mu(\varphi(\sum_{1\leqslant i\leqslant r}\alpha_i A_i - \alpha_p)) \geqslant \mu(A) = b_j$. Therefore the above equality holds as an equality. If $\alpha_1 > 1$, then $\alpha_p > 0$. So for any $u \notin A_i$, there must be $j \in T_0$ such that $u \in M_j$.

So for any $y \in \varphi(A)$, either $\phi^{-1}(y) \cap A_i \cap A = \emptyset$ or $(\phi^{-1}(y) \cap A) \subset A_i$. Similarly, for any $y \notin \varphi(A)$, $\phi^{-1}(y) \cap A_i = \emptyset$ or $\phi^{-1}(y) \subset A_i$.

We continue the proof by discussing two exclusive cases:

(1) $A$ is connected. Let $A' := \{u|\varphi(u) \subset A\}$. We only need to prove that $A'$ can be constructed via $\sum_{1\leqslant i\leqslant r}\alpha_i A_i - \alpha_p \mathcal{U}$. For simplicity, we still call $A'$ as $A$. By such an assumption, there is no $u \notin A$ such that $\varphi(u) \subset \varphi(A)$. Therefore, $A \subset \mathcal{S}_u$. Hence $B := \varphi(A)^c$ is not connected. Let $B_1, ..., B_r$ as all the disconnected branches of $B$. Let $C_k = \varphi(B_k)$, for any $1 \leqslant k \leqslant r$. So $\cup_{k=1}^r C_k = A^c$, $C_{k_1} \cap C_{k_2} = \emptyset$, for any $k_1 \neq k_2$. So each $C_k$ is connected with $A$.

Denote $C^k = \{u|u \in A^c, u \notin C_k\}$.

So $A \cup C^1$, $A \cup C^2$,..., $A \cup C^r$ are sets in $\mathcal{S}_u$. It is also sets in $\mathcal{S}_y^{-1}$ since $C_k = (A \cup C^k)^c$ is connected. Therefore, All these sets are in $\mathcal{S}^*$. And Let $\alpha_i = 1$, $\alpha_p = r - 1$, we could reconstruct the inequality indicated by $A$. And since $r \geqslant 2$, so all the coefficients $\alpha_k \leqslant 1$.

(2) $A$ is not connected. Let $A_1, ..., A_w$ be the connected branches. Let $B = \varphi(A^c)$. Without loss of generality, similar to step (1), we could assume that each $A_i \in \mathcal{S}_u$, $1 \leqslant i \leqslant w$. Assume $B_1, ..., B_k$ is the connected branches of $B$. Let $C_i = \varphi^{-1}(B_i)$, $1 \leqslant i \leqslant k$. Therefore, $C_{i_1} \cap C_{i_2} = \emptyset$, for any $i_1 \neq i_2$. $C_i \cap A \neq \emptyset$, for any $i$. Therefore $C_i$, $1 \leqslant i \leqslant k$ and $A_j$, $1 \leqslant j \leqslant w$ form a bipartite-graph $G_0$. For every $A_i$, let $AC_1, ..., AC_{i_r}$ to be the connect branches of $G_0 - \{A_i\}$. Since the entire graph is connected, so $AC_i$ is connected with $A_i$, $1 \leqslant i \leqslant i_r$. Let $AC^i := \{u|u \notin AC_i\}$. So $AC^i$ is a set in $\mathcal{S}_u \cap \mathcal{S}_y^{-1} = \mathcal{S}^*$. Therefore, the set $A_i$ could be constructed by $\sum_{k=1}^{i_r} AC_k - (i_r - 1)\mathcal{U}$.

If for some set $AC_k$ appears in the different $i$, let $AC$ be such as set such that it appears in $1 \leqslant i \leqslant J$, $J \geqslant 2$. Hence, $A_1, A_2, ..., A_J \subset AC$. Without loss of generality, suppose $C_1, ..., C_q \subset AC$, $q \geqslant 1$, and $C_{q+1}, ..., C_k \cap AC = \emptyset$. For any $1 \leqslant i \leqslant J$, $AC - A_i$ is a connected branch in $G_0 - A_i$, which means that $C_1, ..., C_q$ does not connected with $A - AC$,and $C_{q+1}, ..., C_k$ does not connect with $AC - A_i$. If $J \geqslant 2$, $C_{q+1}, ..., C_k$ does not connect with $AC - A_1$ and $AC - A_2$. But $AC - A_1 \cup AC - A_2 = A$. So $C_{q+1}, ..., C_k$ does not connect with $AC$. And $C_1, ..., C_q$ does not connect with $AC$. So $AC$ and $A$ are not connected! Hence, each $AC_k$ can near appear twice in constructing $A_i$, $1 \leqslant i \leqslant k$. Therefore there exists one way to construct $A$ from $\mathcal{S}^*$ such that all the coefficients $\pi_{ij} \leqslant 1$, for $1 \leqslant j \leqslant p-2$.

Hence, the optimal solution of the problem $\mathcal{R}$ is $s_0$. And $\mathcal{I}_0 = \widehat{\mathcal{I}}$. □

Ye Luo
University of Florida
College of Liberal Arts & Sciences
Department of Economics
PO Box 117140
313 MAT
Gainesville, Florida 32611-7140
USA
E-mail: kurtluo@gmail.com

Hai Wang
Singapore Management University
School of Information Systems
80 Stamford Road
Singapore
E-mail: wanghaimit@gmail.com

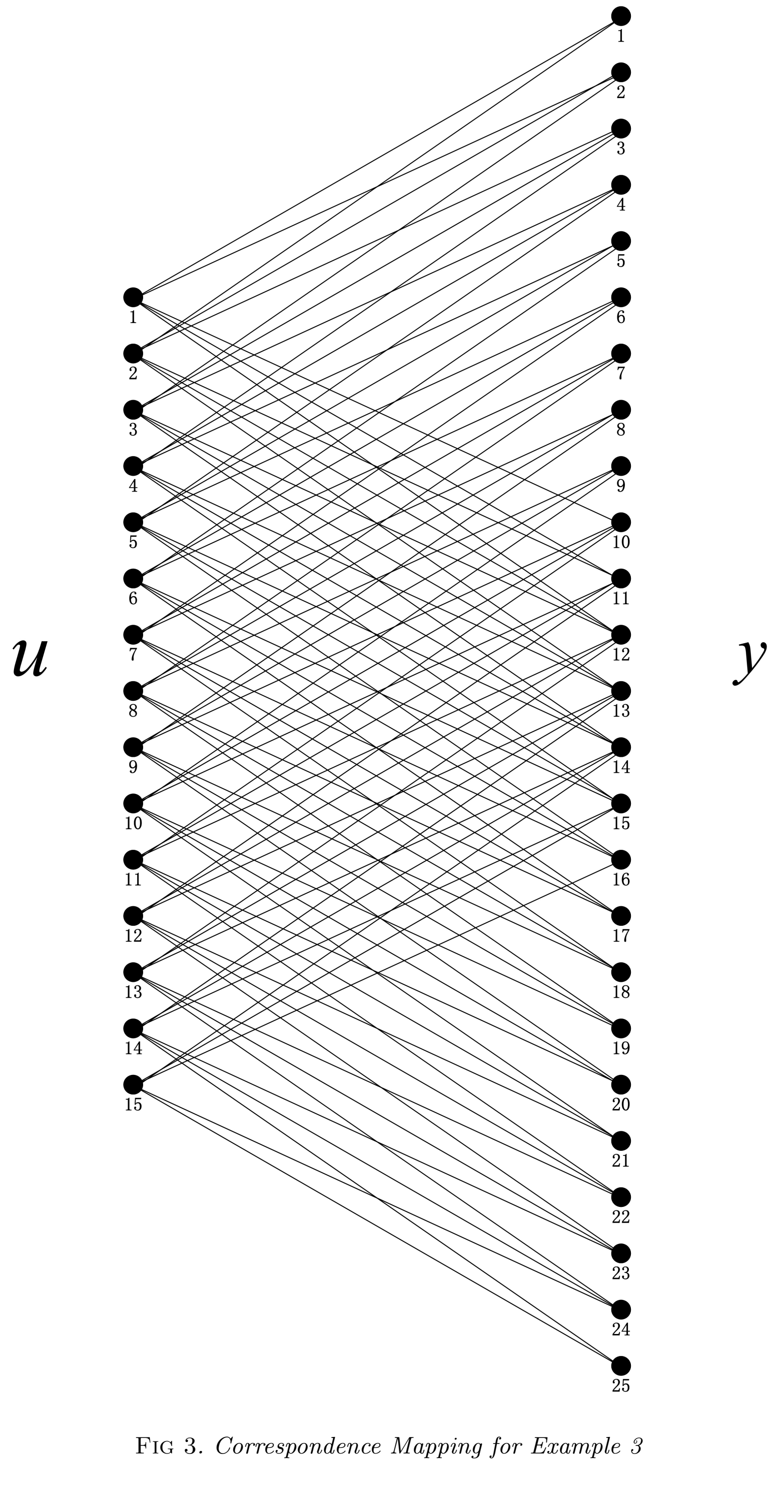

Fig 3. *Correspondence Mapping for Example 3*

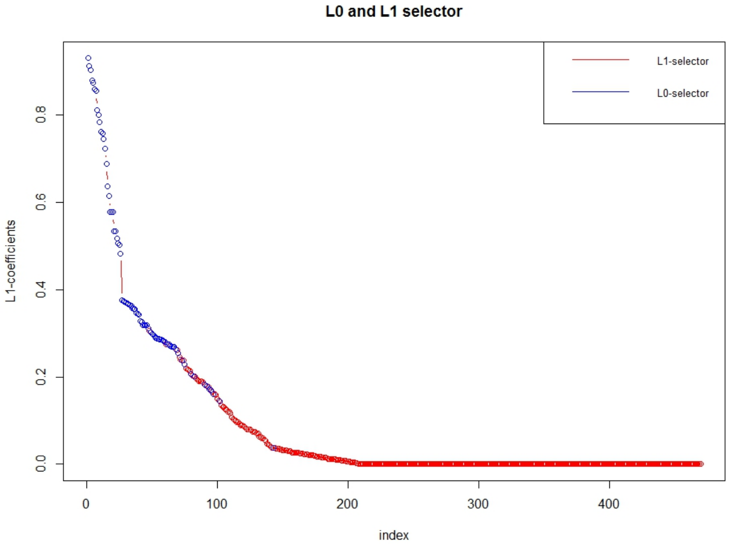

FIG 4. $L^0$ *versus* $L^1$*: with respect to* $L^1$ *Coefficient*

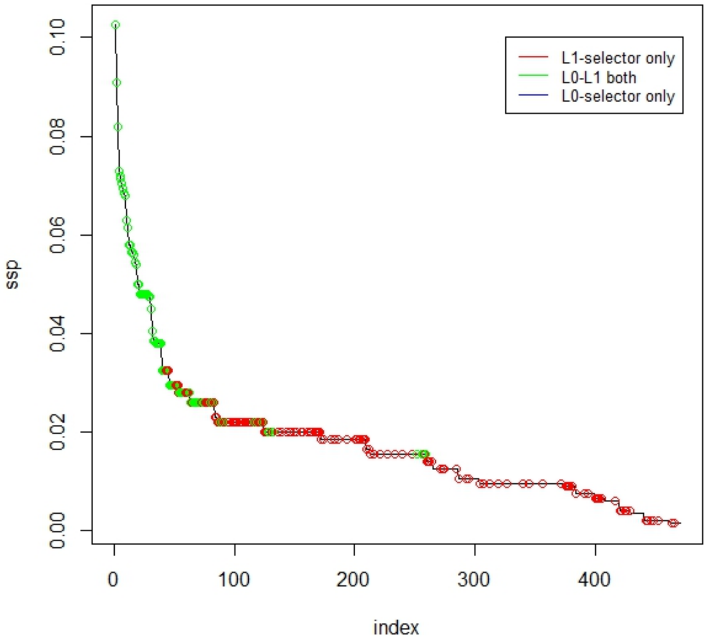

FIG 5. $L^0$ *versus* $L^1$*: with respect to Inequality Separation*

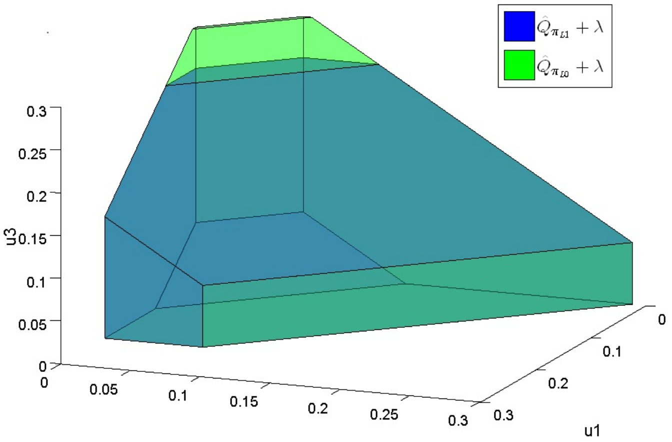

FIG 6. $L^0$ *versus* $L^1$*: Projection onto* $v_1, v_2, v_3$*.*

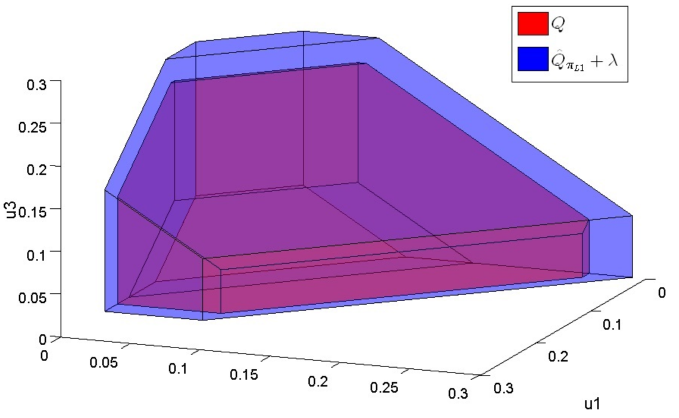

Fig 7. *$L^1$ versus True Feasible Set: Projection onto $v_1, v_2, v_3$.*

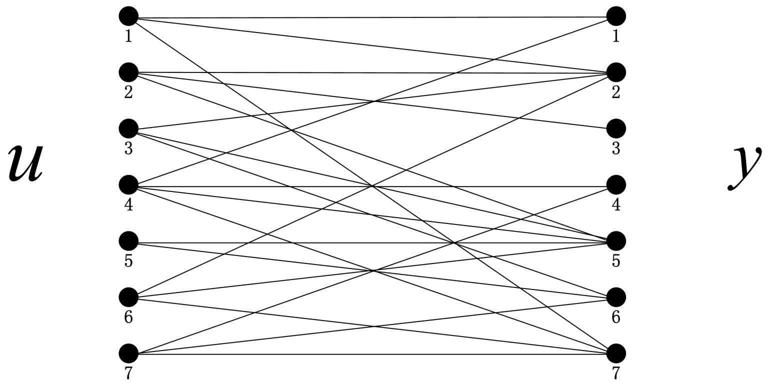

Fig 8. *Correspondence Mapping for Example 4*