ESSAYS IN PARTIAL IDENTIFICATION

Lukáš Lafférs

Dissertation submitted to the
Department of Economics,
NHH - Norwegian School of Economics,
in partial fulfilment of the requirement for the PhD degree.

December 2013

Contents

ACKNOWLEDGMENT

I would like to express my deepest gratitude to my advisor, Gernot Doppelhofer, for his generous guidance and support. Gernot kept me motivated at all times, and it has been a real pleasure to work under his supervision.

I also thank Alexei Onatski, whose inspiring lectures on partial identification captured my interest in the field and eventually led to this dissertation. His invaluable help shaped my research interests at an early stage of my PhD studies.

My thanks also go to the members of the doctoral committee; their time and efforts are highly appreciated.

I sincerely thank all my colleagues and friends at the Department of Economics at NHH for a very friendly and stimulating environment. It has been my greatest pleasure to spend a few wonderful years in Bergen, where the Department made me feel very welcome at all times. I would like to thank the administrative staff, especially Dagny Kristiansen, for all their help throughout the years.

I am very thankful for the friendships and inspiring discussions with my PhD colleagues at NHH; namely Magne, Elias, Tunç, Dada, Kamil, Michal, Peter, Trond, Kiki, Helge, Morten, Steinar, Kristina, Agnes, Sebastian, Grant, Kathrine, Suzanna and Mario, to mention a few.

There were many other people whose comments led to considerable improvements in my work; most notably, Marc Henry, Peter Molnar, Toru Kitagawa, Andrew Chesher, Ivan Sutoris, Adam Rosen, Charles Manski, Erik Sørensen, Konrad Smolinski, Martin Huber, to name a few.

Thanks also to my family and friends, especially to my lovely wife Zuzka for her patience and love.

Thank you!

INTRODUCTION

The partial identification literature relates to situations in which data, together with the economic model, do not uniquely determine the parameter of interest. In such situations, the parameter is *partially* identified. There is a set of possible values of the parameter that cannot be refuted by the economic model, an *identified set*. Traditionally, most economic models include assumptions that guarantee that there exists a unique parameter value that is compatible with the data and the economic model, so the parameter is *point* identified. Assumptions about functional forms or distributional assumptions are often not based on theoretical grounds, but their sole purpose is to guarantee the point identification. It is interesting to consider what we can learn about the parameter of interest from the economic model alone, leaving these ad hoc assumptions aside. This should ultimately lead to a more credible inference, but it often comes at the cost of the loss of the point identification, which may introduce computational and statistical challenges. There is also a trade-off between the strength of the assumptions and the credibility of the analysis.

”The law of decreasing credibility: The credibility of inference decreases with the strength of the assumptions maintained.” Manski (2003).

Policy makers facing a decision may consider an identified set as undesirable, compared with a point-identified model. Yet it seems more prudent to choose a policy from the set of admissible policies according to some transparent rule (e.g., to consider the worst case scenario) than to leave the ad hoc assumptions to make the choice for us instead.

There are two different ways to approach identification. The first is to start with a point-identified model and then examine how different assumptions affect the size of the identified set. The partial identification literature provides useful tools for studying model uncertainty, misspecification, sensitivity analysis and missing data. Knowledge of the strength of the underlying assumptions helps to direct the discussions toward the relevant parts of the economic model. Another option is to examine what can be learned from the data alone and then observe how different assumptions change the identified set, as advocated in Manski (1995, 2003).

Here we only introduce the central concepts in the partial identification literature. Tamer (2010) provides a comprehensive review of the history of thought on the topic.

Partial identification offers clear separation between two important and distinct issues: *identification* and *statistical inference*. Identification considers the following hypothetical situation: if we knew the true probability distribution of the observed variables (or had a data sample of infinite length), what could we learn about the parameter of interest? Once the identification issue is resolved, it is possible to consider statistical inference; that is, how the imperfect data affect the conclusions drawn. The first essay of this thesis addresses the question of identification in models with discrete variables, the second essay considers the statistical inference of the parameter of interest and the third essay applies these results to an empirical application. The fourth chapter is a note that uses the method to point out that conditional and unconditional identifying assumptions are often confused in the applied literature.

The following subsections present the contributions in greater detail.

Chapter 1: Identification in Models with Discrete Variables

Chapter 1 introduces a novel identification method that can determine the identified set in models with discrete variables. This method can replicate some existing results in a straightforward manner, as well as address new problems, and it shows how imperfect instruments affect the size of the identified set, when the strict exogeneity assumption is relaxed. The method is an extension of the partial identification framework of Galichon and Henry (2009a), and it is simple and computationally tractable, and provides a unifying framework that approaches identification in an algorithmic fashion.

Chapter 2: Inference in Partially Identified Models with Discrete Variables

The problem of statistical inference in partially identified models was not addressed in the first chapter. Chapter 2 addresses this problem for a partially identified scalar parameter in models with discrete variables. This paper presents different methods for inference and discusses their advantages and disadvantages. A Monte Carlo simulation study compares the finite sample properties of different methods in economically relevant economic scenarios. The chapter concludes with some practical implementation recommendations on how to implement the inference in this context.

Chapter 3: Bounding Average Treatment Effects using Linear Programming

Chapter 3 shows how the modified identification method from the first chapter can determine the sharp bounds on the average treatment effect under identifying assumptions commonly used in the literature. This method provides a way to conduct sensitivity analysis for the identifying assumptions and missing data in an empirical application concerning the effect of parent's schooling on a child's schooling (de Haan, 2011).

Chapter 4: A Note on Bounding Average Treatment Effects

published in Economics Letters 2013, 120, (3), 424-428

Using the linear programming identification framework from the first paper, it is possible to gain deeper insight into the source of the identification power. Two commonly made assumptions in empirical studies imply bounds on the average treatment effect that differ from those commonly reported in the applied literature. Instead, one of the assumptions should hold conditionally on the value of a specific variable. Chapter 4 explains the difference between the reported bounds and the correct bounds in detail, and shows why it matters, using an empirical example from de Haan (2011). Based on the analysis in this chapter, we recommend that the required conditioning should be stated explicitly.

Chapter 1

IDENTIFICATION IN MODELS WITH DISCRETE VARIABLES

Abstract

This paper provides a novel, simple, and computationally tractable method for determining an identified set that can account for a broad set of economic models when the economic variables are discrete. Using this method, we show using a simple example how imperfect instruments affect the size of the identified set when the assumption of strict exogeneity is relaxed. This knowledge is of great value, as it is interesting to know the extent to which the exogeneity assumption drives results, given it is often a matter of some controversy. Moreover, the flexibility obtained from our newly proposed method suggests that the determination of the identified set need no longer be application specific, with the analysis presenting a unifying framework that algorithmically approaches the question of identification.

JEL: C10, C21, C26, C61.

Keywords: Partial identification, Linear programming, Imperfect instruments.

1.1 Introduction and Motivation

Identification plays a central role in economic research. In most economic models, we introduce latent variables, such as unobserved heterogeneity, ability, or preference shocks, to explain relations of interest, such that the model best mimics reality. Given data that reveal the distribution of observable variables, we would prefer to learn as much as possible about the relations or features of the economic model, information often embedded in an unknown parameter. Unfortunately, as latent variables are not directly observable, we need to make certain assumptions about them in order to use data to say something about an unknown parameter or some feature of interest. Depending on the strength of these assumptions, knowledge of the true data-generating process for the observed variables can then be any of the following: (1) no identifying power, (2) a contraction of the set of potential parameter candidates, such that the model is partially identified, (3) the assumptions are sufficient to identify one potentially true parameter, such that the model is point identified, or (4) the assumptions are too strict and the model can be refuted.

In practice, we often require strong assumptions to guarantee point identification. However, such assumptions could include knowledge of the family of probability distributions of unobserved variables, information we can rarely justify on economic grounds. The only reason is to make inference tractable. It is then interesting to question what would happen had these restrictions not been imposed, and then attempt to develop an inferential procedure that is robust with respect to assumptions that are sometimes controversial or made purely for technical convenience. The first necessary step is to know what set of models (or parameters) are compatible with both the set of assumptions made and the data in situations where we have perfect information on the probability distribution of observable variables, that is, where our data sample is of infinite length. This is the question of identification. Once this is resolved, we can proceed to inference and identify how to use imperfect data to construct confidence regions or hypothesis tests.

The contribution of this paper is threefold. First, we present a new simple identification method. Second, we show how this method nests several existing results from the literature. Third, we show how this method approaches identification in cases when the strict exogeneity of instruments is relaxed. The main advantage over the existing literature is that the economic model is not restricted to the linear form, while at the same time controlling for the degree of violation of the exogeneity assumption.

This paper presents a new method as an extension of an existing framework by Galichon and Henry (2011, 2009a) and Ekeland, Galichon, and Henry (2010) (henceforth, the GH framework) that traces the identified set in a richer set of economic problems when the observed variables are discrete. As a motivating example, we consider the impact of a violation of the strict exogeneity assumption in a single-equation endogenous binary response model. By complementing existing results on imperfect instruments in Nevo and Rosen (2012), Conley et al. (2012), this method can control for departures from the strict exogeneity of the instrument, and permits us to study nonlinear models.

The proposed method is also able to reproduce some other results in the partial identification literature obtained using different approaches. These include the single-equation endogenous binary response model in Chesher (2009) and Chesher (2010), the triangular system of equations with binary dependent variables in Shaikh and Vytlacil (2011), treatment effects in studies with imperfect compliance as in Balke and Pearl (1997), and binary choice models with zero-median restrictions as in Komarova (2013). In the first and fourth examples, the original GH framework[1] also applies, but our extension helps us to formulate the problem in such a way that it is possible to relax the strict exogeneity of instruments more simply, as in Section 1.4. In the remaining examples, the extension is essential, as we cannot formulate some of the assumptions made within the original GH framework. The present extension therefore enriches the set of problems we can address.

The major advantage of this new method is its algorithmic structure: that is, the identifying restrictions enter the setup in a straightforward manner and it employs effective algorithms to determine the identified set. Instead of using distinct strategies for different applications, this method thus provides a unifying framework that is conceptually simple. As the framework presented is not application specific, it thus applies to a wide range of problems including discrete variables when identification is only partial.

[1] With some modification.

Of course, we also recognize several limitations of the proposed method. First, the method describes how we find the identified set given perfect information on the data-generating process of the observed variables, yet we do consider inference here. Second, we restrict the observable variables in the model to be discrete. While we can discretize models with continuous observable variables, this will always bring about some degree of arbitrariness in the problem, and we do not consider the impact of this here. However, while we do *not* restrict the unobservable variables to be discrete, we can always transform a continuous unobservable variable into a discrete form, and we show that this will not affect the identified set.

Manski (1990) initiated the study of partial identification. However, these ideas were not fully appreciated at first. Recent studies include Manski (1995) and Manski (2003), with useful surveys of this literature by Manski (2008) and Tamer (2010). Among the many interesting applications, the most notable include recent work on the returns to schooling (Manski and Pepper, 2000), the demand for fish (Chernozhukov et al., 2009), and discrete choice with social interactions (Brock and Durlauf, 2001). Determination of the identified set is examined in Galichon and Henry (2011, 2009a) by means of an optimal transportation formulation, in Beresteanu and Molinari (2008), Beresteanu et al. (2012, 2011), and Chesher, Rosen, and Smolinski (2013) using random set theory, and in Chesher (2010) using structural quantile functions. Readers interested in statistical inference in the partially identified setting are directed to Galichon and Henry (2011, 2009a), Chernozhukov, Hong, and Tamer (2007), Imbens and Manski (2004), Beresteanu and Molinari (2008), Beresteanu et al. (2012, 2011), Chernozhukov, Lee, and Rosen (2013), Andrews and Shi (2013), Romano and Shaikh (2010),Bugni (2010), and Rosen (2008).

The remainder of the paper is structured as follows. Section 1.2 describes the identification strategy in GH using the proposed extension. In Section 1.3, we provide examples of how the extended framework can nest the different identification approaches. Section 1.4 explains how we can modify one of the examples to consider the impact of imperfect instruments. Section 1.5 concludes the paper and the appendices provide the proofs (Appendix 1.6), technical details on the examples presented (Appendix 1.7), and discussion of some of the implementation issues (Appendix 1.8).

1.2 Methods

This section first explains the basic elements of the partial identification framework in GH, and then motivates and presents our extension.

1.2.1 GH's Framework

Here we present the basic components of the GH identification setup. Let

- $Y \in \mathcal{Y}$ be a random vector of observable variables with probability density function or probability mass function (pdf or pmf) p,
- $U \in \mathcal{U}$ be a random vector of unobservable variables with pdf or pmf ν, and
- $G : \mathcal{U} \mapsto \mathcal{Y}$ be a measurable correspondence[2] that restricts the co-occurrence of pairs (Y,U) to those that are compatible with the economic model at hand, formally $Y \in G(U)$. This represents how economic restrictions are modeled within the GH setup.

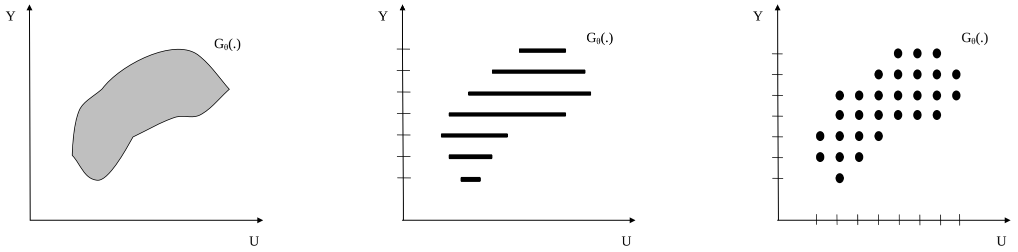
The fact that G is a many-to-many correspondence enables us to work with censored data (for a given Y we contemplate different values of U) or multiple equilibria (for a given U, we consider different values of Y). Figure 1.1 illustrates many-to-many mapping G. Note that point identification is typically achieved if both Y and U are continuous and the inverse of the many-to-many mapping G^{-1} is a function. In this case, knowledge of the probability behavior of the observed variables informs us *exactly* of the probability of the unobserved component.

We first define the concept of a **Structure** that groups all available restrictions.

Definition 1. *A structure S is defined as a triplet $S = (G, \nu, p)$.*

Another important notion to be defined is the **internal consistency** of a structure. The structure is internally consistent if there exists a joint distribution which potentially could have generated the probability of the observed variables p and the latent variables ν and that satisfies the economic restriction defined by G almost surely. If there is no such joint distribution, the structure can clearly be refuted.

[2] Therefore for all open subsets A of $\mathcal{Y}$, $G^{-1}(A) := \{U \in \mathcal{U} : G(U) \cap A \neq \varnothing\}$ is well defined.



(a) Y and U continuous

(b) Y discrete and U continuous

(c) Y and U discrete

Figure 1.1: Illustration of the correspondence G that carries information about the economic model. The joint distribution of (Y, U) is restricted to have support in the gray-shaded area with probability one.

Definition 2. *Structure S is said to be internally consistent if and only if there exists a joint probability distribution π of (Y, U) on $\mathcal{Y} \times \mathcal{U}$ with marginal distributions p and ν, respectively, such that $Pr_\pi(\{Y \in G(U)\}) = 1$.*[3]

In practice, most models are parameterized, so we now consider the situation when $\nu = \nu_\theta$ and $G = G_\theta$ are parameterized with a vector of parameters $\theta \in \Theta$, where $\Theta \subseteq \mathbb{R}^d$.[4] Finally, we define our object of interest, the **identified set**. This is the collection of all parameters θ that guarantee the internal consistency of the structure.

Definition 3. *An identified set for θ, $\Theta_I(p)$, is defined as*

$\Theta_I(p) := \{\theta \in \Theta : (G_\theta, \nu_\theta, p)$ is internally consistent$\}$.[5]

Note that all members of the identified set correspond to structures that could have generated the probability of the observed variables p. In this sense, they are *observationally equivalent*, and no amount of data would ever help us to distinguish between them. The identified set

- could be empty: $\Theta_I(p) = \{\varnothing\}$, hence the structure $(G_\theta, \nu_\theta, p)$ is *refuted* for all $\theta \in \Theta$,
- may consist of a single point: $\Theta_I(p) = \{\theta\}$, in this case θ is *point identified,*
- can be a subset of Θ: $\Theta_I(p) = \{I \subset \Theta\}$ and θ is *partially identified,* or

[3] Definition 1 in Galichon and Henry (2009a).

[4] The parameter θ may consist of two parts, $\theta = [\theta_1, \theta_2]$, so we can have G_{θ_1} and ν_{θ_2}.

[5] Definition 2 in Galichon and Henry (2009a), where the dependence of the identified set $\Theta_I(p)$ on the distribution of observable variables p is made explicit.

- may not shrink Θ at all: $\Theta_I(p) = \Theta$, so the structure $(G_\theta, \nu_\theta, p)$ places *no identifying* restrictions on θ.

For a fixed parameter θ, if all variables in the model are discrete, the problem of finding a joint distribution of (Y, U) compatible with the economic model described by G_θ with appropriate marginals can be formulated as a linear program as shown. Note that in most economic applications, the latent component U is continuous. If the observed variables are discrete, it is, however, possible to discretize U in a way that leaves the identified set unchanged as proved in Galichon and Henry (2011). Suppose $\mathcal{Y} = \{y_1, ..., y_i, ..., y_n\}$ with corresponding probabilities p_i, $\mathcal{U} = \{u_1, ..., u_j, ..., u_m\}$ with probabilities ν_j. The economic model enters the problem as a set of restrictions on the support of (Y, U). Let us define a zero-one penalty on the support of all joint probabilities on $\mathcal{Y} \times \mathcal{U}$:

$$c_{ij} = 1(y_i \notin G_\theta(u_j)) = \begin{cases} 0, \text{ if } y_i \in G_\theta(u_j), \\ 1, \text{ otherwise,} \end{cases}$$

so a penalty is put on those pairs (Y, U) that are incompatible with the economic model. The $n \times m$ matrix of the zero-one penalties $\{c_{ij}\}$ carries the same information as the mapping $G_\theta(.)$ and we denote the nm vector of this stacked matrix as $\boldsymbol{c}$.

Now, the question of the existence of a joint probability distribution that assures internal consistency can be answered by means of the following linear program:[6]

$$\begin{aligned} &\min\nolimits_{(\pi)} \textstyle\sum_{i,j} \pi_{ij} c_{ij} \\ &\text{s.t.} \\ &\textstyle\sum_j \pi_{ij} = p_i, \qquad \forall i \qquad (1.1) \\ &\textstyle\sum_i \pi_{ij} = \nu_j, \qquad \forall j \qquad (1.2) \\ &\pi_{ij} \geq 0, \qquad \forall i, j, \qquad (1.3) \end{aligned}$$

where the minimum is taken across all joint probability distributions π (nm vector of the stacked $n \times m$ matrix with elements $\{\pi_{ij}\}$). A structure is internally consistent if and only if the optimized value of the objective function is equal to 0. If this is the case, it means that we have found a proper joint distribution π that is compatible

[6] The dependence of c_{ij} and ν_j on parameter θ is omitted for the sake of brevity.

with the data (1.1) and the assumptions made on the latent variables (1.2), and the probability of an event not being compatible with the economic model is zero.

The necessary and sufficient condition for the inclusion of the parameter θ in the identified set is:

$$0 = \max_{A \subset \mathcal{Y}} (Pr(A) - \nu_\theta(G_\theta^{-1}(A))), \tag{1.4}$$

where the maximum is taken across all possible subsets of $\mathcal{Y}$. A similar result was first proven by Artstein (1983), and is based on an extension of the Marriage Lemma. Alternative proofs of (1.4) were given in Galichon and Henry (2009a), which relied on optimal transportation theory, and in Henry et al. (2011) based on combinatorial optimization methods. Equation (1.4) can then also be used for hypothesis testing or building confidence regions for θ, as proposed in Galichon and Henry (2009a) and Henry et al. (2011). The latter allows for efficient confidence regions construction using a combinatorial bootstrap.

The properties of the approach are as follows.

- It offers a flexible procedure to access many problems when partial identification occurs.
- For discrete cases, its linear program nature makes it computationally convenient.
- If only U is continuous, the problem can be transformed into a discrete form.
- The economic model is described by restrictions on the support of observables and unobservables.

1.2.2 Extension of the GH Framework

We aim to extend the GH method to entertain additional distribution restrictions. Even though the GH setup can address many problems, we are unable to formulate certain types of problems within the GH framework. There are two ways in which our prior information can enter the structure: first, via the marginal distribution of the unobservables ν, and second, through the support of (Y, U) via the correspondence G (or equivalently $\boldsymbol{c}$). However, not all distributional assumptions we can conceive can enter the structure, because in many economic models some notion of independency is assumed.[7]

Because the problem is accessed at the lowest level, by constructing a joint distribution compatible with all the information a researcher may have, it is possible to restrict this joint distribution to satisfy *any* type of distributional assumptions one may wish to make. If the distributional assumption can be written as a linear function of the joint probability π, the problem remains computationally attractive. Modeling the joint distribution gives full control over utilizing the information at hand. This flexibility delivers a solution to cases where the GH setup is too restrictive, and this is the main contribution of the present analysis.

For illustrative purposes, suppose that in addition to information about G, we know that $E(\phi_\theta(Y,U)) = 0$ and $|cov(Y,U)| \leq 0.1$. Such assumptions simply cannot be formulated as a restriction on the support of (Y,U), so there is no way that these assumptions can be embedded into the framework via G or ν. In this sense, the original GH framework is too restrictive. Instead, the way to incorporate these assumptions is simply to restrict the set of joint distributions (all π-s) to only those that are compatible with this piece of information.

The question of whether the extended set of restrictions is compatible with the observed data then reduces to checking whether the optimized value is equal to zero

[7] We may be willing to make some assumptions about the distribution of variables in the form of moment equality or inequality. It is important to note here that the GH setup can handle moment inequalities $E(\phi(Y)) \leq 0$ if $E(m(U)) = 0$ is assumed (Ekeland et al. (2010) and Henry and Mourifié (2013)). In this case, the correspondence G is restricted to take a specific form. However, within the GH framework, it is not possible to consider moment inequality *and* further information given by G.

in the following linear program:

$$
\begin{aligned}
&\min_{(\pi)} \textstyle\sum_{i,j} \pi_{ij} c_{ij} \\
&\text{s.t.} \\
&\textstyle\sum_j \pi_{ij} = p_i, && \forall i \\
&\textstyle\sum_i \pi_{ij} = \nu_j, && \forall j \\
&\textstyle\sum_{i,j} \pi_{ij} \phi_\theta(y_i, u_j) = 0, && (1.5) \\
&\textstyle\sum_{i,j} \pi_{ij} y_i u_j - \sum_i p_i y_i \sum_j \nu_j u_j \leq 0.1, && (1.6) \\
&-\textstyle\sum_{i,j} \pi_{ij} y_i u_j + \sum_i p_i y_i \sum_j \nu_j u_j \leq 0.1, && (1.7) \\
&\pi_{ij} \geq 0, && \forall i, j.
\end{aligned}
$$

Equation (1.5) restricts the joint distribution π to satisfy $E(\phi_\theta(Y, U)) = 0$, whereas inequalities (1.6) and (1.7) ensure that $|cov(Y, U)| \leq 0.1$ is satisfied.

As another example, suppose we have two observed variables $Y = (X, Z)$ with probabilities p_{ij} and an unobserved variable U, but instead of assuming full knowledge of its distribution, we assume that it has zero mean, that its 75% quantile is 0.8, and that it is independent of Z. We now formulate the problem as follows:

$$
\begin{aligned}
&\min_{(\pi)} \textstyle\sum_{i,j,k} \pi_{ijk} c_{ijk} \\
&\text{s.t.} \\
&\textstyle\sum_k \pi_{ijk} = p_{ij}, && \forall i, j \\
&\textstyle\sum_{i,j,k} \pi_{ijk} u_k = 0, \\
&\textstyle\sum_{i,j,k} \pi_{ijk} 1(u_k \leq 0.8) = 0.75, \\
&\textstyle\sum_i \pi_{ijk} - \sum_i p_{ij} \sum_{i,j} \pi_{ijk} = 0, && \forall j, k \\
&\pi_{ijk} \geq 0, && \forall i, j, k.
\end{aligned}
$$

These examples are somewhat artificial, but explain the main point well. Economically interesting examples follow in Section 1.3. It is important to note that if the additional constraints are such that the problem lies within the linear programming framework, it remains computationally feasible.

The crucial step is to prove that the discretization of the unobserved variables is possible *even* when additional distributional restrictions are entertained. This is done for a certain class of distributional restrictions, and is discussed in detail in Subsection 1.2.3, with the proof given in Appendix 1.6.

We now state the proposed extension formally. We recall that $\mathcal{Y}$ and $\mathcal{U}$ are the supports of the discrete observable variable and the continuous or discrete unobservable variables, respectively. The set of all probability distributions on $\mathcal{Y} \times \mathcal{U}$ is denoted by $\mathbf{\Pi}(\mathcal{Y}, \mathcal{U})$, and $\psi_\theta(\mathcal{Y}, \mathcal{U}, p, \nu)$ is the set of all $\pi \in \mathbf{\Pi}(\mathcal{Y}, \mathcal{U})$ satisfying additional restrictions imposed. If information about the probability distribution ν of unobserved variables is not available, we have $\psi_\theta(\mathcal{Y}, \mathcal{U}, p, .)$. The set of all restrictions imposed is then compatible with the data if and only if the optimal solution for the following optimization procedure is zero:

$$\begin{gathered}\min_{(\pi)} \pi\{1(Y \notin G_\theta(U))\} \\ \text{s.t.} \\ \pi\{1(Y = y_i)\} = p_i, \qquad \forall i \\ \pi \in \psi_\theta(\mathcal{Y}, \mathcal{U}, p, \nu).\end{gathered}$$

Note that if U is discrete and the set ψ_θ consists of restrictions that are linear in π, linear programming routines may be used.

The additional restrictions for the two examples given above are:

$$\psi_\theta(\mathcal{Y}, \mathcal{U}, p, \nu) = \left\{ \pi \in \mathbf{\Pi}(\mathcal{Y}, \mathcal{U}) : \begin{array}{l} \forall u \in \mathcal{U} : \pi\{1(U = u)\} = \nu(u), \\ E_\pi \phi_\theta(Y, U) = 0, \\ \left| E_\pi YU - \sum_i p_i y_i \sum_j \nu_j u_j \right| \leq 0.1 \end{array} \right\} \tag{1.8}$$

and:

$$\psi(\mathcal{X} \times \mathcal{Z}, \mathcal{U}, p, .) = \left\{ \pi \in \mathbf{\Pi}(\mathcal{X} \times \mathcal{Z}, \mathcal{U}) : \begin{array}{r} E_\pi U = 0, \\ E_\pi 1(U \leq 0.8) = 0.75, \\ \forall z \in \mathcal{Z}, \forall u \in U : \pi\{1(Z = z, U = u)\} = \\ = \sum_i p_{ij} \pi\{1(U = u) \ \} \end{array} \right\} \tag{1.9}$$

where in the second example $\mathcal{Y} = \mathcal{X} \times \mathcal{Z}$ and ψ does not depend on θ.

We now redefine the notion of structure and the identified set. To enrich the concept of the original structure, we denote a triplet (G, ψ, p) as a **Generalized Structure**, which groups all the restrictions placed on π.

Definition 4. *A Generalized Structure S is defined as a triplet* $S = (G, \psi, p)$*.*

Internal consistency and **identified set** are then defined similarly as in definitions 2 and 3.

Definition 5. *A Generalized Structure S is said to be internally consistent if and only if there exists a joint probability distribution* π *of* (Y, U) *on* $\mathcal{Y} \times \mathcal{U}$ *in* $\psi(\mathcal{Y}, \mathcal{U}, p)$ *with Y-marginal distribution p such that* $\pi(\{Y \in G(U)\}) = 1$*.*

Definition 6. *An identified set for* θ*,* $\Theta_I(p)$*, is defined as*

$\Theta_I(p) := \{\theta \in \Theta : (G_\theta, \psi_\theta, p)$ is internally consistent$\}$.

We refer to this formulation as the *extended GH framework*. If the latent variable U is discrete and the set ψ can be written as linear restrictions in π, we can employ effective algorithms to solve this linear program.

1.2.3 Discretization of Unobserved Variables

In most economic problems, the unobserved component is continuous. Hence, in order to make the search in the space of joint probability functions tractable, it may be convenient to discretize the unobserved component and then show that this discretization leaves the identified set unaffected. This is not true in general. We show that if the distributional restrictions ψ take specific forms that nest all of the examples presented in this paper, the discretization of the unobserved variable is possible and harmless. These sets of restrictions for problems with continuous unobservables are:

$$\psi_1(\mathcal{Y},\mathcal{U},p,\nu) = \left\{ \pi \in \mathbf{\Pi}(\mathcal{Y},\mathcal{U}) : \begin{array}{c} \forall u \in \mathcal{U} : \pi\{1(U=u)\} = \nu(u), \\ \forall I \in \mathbb{I}; \forall u \in \mathcal{U} : \\ |\sum_{i\in I} \pi(y_i,u) - \sum_{i\in I} p_i \nu(u)| \leq \alpha \sum_{i\in I} p_i \nu(u) \end{array} \right\}, \tag{R1}$$

and:

$$\psi_2(\mathcal{Y},\mathcal{U},p,\cdot) = \left\{ \pi \in \mathbf{\Pi}(\mathcal{Y},\mathcal{U}) : \begin{array}{l} E_\pi \phi(U) = 0, \\ \forall I \in \mathbb{I}; \forall u \in \mathcal{U} : | \sum_{i\in I} \pi(y_i,u) - \\ - \sum_{i\in I} p_i \pi\{1(U=u)\}| \leq \alpha \sum_{i\in I} p_i \pi\{1(U=u)\} \end{array} \right\}, \tag{R2}$$

where $\phi : \mathcal{U} \mapsto \mathcal{M}$ has a finite range $\mathcal{M}$ and $\mathbb{I}$ is a fixed set of indices.[8]

The first restriction (R1) requires π to be compatible with the assumed distribution of unobserved variables and hence nests the original GH framework. The second restriction helps us to restrict part of the observed component to be independent or "close to being independent" of the unobserved component, while the first line in (R2) permits us to work with quantiles of U.[9]

Let us denote the question of the internal consistency of a generalized structure (G,ψ,p) with a continuous unobserved variable as $\mathcal{P}_1$:

Y discrete with support $\mathcal{Y} = \{y_1,...,y_n\}$ and with probability $p = \{p_1,...,p_n\}$,
U continuous with support $\mathcal{U}$ (and with positive probability density ν for (R1)),
$G : \mathcal{U} \mapsto \mathcal{Y}$.

The aim is to find a function $\pi_1 : \mathcal{Y} \times \mathcal{U} \mapsto [0,1]$ that satisfies:

[8] If the observed variable is multidimensional we can stack it into a single vector. Summing across some sets of indices then allows us to formulate a restriction for only one dimension. As an example, suppose that the observed variables are (Y,X,Z); then, we can place a restriction on X only, so that X is independent of U.

[9] The manner in which the independency restriction is relaxed is discussed in Section 1.4.

$$\begin{aligned} & \sum_{i=1}^{n} \int_{u \in \mathcal{U}} \pi_1(y_i, u) 1(y_i \in G(u)) du = 1, \\ \forall i = 1, ..., n : \quad & \int_{u \in \mathcal{U}} \pi_1(y_i, u) du = p_i, \\ & \pi_1 \in \psi(\mathcal{Y}, \mathcal{U}, p, \cdot). \end{aligned}$$

Problem $\mathcal{P}_1$ is computationally unfeasible because of its continuous component U. We can, however, transform the problem $\mathcal{P}_1$ with a continuous U to the problem $\mathcal{P}_2$ with a discrete U, such that it will not affect the identified set.

We partition $\mathcal{U}$ into subsets that deliver the same $G(U)$ for the set of restrictions (R1) and into those that deliver the same $G(U)$ and $\phi(U)$ for (R2). It is then easy to show that if we group all Us in these subsets into atoms and proceed as if U were discrete, the identified set remains unchanged.

Formally, the partitioning of the $\mathcal{U}$ space is the following:

$$\mathcal{G} \equiv \{\Delta^* \subset \mathcal{U} : \forall g_I \in \Delta^*, \forall g_{NI} \in \Delta^{*C} : G(g_I) \neq G(g_{NI})\} \tag{PartU1}$$

for (R1) and:

$$\mathcal{S} \equiv \{\Delta^* \subset \mathcal{U} : \forall s_I \in \Delta^*, \forall s_{NI} \in \Delta^{*C} : G(s_I) \neq G(s_{NI}),\ \phi(s_I) \neq \phi(s_{NI})\} \tag{PartU2}$$

for (R2).

The assumption of a finite range of ϕ is crucial, as it implies a finite $\mathcal{S}$. Let m denote the cardinality of either $\mathcal{G}$ or $\mathcal{S}$, depending on which is in use. Then, a new random variable U^* is defined. For every $j \in \{1, ..., m\}$, we choose a point of support u_j^* to be any $u \in \Delta_j^*$, a representative of the set Δ_j^*:

$$\mathcal{U}^* \in \Delta_1^* \times \cdots \times \Delta_m^*. \tag{U}$$

To obtain a probability distribution ν^* of U^*, needed for restrictions (R1), we integrate $\nu(u)$ across the corresponding regions Δ_j^* of $\mathcal{U}$:

$$\forall j = 1, ..., m: \qquad v_j^* \equiv \int_{\Delta_j^*} v(u)du. \tag{P}$$

The discretized problem $\mathcal{P}_2$ is the following:

Y with support $\mathcal{Y} = \{y_1, ..., y_n\}$ with probability $p = \{p_1, ..., p_n\}$

U^* with support $\mathcal{U}^* = \{u_1^*, ..., u_m^*\}$ (with probability $v^* = \{v_1^*, ...v_m^*\}$ for (R1))

$G : \mathcal{U}^* \mapsto \mathcal{Y}$.

The question is then whether there exists a function $\pi_2 : \mathcal{Y} \times \mathcal{U}^* \mapsto [0, 1]$ such that:

$$\sum_{i=1}^{n} \sum_{j=1}^{m} \pi_2(y_i, u_j^*) 1(y_i \in G(u_j^*))du = 1,$$

$$\forall i = 1, ..., n: \qquad \sum_{j=1}^{m} \pi_2(y_i, u_j^*) = p_i,$$

$$\pi_2 \in \psi(\mathcal{Y}, \mathcal{U}^*, p, \cdot).$$

Lemma 1. *If* (R1),(P) *and* (PartU1) *hold, then a generalized structure $(G, \psi(\mathcal{Y}, \mathcal{U}, p, v), p)$ is internally consistent if and only if a generalized structure $(G, \psi(\mathcal{Y}, \mathcal{U}^*, p, v^*), p)$ is internally consistent.*

Lemma 2. *If* (R2) *and* (PartU2) *hold, then a generalized structure $(G, \psi(\mathcal{Y}, \mathcal{U}, p, \cdot), p)$ is internally consistent if and only if a generalized structure $(G, \psi(\mathcal{Y}, \mathcal{U}^*, p, \cdot), p)$ is internally consistent.*

The proofs are in Appendix 1.6. Lemmata 1 and 2 state that for the internal consistency of a generalized structure, the proposed discretization is harmless.

It immediately follows that if G and ψ were parameterized by some $\theta \in \Theta$, problem $\mathcal{P}_1$ and problem $\mathcal{P}_2$ would lead to the *same* identified set for θ for both (R1) and (R2).

1.3 Motivating Examples

This section introduces some examples of how the extended GH framework applies to problems in the partial identification literature. The replication of existing results illustrates that the proposed extension indeed works. That said, there is no computational gain from employing the present method over the other frameworks that derive analytical solutions. Rather, the greatest advantage of this method is its *generality*. Instead of deriving the identified set and proving that it is sharp from case to case, we propose a single unifying framework that traces the identified set regardless of the application. It is then sufficient to formulate the economic model with restrictions in the extended GH setup and let the computer do the work. Further, if additional information becomes available, it is straightforward to incorporate this into the setup. Unlike the existing application-specific approaches, where incorporating further restrictions or changing the existing restriction may cause significant difficulties for tracing the identified set, adding additional assumptions or changing existing assumptions in the extended GH framework is trivial. Moreover, if the distributional restrictions are linear in the joint probability π, we can employ linear programming routines. This is particularly interesting, as linear programming is widely understood and ready-to-use computer codes are readily available.

The four examples presented in this section not only demonstrate that the method nests several existing identification strategies and can thus replicate their results, but also illustrate how to formulate the economic problem at hand in the extended GH framework.

The four considered examples include the single-equation endogenous binary response model in Chesher (2009, 2010), the bounds on treatment effects in triangular models with binary dependent variables (Shaikh and Vytlacil, 2011), studies with imperfect compliance as in Balke and Pearl (1997), and binary choice models with zero-median restrictions as in Komarova (2013).

For each example, we first introduce the problem and the notation. We then present the discretization of the unobserved variables. Afterwards, we formulate the problem in the extended GH framework. Finally, we compare the results. The original identification strategy is briefly outlined in Appendix 1.7, together with selected technical details on the examples.

Example 1: Single-equation Endogenous Binary Response Model

The illustrative example of a single-equation endogenous binary response model is from Chesher (2010). Consider a probit model where the discrete explanatory variable X is possibly correlated with an unobserved U and an instrument Z, which is independent of U, is available.[10] Such a model is in general not point identified.

Suppose that the set of assumptions that define our model is the following:

- $Y = 1(U > t(X))$ (1.10)
- $U \perp\!\!\!\perp Z$ – the unobserved U is independent of the instrument Z
- $U \sim Unif(0,1)$ – U is uniformly distributed on $[0,1]$ interval
- $t(X) = \Phi(-\theta_0 - \theta_1 X)$ – the threshold-crossing function is assumed to take a particular form, where $\Phi(.)$ is a cumulative distribution function of the standard normal distribution.[11]

An interesting question we may ask is the following. Given that we have perfect information on the distribution of the observables, what can we say about the function $t(X)$, or equivalently, about the coefficient $\theta = (\theta_0, \theta_1)$, from our economic model?

Discretization of Unobservables

The discretization as explained in Section 1.2.3 in this case boils down to the discretization employed by Galichon and Henry (2011) in the original GH setup. This is because the additional assumption $E(\phi(U)) = 0$ is not present. We demonstrate this for illustrative purposes.

Suppose that $\theta_1 > 0$. Then, the only subsets of (Y, X) that are compatible with (1.10) are $\{(0,0),(0,1)\}$ for $U \leq t(1)$, $\{(0,0),(1,1)\}$ for $t(1) < U \leq t(0)$ and $\{(1,0),(1,1)\}$ for $U < t(0)$. We assign to these three sets of Us three points (u_1^*, u_2^*, u_3^*) with probabilities $(t(1), t(0) - t(1), 1 - t(0))$. A similar procedure applies for $\theta_1 < 0$. In Figure 1.2, we can see the case for $\theta_1 > 0$ on the left-hand side and for $\theta_1 < 0$ on the right-hand

[10] In the case when X is continuous, the parameter is point identified and could be obtained by e.g. STATA's `ivprobit`.

[11] It is possible to determine the lower and upper bound of the threshold-crossing function $t(X)$ without making this parametric assumption as in Chesher (2009), but instead assuming the monotonicity of $t(X)$. For the sake of simplicity, we present the parametric example.

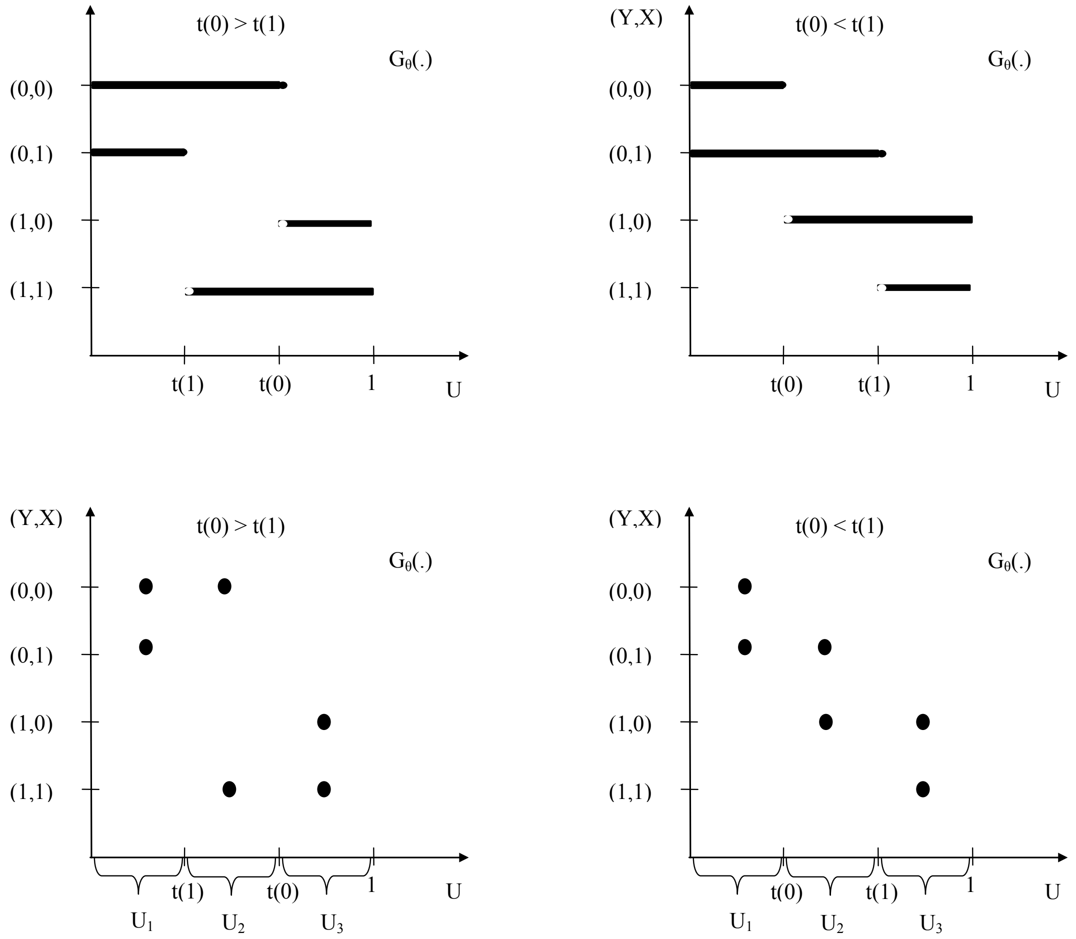


Figure 1.2: Discretization of unobservables in example 1.3. The top left-hand-side panel is for $\theta_1 > 0$, while the top right-hand-side panel is for $\theta_1 < 0$. The discretized counterpart is shown immediately below the original continuous formulation of G_θ in each instance.

side. The upper panes depict the original support restriction G_θ and the lower panes depict their discrete counterparts.

Formulation in the Extended GH Framework

The distribution of observables (Y, X, Z) is assumed known and is denoted p_{ijk}, and U is assumed to be uniformly distributed $[0, 1]$.[12]

For a given (θ_0, θ_1), the aim is to find the joint probability π_{ijkl} of (Y, X, Z, U) that is compatible with the support restrictions and the distributional restrictions, where the marginals of π_{ijkl} are p_{ijk} and ν_l, respectively, and Z and U are independent.

[12] We could also assume that we observe the probability of Y, X given Z, such that for the sake of exposition, the probability of (Y, X, Z) is known.

We define the support restrictions as follows:

$$c_{ijkl} = 1(y_i \neq 1(u_l > t(x_j))) = \begin{cases} 0, \ y_i = 1(u_l > t(x_j)), \\ 1, \ \text{otherwise.} \end{cases} \tag{1.11}$$

Therefore, basically, (Y, X, Z, U)s are restricted to those that satisfy (1.10).

We now convert the formulation of the problem into the extended GH framework:

$$\begin{aligned} &\min_{(\pi)} \textstyle\sum_{i,j,k,l} \pi_{ijkl} c_{ijkl} \\ &\qquad \text{s.t.} \\ &\textstyle\sum_l \pi_{ijkl} = p_{ijk}, && \forall i, j, k \\ &\textstyle\sum_{i,j,k} \pi_{ijkl} = \nu_l, && \forall l \\ &\textstyle\sum_{i,j} \pi_{ijkl} = \sum_{i,j} p_{ijk} \nu_l, && \forall k, l \\ &\pi_{ijkl} \geq 0, && \forall i, j, k, l. \end{aligned} \tag{1.12}$$

If for a given (θ_0, θ_1) the optimum is achieved at 0, this (θ_0, θ_1) is added into the identified set.[13] [14]

Results (binary X)

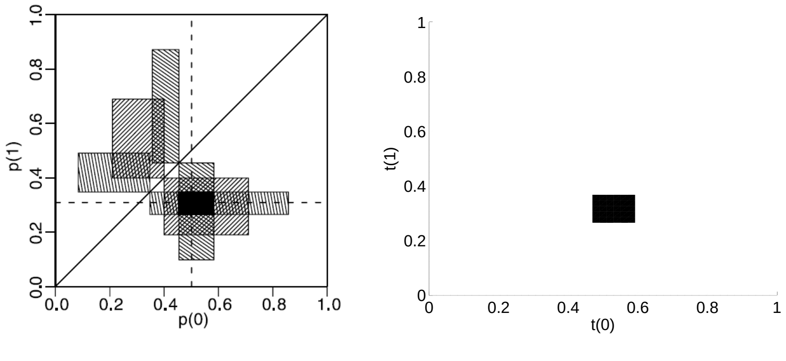


Figure 1.3: Identified set obtained by Chesher's approach (Chesher, 2010) is compared with our solution.

[13] In this case, parameter θ affects the support restrictions (1.10) only.

[14] Note that even though π is four dimensional, the problem still lies within the linear programming framework, as the elements of π can be stacked to make a vector of size $n_Y \cdot n_X \cdot n_Z \cdot n_U$.

The identified set is expressed in terms of the threshold-crossing function at 0 and 1, $t(0)$ and $t(1)$, rather than in the parameter space.[15] Figure 1.3 illustrates that the extended GH setup works for instruments in the case of a binary endogenous variable.

Results (continuous X discretized)

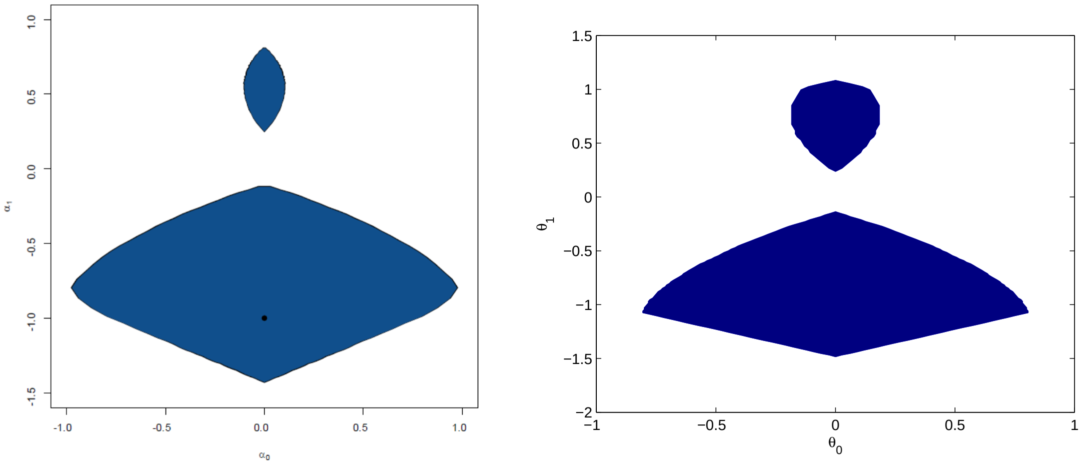


Figure 1.4: Chesher's result Chesher (2009) (Figure 8, p. 37) for problem (1.10) with parameters given by (1.23) compared with the result obtained by the extended GH approach.

Figure 1.4 compares the results obtained by Chesher (2009) and the extended GH framework. Note that even though the shapes of the identified sets are similar, they differ. We need to develop methods for the discrete approximation of continuous observed variables in order to obtain reliable results.

Identifying the Power of the Independency Restriction

We can now consider the identifying strength of the independency condition itself.

Figure 1.5 shows the strength of the independency restriction.[16] It is clear that this extra information shrinks the identified region. It is also worth noting that even if the instruments are entirely endogenous, we exclude some parameter values from the

[15] In order to avoid confusion with the probabilities p_{ijk} of the observed variables, the threshold-crossing function is denoted $t(.)$ unlike in Chesher (2009), who set it as $p(.)$.

[16] The meaning with the second-last restriction is omitted: $\sum_{i,j} \pi_{ijkl} = \sum_{i,j} p_{ijk}\nu_l \;\; \forall k, l$.

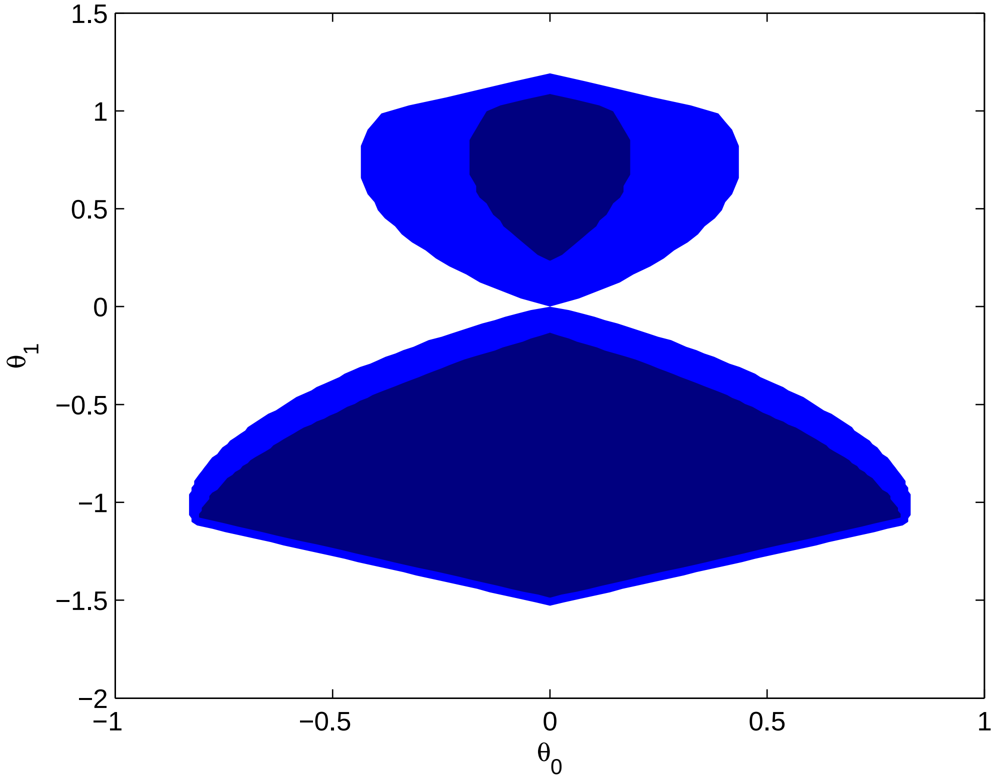


Figure 1.5: Dark blue – with independency restriction, light blue – without assuming independency.

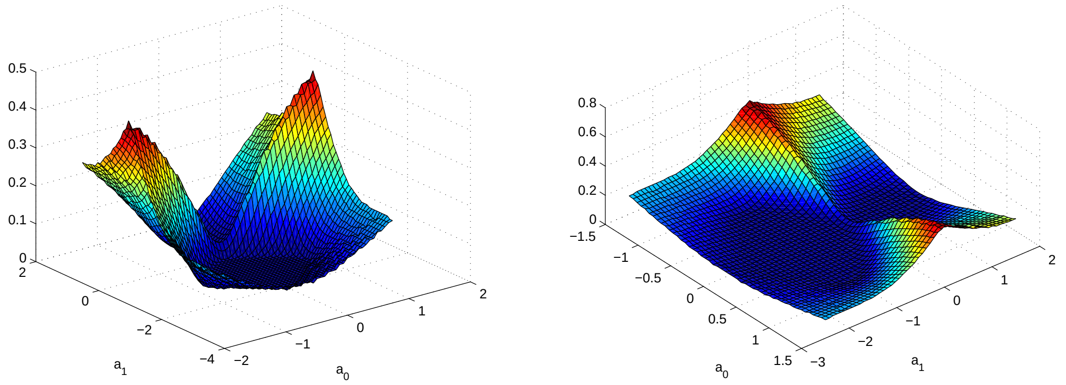


Figure 1.6: Minimized objective function.

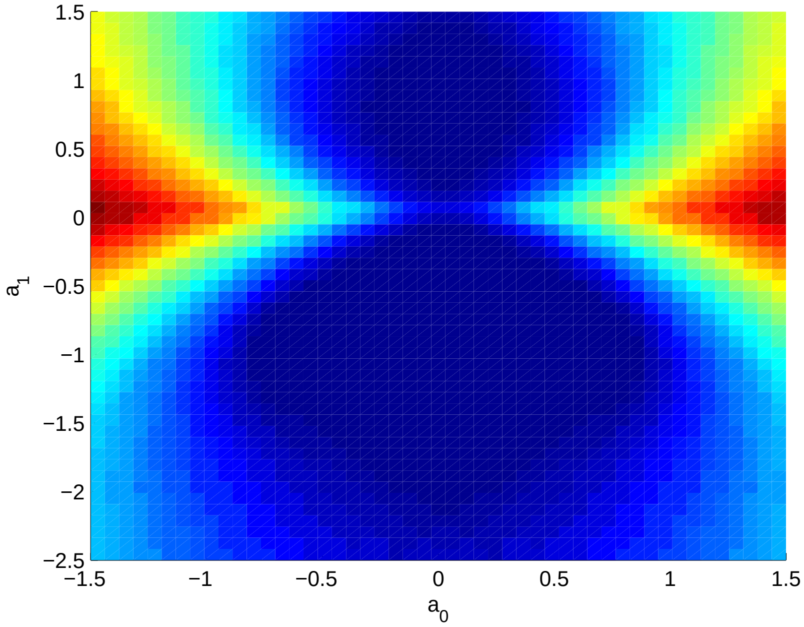


Figure 1.7: Contours of the minimized objective function.

identified set. For these, no joint probability π_{ijkl} of observables and unobservables exists that is compatible with the data-generating process p_{ijk} and with ν_l.

Objective Function

In Figures 1.6 and 1.7, the minimized objective function and its contours are shown.

The zeros of this function correspond to the identified set. However, the values outside the identified set also have an interesting interpretation in that they represent the smallest probability of an event incompatible with the economic model. If, for instance, for a certain parameter value the minimized value of the objective function is 0.2, this means that for any data-generating process, at least 20% of the pairs of observed and unobserved variables violate the support restrictions.[17] This may serve as an appealing measure of misspecification with respect to the support restrictions.

Example 2: Triangular System of Equations with Binary Dependent Variables

Following Shaikh and Vytlacil (2011), the object of interest is the Average Treatment Effect (ATE) in the triangular system of equations.

The collection of assumptions is as follows:

- $Y = 1(\alpha D + \beta X - \epsilon_1 \geq 0),$ (1.13)
- $D = 1(\delta Z - \epsilon_2 \geq 0),$ (1.14)
- $(X, Z) \perp\!\!\!\perp (\epsilon_1, \epsilon_2),$

where Y is a binary outcome variable, D is a treatment identifier, X is an exogenous covariate, and Z is an instrument. Note that no parametric distributional assumptions on (ϵ_1, ϵ_2) are made.

Formulation in the Extended GH Framework

We have four observed variables (Y, X, D, Z) with probabilities p_{ijkl}, and two unobserved variables (ϵ_1, ϵ_2). The discretization of unobserved (ϵ_1, ϵ_2) into (u_1, u_2) is according to Lemma 2. Let us denote $\pi_{ijklmn} = Pr(Y = y_i, X = x_j, D = d_k, Z = z_l, \epsilon_1 = u^1_m, \epsilon_2 = u^2_n)$. The penalty on the points of support not compatible with the economic restrictions G is given by:

$$c_{ijklmn} = \begin{cases} 0, \ (y_i, x_j, d_k, z_l, u^1_m, u^2_n) : y_i = 1(\alpha d_k - u^1_m \geq 0) \text{ and } d_k = 1(\delta z_l - u^2_n \geq 0), \\ 1, \ \text{otherwise.} \end{cases}$$

A particular value of $ATE = \theta$ is compatible with the list of assumptions and with data (p_{ijkl}) if and only if zero is the optimal solution of the following optimization

[17] From Lemma 2, we can see that this interpretation is unaffected by the discretization of the unobserved variables.

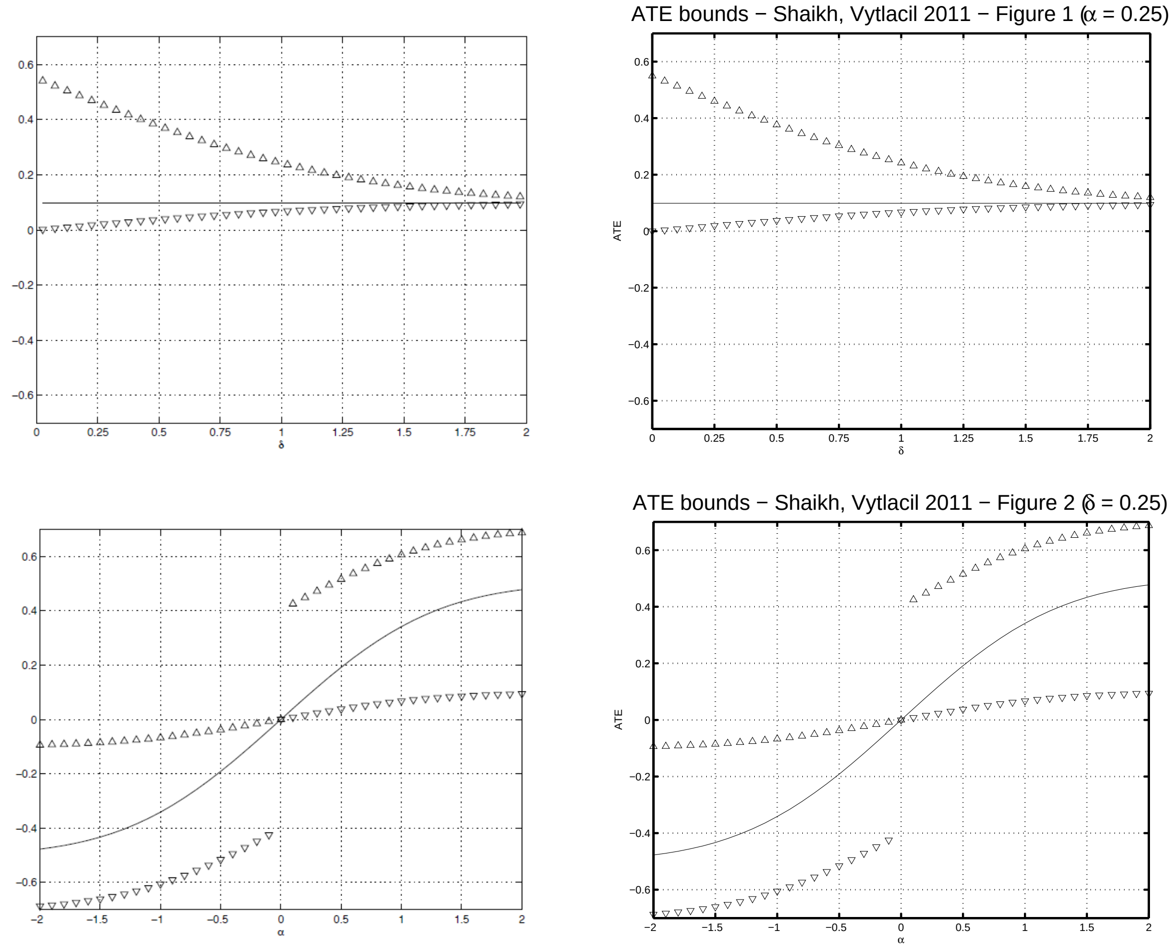


Figure 1.8: The bounds on the ATE are compared using the Shaikh and Vytlacil (2011) approach (left-hand side) and the extended GH framework (right-hand side), with X fixed ($X = 0$) and α fixed ($\alpha = 0.25$, upper pane) or δ fixed ($\delta = 0.25$, lower pane).

problem:

$$\min_{(\pi)} \sum_{i,j,k,l,m,n} \pi_{ijklmn} c_{ijklmn}$$

$$\text{s.t.}$$

$$\sum_{m,n} \pi_{ijklmn} = p_{ijkl}, \qquad \forall i,j,k,l$$

$$\sum_{i,k} \pi_{ijklmn} = \sum_{i,k} p_{ijkl} \sum_{i,j,k,l} \pi_{ijklmn}, \qquad \forall k,l,m,n$$

$$\sum_m \left[1(\alpha \geq u_m^1) - 1(0 \geq u_m^1)\right] \sum_{i,j,k,l,n} \pi_{ijklmn} = \theta,$$

$$\pi_{ijklmn} \geq 0, \qquad \forall i,j,k,l,m,n.$$

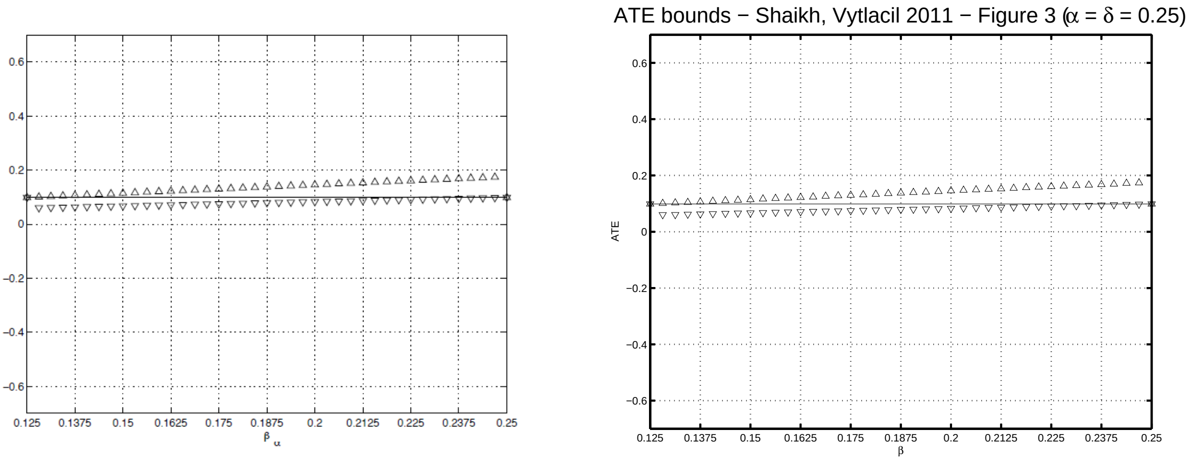


Figure 1.9: Bounds on ATE are compared using the Shaikh and Vytlacil (2011) approach (left) and extended GH framework (right-hand side), with variation in X ($supp(X) = \{-2, -1, 0, 1, 2\}$) and $\alpha = \beta = 0.25$ fixed.

Results

Figures 1.8 and 1.9 compare the results of Shaikh and Vytlacil (2011) with the extended GH framework.

Example 3: Bounds on Treatment Effects with Imperfect Compliance

The following subsection shows how the extended GH framework can determine sharp bounds on the average causal effect when imperfect compliance is present. This was done in the celebrated works of Balke and Pearl (1997, 1994), and this section replicates their results.

Consider three types of observed variables: $Y \in \{y_0, y_1\}$ is an outcome variable where y_0 is for a positive observed response, $D \in \{d_0, d_1\}$ is whether treatment was received (d_1) or not (d_0), and $Z \in \{z_0, z_1\}$ is whether treatment was offered (z_1) or not (z_0). We also assume the existence of an unobserved U that captures individual characteristics affecting the receipt of treatment and the outcome variable. The quantity of interest is the average causal effect of D on Y, denoted as:

$$ACE(D \to Y) = Pr(Y = y_1 | D = d_1) - Pr(Y = y_1 | D = d_0). \tag{1.15}$$

Restrictions that are imposed:

- $Z \perp\!\!\!\perp Y|\{D, U\}$, treatment assignment only affects the outcome variable through actual treatment D.
- $Z \perp\!\!\!\perp U$, Z and U are independent, randomization of the treatment assignments Z may deliver this property.
- no interactions between individuals or Stable Unit Treatment Value Assumption (known as the SUTVA Assumption (Rubin, 1974)).

Formulation in the Extended GH Framework

Following the notation of Balke and Pearl (1994), the unobserved type U of an individual is decomposed into two response function variables $R_D \in \{0, 1, 2, 3\}$ and $R_Y \in \{0, 1, 2, 3\}$. Pair (R_Y, R_D) is now the unobserved type (U) of the individual. Treatment D is a deterministic function of Z and R_D:

$$D = f_D(Z, R_D)$$

, where

$$\begin{aligned} f_D(z, 0) = d_0 \ \ , \ \ f_D(z, 1) \quad &= \quad \begin{cases} d_0, & \text{if } z = z_0, \\ d_1, & \text{if } z = z_1, \end{cases} \\ f_D(z, 2) = d_1 \ \ , \ \ f_D(z, 3) \quad &= \quad \begin{cases} d_1, & \text{if } z = z_0, \\ d_0, & \text{if } z = z_1. \end{cases} \end{aligned}$$

Similarly, the outcome Y is a deterministic function of D and R_Y:

$$D = f_Y(D, r_Y)$$

, where

$$\begin{aligned} f_Y(d, 0) = y_0 \ \ , \ \ f_Y(d, 1) \quad &= \quad \begin{cases} y_0, & \text{if } d = d_0, \\ y_1, & \text{if } d = d_1, \end{cases} \\ f_Y(d, 2) = y_1 \ \ , \ \ f_Y(d, 3) \quad &= \quad \begin{cases} y_1, & \text{if } d = d_0, \\ y_0, & \text{if } d = d_1. \end{cases} \end{aligned}$$

This is basically a discretization of the unobserved component U into the discrete (R_Y, R_D).

The quantity of interest is the Average Causal Effect $\theta = ACE(D \to Y) = Pr(R_Y = 1) - Pr(R_Y = 3)$. We would like to find sharp bounds on θ given $Pr(Y, D, Z)$.[18]. We denote the probability of observed variables $p_{ijk} = Pr(Y = y_i, D = d_j, Z = z_j)$. There are 5 variables in the model: observed Y, D, Z and unobserved R_Y, R_D. The mapping G between unobserved variables and observed variables is defined as

$$G(R_Y, R_D) = \{(Y, D, Z) : f_D(Z, R_D) = D, f_Y(D, R_Y) = Y\}.$$

Now, we denote the joint probability distribution of observed and unobserved variables as $\pi_{ijklm} = Pr(Y = y_i, D = d_j, Z = z_k, R_Y = l, R_D = m)$.

The penalty on the points of support not compatible with G is given by:

$$c_{ijklm} = \begin{cases} 0, \ (y_i, d_j, z_k) \in G(l, m), \\ 1, \ \text{otherwise.} \end{cases}$$

Finally, parameter θ is included in the identified set if and only if the optimized value of the following problem is equal to zero:

$$\begin{aligned} \min_{(\pi)} \textstyle\sum_{ijklm} \pi_{ijklm} c_{ijklm} & \\ \text{s.t.} & \\ \textstyle\sum_{lm} \pi_{ijklm} = p_{ijk}, & \quad \forall i, j, k \\ \pi_{ijklm} \textstyle\sum_{ik} \pi_{ijklm} = \sum_i \pi_{ijklm} \sum_k \pi_{ijklm}, & \quad \forall i, j, k, l, m \\ \textstyle\sum_{ij} \pi_{ijklm} = \sum_{ij} p_{ijk} \sum_{ijk} \pi_{ijklm}, & \quad \forall i, j, k, \\ \textstyle\sum_{ijkm} \pi_{ijk1m} - \sum_{ijkm} \pi_{ijk3m} = \theta, & \\ \pi_{ijklm} \geq 0, & \quad \forall i, j. \end{aligned}$$

The first restriction states that the π_{ijklm} has to be compatible with p_{ijk}, which is observed from the data. The second equality states that when fixing D, R_Y, R_D

[18] $ACE(D \to Y) = Pr(Y = y_1 | D = d_1) - Pr(Y = y_1 | D = d_0) = Pr(R_Y = 1) + Pr(R_Y = 2) - (Pr(R_Y = 2) + Pr(R_Y = 3)) = Pr(R_Y = 1) - Pr(R_Y = 3)$

(equivalent to fixing D, U), Z is independent of Y.[19] The third equation ensures that Z is marginally independent of (R_Y, R_D), whereas the fourth restricts the space of joint distributions to those that have $ACE(D \rightarrow Y)$ equal to θ.

Note that the second restriction is quadratic, so the whole problem is not a linear program. Quadratic restrictions may give rise to the use of semidefinite programming routines.

Although the nonlinear constraint causes significant computational difficulties, results in Balke and Pearl (1997) can be replicated to a reasonable degree of precision (10^{-4}).

Example 4: Binary Choice Model with Zero-median Restriction

This subsection aims to capture the identification setup of the binary choice model with discrete explanatory variables within the extended GH framework. Identification for this type of problem has been studied extensively in recent work by Komarova (2013). It is well known that if all explanatory variables in a binary choice model are discrete, the parameters of the model are in general set rather than point identified. An identification strategy was outlined earlier (Manski and Thompson, 1986), and in Komarova (2013) a computationally attractive recursive procedure is outlined that determines sharp bounds on the identified set.

The problem that is studied takes the following form:

- $Y = 1(X\beta + U \geq 0)$ (1.16)
- $Pr(U \leq 0|X = x) = 0.5 \quad \forall x \in \mathcal{X}$ (1.17) ,

where Y is the outcome variable, X is a k-dimensional random variable with discrete support $\mathcal{X}$, β is a k-dimensional parameter of interest, and U is an unobservable scalar vector variable. The only distributional assumption about U that is made is that the median of U is zero conditional on X.

Discretization of Unobservables

The observed variables X are exogenous in this setup, so the analysis is done conditional on a particular x. The identified set for β will therefore be an intersection of

[19] Instrument Z only affects Y via D: $Pr(Y|D, Z, R_Y, R_D) = Pr(Y|D, R_Y, R_D)$, and this equation can be reformulated as $Pr(Y, D, Z, r_Y, r_D)Pr(D, R_Y, R_D) = Pr(Y, D, r_Y, r_D)Pr(D, Z, R_Y, R_D)$.

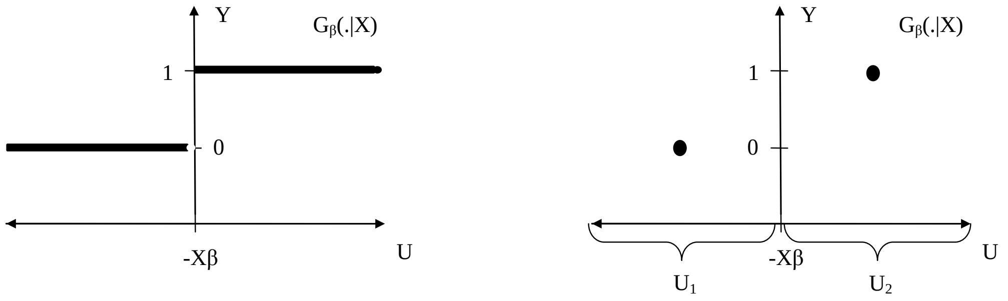


Figure 1.10: The figure on the left-hand side depicts the support restrictions and the figure on the right-hand side is a result of naive discretization.

bounds created by conditioning on all values of X that have nonzero probability.[20] The only restriction put on the unobservable variable U is the zero-median restriction, which has to be taken into account when finding a suitable discretization of U. Naive discretization is presented in Figure 1.10 and does not allow the unobservables to meet the conditional zero-median condition. When the discretization is done by virtue of Lemma 2 so that further distributional restrictions are taken into account as shown in Figure 1.11, the discretization is sufficiently rich to allow us to formulate the conditional zero-median condition. Note that Lemma 2 proves that this discretization leaves the identified set unaffected.

Formulation in the Extended GH Framework

Let $X = x$ be fixed and $p_i = Pr(Y = y_i|X = x)$, where $y_1 = 0$ and $y_2 = 1$. A penalty c_{ij}:

$$c_{ij} = \begin{cases} 0, \text{ if } y_i = 1(x\beta + u_j \geq 0), \\ 1, \text{ otherwise,} \end{cases}$$

carries the information on support restrictions.

The problem can now be formulated as:

[20] As with exogenous instruments, the marginal distribution of X does not have any identifying power.

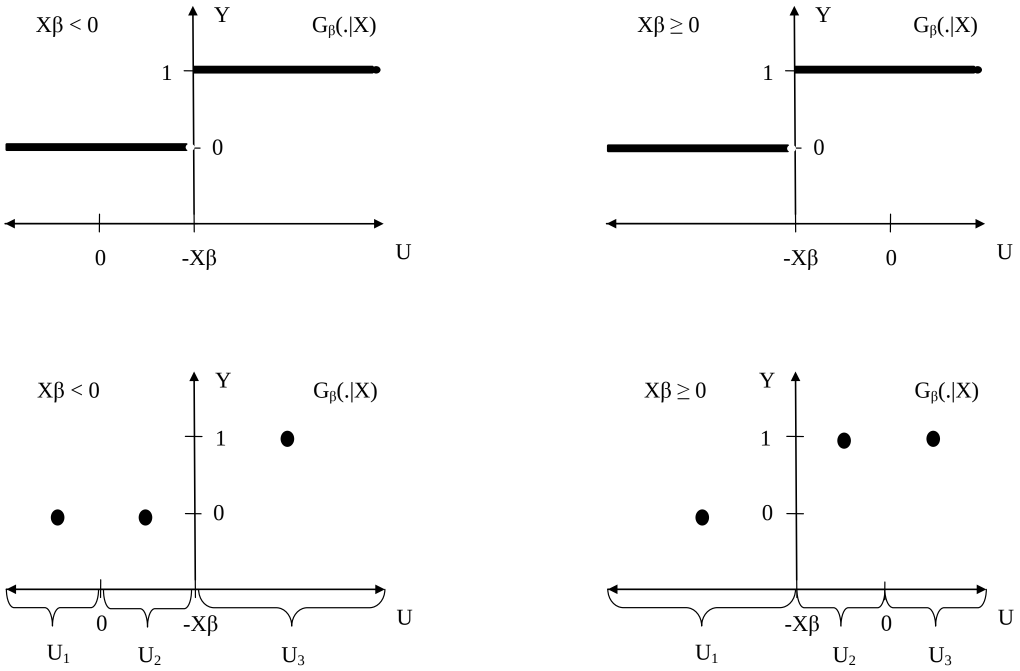


Figure 1.11: The two panes on the left (right)-hand side represent suitable discretization when $X\beta < 0$ ($X\beta \geq 0$). The discretization was obtained using Lemma 1.

$$
\begin{aligned}
&\min_{(\pi)} \textstyle\sum_{i,j} \pi_{ij} c_{ij} \\
&\text{s.t.} \\
&\textstyle\sum_j \pi_{ij} = p_i, \qquad \forall i \\
&\textstyle\sum_i \pi_{i1} = \sum_i \pi_{i2} + \sum_i \pi_{i3}, \\
&\pi_{ij} \geq 0, \qquad \forall i, j,
\end{aligned}
$$

whenever $X\beta < 0$ and:

$$
\begin{aligned}
&\min_{(\pi)} \textstyle\sum_{i,j} \pi_{ij} c_{ij} \\
&\text{s.t.} \\
&\textstyle\sum_j \pi_{ij} = p_i, \qquad \forall i \\
&\textstyle\sum_i \pi_{i1} + \sum_i \pi_{i2} = \sum_i \pi_{i3}, \\
&\pi_{ij} \geq 0, \qquad \forall i, j
\end{aligned}
$$

when $X\beta \geq 0$. The first set of equalities states that the joint distribution π is compatible with observed data p_i, while the second equality restricts U to have zero median.[21] As in previous examples, parameter β is included in the identified set if the optimized value of the problem is equal to 0.

To simplify the notation, all probabilities are implicitly conditioned on $X = x$. If $X\beta < 0$, one can immediately see that $Pr(U = u_3) = Pr(Y = 1) = p_2$ and $Pr(U = u_1) + Pr(U = u_2) = Pr(Y = 0) = p_1$. Therefore, $Pr(U = u_1) = Pr(U = u_2) + Pr(U = u_3) = 0.5$ implies that a proper distribution on U exists if and only if $Pr(U = u_3) = Pr(Y = 1) < 0.5$. On the other hand, if $X\beta \geq 0$, then $Pr(U = u_2) + Pr(U = u_3) = Pr(Y = 1) = p_2$ and $Pr(U = u_1) = Pr(Y = 0) = p_1$ together with $Pr(U = u_1) + Pr(U = u_2) = Pr(U = u_3) = 0.5$ imply that $Pr(U = u_3) = Pr(Y = 1) < 0.5$, so we obtain precisely the same result as (1.25).

This example is simple, but shows how we can easily approach identification in a systematic manner.

1.4 Imperfect Instruments in a Single-equation Endogenous Binary Response Model

As opposed to the previous section, this section demonstrates how the extended GH framework can work in a problem not studied before. As shown in the example with imperfect instruments, we seek to find how the flexibility of adding extra distributional constraints can help us access this problem. The extension plays a crucial role in that we cannot apply the original GH framework.

Identification based on instrumental variables has become the workhorse of applied research, given that we are unable to test the exogeneity of instruments in the just-identified case. It is then of great interest to know the identifying power of this assumption. This information can serve as a sensitivity analysis, such that when relaxing this assumption we can see how the identified set expands. If the identified set becomes substantially larger when exogeneity is only slightly relaxed, we should focus more attention on discussion of this assumption. One may then need to defend

[21] If $X\beta < 0$ equation (1.17) is equivalent to $Pr(U = u_1|X = x) = Pr(U = u_2|X = x) + Pr(U = u_3|X = x)$, and if $X\beta \geq 0$ equation (1.17) can be rewritten as $Pr(U = u_1|X = x) + Pr(U = u_2|X = x) = Pr(U = u_3|X = x)$. Note that this restriction can be rewritten as $\sum_{i,j} \pi_{ij}(1(U \leq 0) - 0.5) = 0$.

the assumption of exogeneity very well for the results to be credible. Conversely, if the exogeneity of instruments is shown not to have great identifying power, the analysis could be said to be robust to some departures from the exogeneity.

Different approaches have been employed in the literature to address the issue of imperfect instruments. For example, Conley, Hansen, and Rossi (2012) parameterize the amount of instrument endogeneity and derive the identified set in the linear regression model. Elsewhere, Hahn and Hausman (2005), rather than deriving the identified set, compare the properties of the ordinary and two-stage least squares estimators, while Manski and Pepper (2000) make use of the monotonicity of the instrumental variables instead of an exogeneity assumption. Lastly, Nevo and Rosen (2012) derive sharp bounds on the parameters under the assumption that the correlation between the instrument and the error term has the same sign as the correlation between the endogenous regressor and the error term, and that the instrument is assumed to be less correlated with the error term than the endogenous regressor.

The example of the single-equation endogenous binary response model from Section 1.3 demonstrates how we can use the extended GH setup to trace the identified set if the strict exogeneity condition is relaxed. The way this assumption is relaxed is as follows: under the strict exogeneity restriction, $Pr(Z) \cdot Pr(U) = Pr(Z \cap U)$ for all pairs (Z, U). The distribution $Pr(Z) \cdot Pr(U)$ can be represented as a point in the $n_Z \times n_U$-dimensional unit simplex. Instead of restricting $Pr(Z \cap U)$ to be exactly equal to $Pr(Z) \cdot Pr(U)$, we will assume that the difference $Pr(Z \cap U) - Pr(Z) \cdot Pr(U)$ has to be less or equal to $\alpha Pr(Z) \cdot Pr(U)$ in absolute value for some fixed $\alpha > 0$ and all (Z, U). The parameter α hence controls the amount of endogeneity in the instruments. We can model the departure from exogeneity in many ways. However, we select this somewhat ad hoc way of relaxing strict exogeneity so that the problem still lies within the linear programming framework and so that discretization is possible.

The model under the study is (1.10) with support restrictions (1.11). In addition, we assume the instruments are not strictly exogenous. We formulate the problem within the extended GH framework in the following way:

$$
\begin{aligned}
&\min_{(\pi)} \textstyle\sum_{i,j,k,l} \pi_{ijkl} c_{ijkl} && \\
&\text{s.t.} && \\
&\textstyle\sum_l \pi_{ijkl} = p_{ijk}, && \forall i,j,k \\
&\textstyle\sum_{i,j,k} \pi_{ijkl} = \nu_l, && \forall l \\
&\textstyle\sum_{i,j} \pi_{ijkl} - \sum_{i,j} p_{ijk}\nu_l \leq \alpha \sum_{i,j} p_{ijk}\nu_l, && \forall k,l \\
&-\textstyle\sum_{i,j} \pi_{ijkl} + \sum_{i,j} p_{ijk}\nu_l \leq \alpha \sum_{i,j} p_{ijk}\nu_l, && \forall k,l \\
&\pi_{ijkl} \geq 0, && \forall i,j,k,l.
\end{aligned}
\tag{1.18}
$$

As in (1.3), we generate the probabilities of the observed variables according to (1.21), with Z having support on $\{-0.75, 0, 0.75\}$ with probabilities $\left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$.

Results

The results of this illustration are presented in Figures 1.12 and 1.13. We can see how the identified set becomes larger as the departure from strict exogeneity increases. As the identified set with the weak instrument is larger than the identified set with the stronger instrument, it is less sensitive to the violation of the strict exogeneity assumption because it is much closer to the identified set with the instrument that is completely endogenous.

1.5 Conclusion

In this paper, we proposed a new method to obtain the identified set as a simple extension of the GH identification strategy so that a broader class of problems can be solved. A considerable advantage of this new method is its algorithmic structure, such that we need not derive the sharp bounds of the identified set from case to case, but rather efficient algorithms can be employed to trace the identified set independently of the structure of the problem. Some existing identification results were replicated in a straightforward manner. Moreover, the new method allowed us to consider the im-

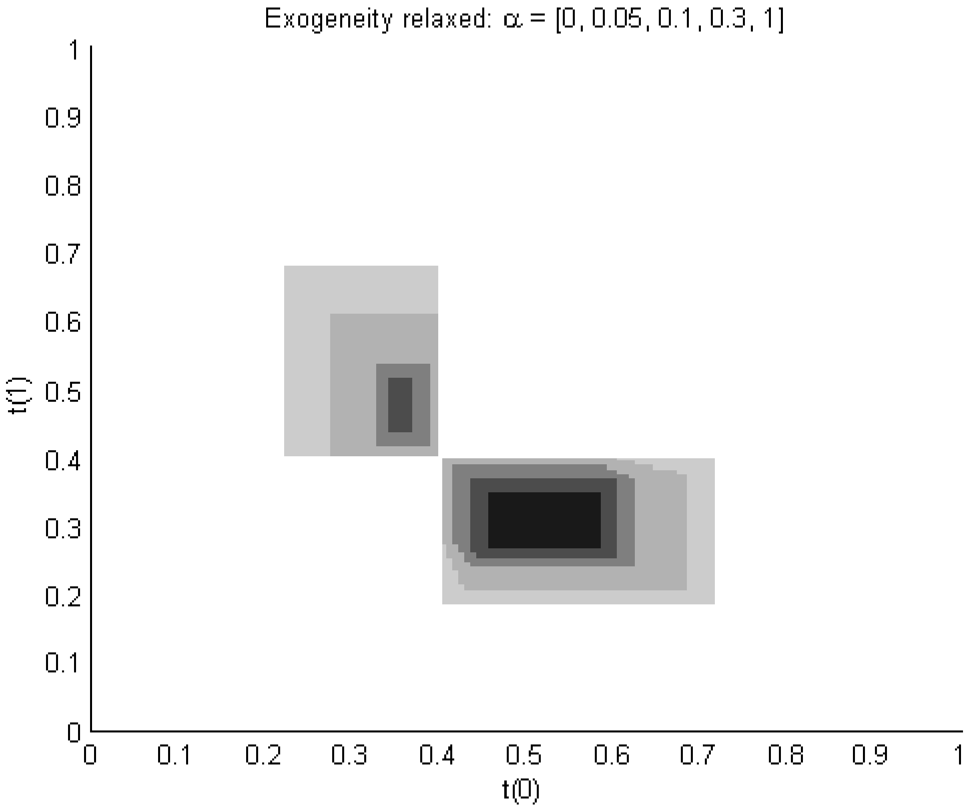


Figure 1.12: Identified sets corresponding to different values of the parameter α; the case with a strong instrument. The darker-shaded areas indicate stronger exogeneity. Note that the observed probabilities together with the assumption of uniform U and support restrictions given by the economic model do have some identifying power, even if the instrument is completely endogenous.

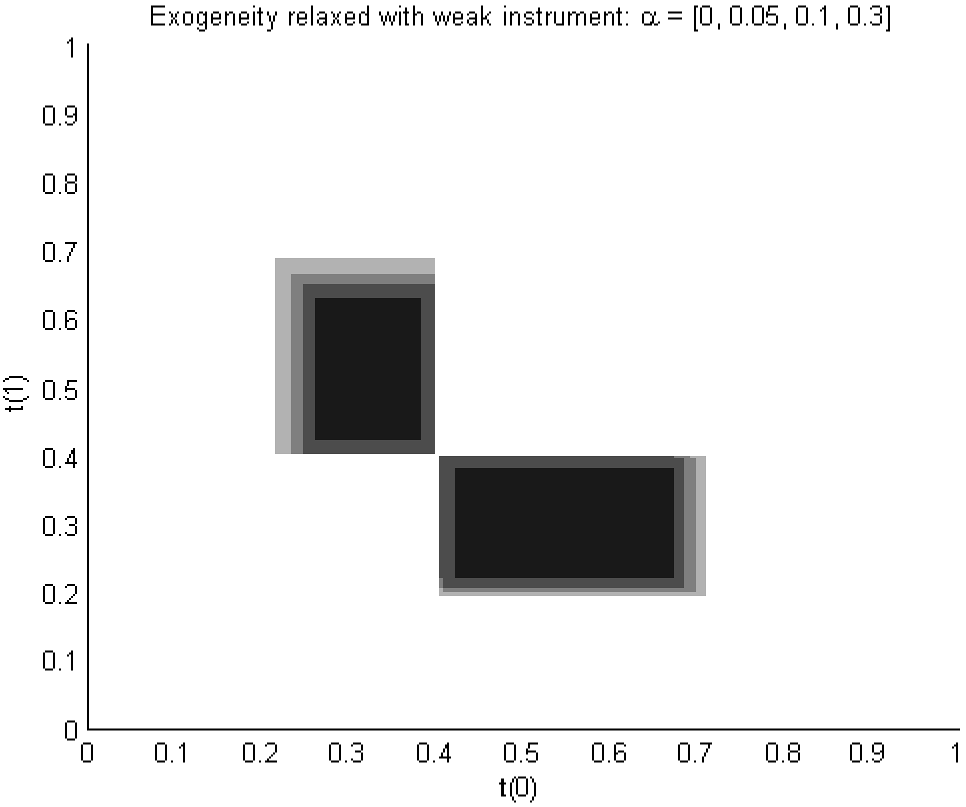


Figure 1.13: Identified sets corresponding to different values of the parameter α; the case with a weak instrument. The darker-shaded areas indicate stronger exogeneity.

pact of relaxing the assumption of strict exogeneity in nonlinear models with discrete variables.

The main finding is that if the observed variables are discrete, identification can be attacked at its lowest level by searching in the space of the joint distribution functions for the observed and unobserved variables. This delivers greater flexibility when studying the identifying power of different sets of assumptions. How to make this method operational in a continuous case, e.g. as an analog of condition (1.4), and how to undertake statistical inference remain open questions. The iterative subsampling scheme in Romano and Shaikh (2010) and the intersection bounds in Chernozhukov et al. (2013) appear to be useful steps forward. Further research is warranted.

APPENDIX

1.6 Proofs

1.6.1 Proof of Lemma 1

Proof. We need to show that there exists $\pi_1 : \mathcal{Y} \times \mathcal{U} \mapsto [0,1]$ satisfying:

$$\begin{aligned}
& & \sum_{i=1}^{n} \int_{u \in \mathcal{U}} \pi_1(y_i, u) 1(y_i \in G(u)) du = 1, \qquad & \text{(C1)} \\
\forall i = 1, ..., n : & & \int_{u \in \mathcal{U}} \pi_1(y_i, u) du = p_i, \qquad & \text{(C2)} \\
\forall u \in \mathcal{U} : & & \sum_{i=1}^{n} \pi_1(y_i, u) = \nu(u), \qquad & \text{(C3)} \\
\forall I \in \mathbb{I}; \forall u \in \mathcal{U} : & & \left| \sum_{i \in I} \pi_1(y_i, u) - \sum_{i \in I} p_i \nu(u) \right| \leq \alpha \sum_{i \in I} p_i \nu(u). \qquad & \text{(C4)} \\
\forall i = 1, ..., n; \forall u \in \mathcal{U} : & & \pi_1(y_i, u) \geq 0 \qquad & \text{(C5)}
\end{aligned}$$

if and only if there exists $\pi_2 : \mathcal{Y} \times \mathcal{U}^* \mapsto [0,1]$ satisfying:

$$\sum_{i=1}^{n}\sum_{j=1}^{m}\pi_2(y_i,u_j^*)1(y_i\in G(u_j^*))du=1, \qquad \text{(D1)}$$

$$\forall i=1,...,n: \qquad \sum_{j=1}^{m}\pi_2(y_i,u_j^*)=p_i, \qquad \text{(D2)}$$

$$\forall j=1,...,m: \qquad \sum_{i=1}^{n}\pi_2(y_i,u_j^*)=\nu^*(u_j^*), \qquad \text{(D3)}$$

$$\forall I\in\mathbb{I};\forall j=1,...,m: \qquad \left|\sum_{i\in I}\pi_2(y_i,u_j^*)-\sum_{i\in I}p_i\nu^*(u_j^*)\right|\leq\alpha\sum_{i\in I}p_i\nu^*(u_j^*). \qquad \text{(D4)}$$

$$\forall i=1,...,n;\forall j=1,...,m: \qquad \pi_2(y_i,u_j^*)\geq 0 \qquad \text{(D5)}$$

"$(\Rightarrow)$" - Given π_1, we construct π_2 according to:

$$\forall i=1,...,n;\forall j=1,...,m: \qquad \pi_2(y_i,u_j^*)=\int\limits_{\Delta_j}\pi_1(y_i,u)du, \qquad (\Pi_2)$$

and this will ensure that {(C1),(C2),(C3M),(C4M),(C5)} implies {(D1),(D2),(D3M),(D4M),(D5)} as shown below:

$$\sum_{i=1}^{n}\sum_{j=1}^{m}\pi_2(y_i,u_j^*)1(y_i\in G(u_j^*))\overset{(\Pi_2)}{=}\sum_{i=1}^{n}\sum_{j=1}^{m}\int\limits_{\Delta_j}\pi_1(y_i,u)du1(y_i\in G(u_j^*))\overset{\text{(PartU1)}}{=}$$

$$\overset{\text{(PartU1)}}{=}\sum_{i=1}^{n}\sum_{j=1}^{m}\int\limits_{\Delta_j}\pi_1(y_i,u)1(y_i\in G(u))du=\sum_{i=1}^{n}\int\limits_{u\in\mathcal{U}}\pi_1(y_i,u)1(y_i\in G(u))du\overset{\text{(C1)}}{=}1,$$

$$\forall i: \qquad \sum_{j=1}^{m}\pi_2(y_i,u_j^*)\overset{(\Pi_2)}{=}\sum_{j=1}^{m}\int\limits_{\Delta_j}\pi_1(y_i,u)du=\int\limits_{u\in\mathcal{U}}\pi_1(y_i,u)du\overset{\text{(C2)}}{=}p_i,$$

$$\forall j: \qquad \sum_{i=1}^{n}\pi_2(y_i,u_j^*)\overset{(\Pi_2)}{=}\sum_{i=1}^{n}\int\limits_{\Delta_j}\pi_1(y_i,u)du=\int\limits_{\Delta_j}\sum_{i=1}^{n}\pi_1(y_i,u)du\overset{\text{(C3)}}{=}\int\limits_{\Delta_j}\nu(u)du\overset{\text{(P)}}{=}\nu^*(u_j^*),$$

$$\forall j,\forall I: \qquad \left|\sum_{i\in I}\pi_2(y_i,u_j^*)-\sum_{i\in I}p_i\nu^*(u_j^*)\right|\overset{(\Pi_2),\text{(P)}}{=}\left|\sum_{i\in I}\int\limits_{\Delta_j}\pi_1(y_i,u)du-\sum_{i\in I}p_i\int\limits_{\Delta_j}\nu(u)du\right|=$$

$$=\left|\int\limits_{\Delta_j}\left(\sum_{i\in I}\pi_1(y_i,u)-\sum_{i\in I}p_i\nu(u)\right)du\right|\overset{\text{(C4)}}{\leq}\left|\int\limits_{\Delta_j}\alpha\sum_{i\in I}p_i\nu(u)du\right|=\alpha\sum_{i\in I}p_i\nu^*(u_j^*).$$

$$\forall i, \forall j: \qquad \pi_2(y_i, u_j^*) \overset{(\Pi_2)}{=} \int\limits_{\Delta_j} \pi_1(y_i, u)du \overset{(C5)}{\geq} \int\limits_{\Delta_j} 0du = 0.$$

"($\Leftarrow$)" - If we know π_2, we obtain π_1 using:

$$\forall i = 1, ..., n; \forall j = 1, ..., m; \forall u \in \Delta_j: \quad \pi_1(y_i, u) = \pi_2(y_i, u_j^*)\frac{\nu(u)}{\nu^*(u_j^*)}, \qquad (\Pi_1)$$

(note that (Π_1) implies (Π_2)) and we now show that {(D1),(D2),(D3M),(D4M),(D5)} implies {(C1),(C2),(C3M),(C4M),(C5)}:

$$\begin{aligned}
&\sum_{i=1}^{n} \int\limits_{u \in \mathcal{U}} \pi_1(y_i, u)1(y_i \in G(u))du = \sum_{i=1}^{n}\sum_{j=1}^{m} \int\limits_{\Delta_j} \pi_1(y_i, u)1(y_i \in G(u))du \overset{\text{(PartU1)}}{=} \\
&\overset{\text{(PartU1)}}{=} \sum_{i=1}^{n}\sum_{j=1}^{m} \int\limits_{\Delta_j} \pi_1(y_i, u)du1(y_i \in G(u_j^*)) \overset{(\Pi_1)}{=} \sum_{i=1}^{n}\sum_{j=1}^{m} \pi_2(y_i, u_j^*)1(y_i \in G(u_j^*)) \overset{\text{(D1)}}{=} 1,
\end{aligned}$$

$$\forall i: \qquad \int\limits_{u \in \mathcal{U}} \pi_1(y_i, u)du = \sum_{j=1}^{m} \int\limits_{\Delta_j} \pi_1(y_i, u)du \overset{(\Pi_1)}{=} \sum_{j=1}^{m} \pi_2(y_i, u_j^*) \overset{\text{(D2)}}{=} p_i,$$

$$\forall j, \qquad \forall u \in \Delta_j: \sum_{i=1}^{n} \pi_1(y_i, u) \overset{(\Pi_1)}{=} \sum_{i=1}^{n} \pi_2(y_i, u_j^*)\frac{\nu(u)}{\nu^*(u_j^*)} \overset{\text{(D3)}}{=} \nu(u),$$

$$\begin{aligned}
\forall j, \forall I, \quad \forall u \in \Delta_j: &\left|\sum_{i \in I} \pi_1(y_i, u) - \sum_{i \in I} p_i \nu(u)\right| \overset{(\Pi_1)}{=} \left|\sum_{i \in I} \pi_2(y_i, u_j^*)\frac{\nu(u)}{\nu^*(u_j^*)} - \sum_{i \in I} p_i \frac{\nu(u)}{\nu^*(u_j^*)}\nu^*(u_j^*)\right| = \\
&= \left|\frac{\nu(u)}{\nu^*(u_j^*)}\left(\sum_{i \in I} \pi_2(y_i, u_j^*) - \sum_{i \in I} p_i \nu_j^*\right)\right| \overset{\text{(D4)}}{\leq} \left|\alpha \sum_{i \in I} p_i \nu(u)\right| = \alpha \sum_{i \in I} p_i \nu(u),
\end{aligned}$$

$$\forall i, \forall j, \qquad \forall u \in \Delta_j: \pi_1(y_i, u_j) \overset{(\Pi_1)}{=} \pi_2(y_i, u_j^*)\frac{\nu(u)}{\nu^*(u_j^*)} \overset{\text{(D5)}}{\geq} 0.$$

□

1.6.2 Proof of Lemma 2

Proof. Similarly to the proof of Lemma 1, we need to show that there exists $\pi_1 : \mathcal{Y} \times \mathcal{U} \mapsto [0, 1]$ satisfying (C1),(C2),(C5) and:

$$\sum_{i=1}^{n}\int_{u\in\mathcal{U}}\pi_1(y_i,u)\phi(u)du=0, \qquad \text{(C3M)}$$

$$\forall I\in\mathbb{I};\forall u\in\mathcal{U}: \quad \left|\sum_{i\in I}\pi_1(y_i,u)-\sum_{i\in I}p_i\sum_{i=1}^{n}\pi_1(y_i,u)\right|\leq\alpha\sum_{i\in I}p_i\sum_{i=1}^{n}\pi_1(y_i,u) \qquad \text{(C4M)}$$

if and only if there exists $\pi_2:\mathcal{Y}\times\mathcal{U}^*\mapsto[0,1]$ satisfying (D1),(D2),(D5) and:

$$\sum_{i=1}^{n}\sum_{j=1}^{m}\pi_2(y_i,u_j^*)\phi(u_j^*)du=0, \qquad \text{(D3M)}$$

$$\forall I\in\mathbb{I};\forall j=1,...,m: \quad \left|\sum_{i\in I}\pi_2(y_i,u_j^*)-\sum_{i\in I}p_i\sum_{i=1}^{n}\pi_2(y_i,u_j^*)\right|\leq$$
$$\leq\alpha\sum_{i\in I}p_i\sum_{i=1}^{n}\pi_2(y_i,u_j^*). \qquad \text{(D4M)}$$

"$(\Rightarrow)$" - Given π_1, we construct π_2 according to:

$$\forall i=1,...,n;\forall j=1,...,m: \quad \pi_2(y_i,u_j^*)=\int_{\Delta_j}\pi_1(y_i,u)du, \qquad (\Pi_2)$$

and this will ensure that {(C1),(C2),(C3M),(C4M),(C5)} imply {(D1),(D2),(D3M), (D4M),(D5)}. Because the partitioning of the $\mathcal{U}$ space using (PartU2) is finer than that using (PartU1), we find that {(C1),(C2),(C5)}, implying {(D1),(D2),(D5)} immediately using the proof of Lemma 1. It is therefore sufficient to show that {(C3M),(C4M)} imply {(D3M),(D4M)}:

$$\sum_{i=1}^{n}\sum_{j=1}^{m}\pi_2(y_i,u_j^*)\phi(u_j^*)\overset{(\Pi_2)}{=}\sum_{i=1}^{n}\sum_{j=1}^{m}\int_{\Delta_j}\pi_1(y_i,u)du\;\phi(u_j^*)=$$
$$=\sum_{i=1}^{n}\sum_{j=1}^{m}\int_{\Delta_j}\pi_1(y_i,u)\phi(u^*)du\overset{\text{(PartU2)}}{=}\sum_{i=1}^{n}\sum_{j=1}^{m}\int_{\Delta_j}\pi_1(y_i,u)\phi(u)du=$$
$$=\sum_{i=1}^{n}\int_{u\in\mathcal{U}}\pi_1(y_i,u)\phi(u)du\overset{\text{(C3M)}}{=}0,$$

$$\forall I \in \mathbb{I}; \forall j = 1, ..., m : \qquad \left| \sum_{i \in I} \pi_2(y_i, u_j^*) - \sum_{i \in I} p_i \sum_{i=1}^{n} \pi_2(y_i, u_j^*) \right| \overset{(\Pi_2)}{=}$$

$$\overset{(\Pi_2)}{=} \left| \sum_{i \in I} \int\limits_{\Delta_j} \pi_1(y_i, u) du - \sum_{i \in I} p_i \sum_{i=1}^{n} \int\limits_{\Delta_j} \pi_1(y_i, u) du \right| =$$

$$= \left| \int\limits_{\Delta_j} \left(\sum_{i \in I} \pi_1(y_i, u) - \sum_{i \in I} p_i \sum_{i=1}^{n} \pi_1(y_i, u) \right) du \right| \overset{\text{(C4M)},(\Pi_2)}{\leq}$$

$$\overset{\text{(C4M)},(\Pi_2)}{\leq} \left| \alpha \sum_{i \in I} p_i \sum_{i=1}^{n} \pi_2(y_i, u_j^*) \right| = \alpha \sum_{i \in I} p_i \sum_{i=1}^{n} \pi_2(y_i, u_j^*).$$

"$(\Leftarrow)$" - Knowing π_2, we obtain π_1 using:

$$\forall i = 1, ..., n; \forall j = 1, ..., m; \forall u \in \Delta_j : \quad \pi_1(y_i, u) = \pi_2(y_i, u_j^*) \frac{\gamma(u)}{\int\limits_{u \in \Delta_j} \gamma(u) du}, \qquad (\Pi_1)$$

where γ is an arbitrary strictly positive probability density function. It is now sufficient to show that {(D3M),(D4M) (D5)} imply {(C3M),(C4M),(C5)}, because the proof of Lemma 1 reveals that {(C1),(C2)} imply {(D1),(D2)} and (PartU2) provides a finer discretization of $\mathcal{U}$ than does (PartU1):

$$\sum_{i=1}^{n} \int\limits_{u \in \mathcal{U}} \pi_1(y_i, u) \phi(u) du = \sum_{i=1}^{n} \sum_{j=1}^{m} \int\limits_{\Delta_j} \pi_1(y_i, u) \phi(u) du \overset{\text{(PartU2)}}{=}$$

$$\overset{\text{(PartU2)}}{=} \sum_{i=1}^{n} \sum_{j=1}^{m} \int\limits_{\Delta_j} \pi_1(y_i, u) \phi(u_j^*) du = \sum_{i=1}^{n} \sum_{j=1}^{m} \int\limits_{\Delta_j} \pi_1(y_i, u) du \ \phi(u_j^*) \overset{(\Pi_1)}{=}$$

$$\overset{(\Pi_1)}{=} \sum_{i=1}^{n} \sum_{j=1}^{m} \pi_2(y_i, u_j^*) \phi(u_j^*) \overset{\text{(D3M)}}{=} 0,$$

$$\begin{aligned}
\forall j, \forall I, \qquad & \forall u \in \Delta_j : \left| \sum_{i \in I} \pi_1(y_i, u) - \sum_{i \in I} p_i \sum_{i=1}^{n} \pi_1(y_i, u) \right| \overset{(\Pi_1)}{=} \\
& \overset{(\Pi_1)}{=} \left| \sum_{i \in I} \pi_2(y_i, u_j^*) \frac{\gamma(u)}{\int\limits_{u \in \Delta_j} \gamma(u) du} - \sum_{i \in I} p_i \sum_{i=1}^{n} \pi_2(y_i, u_j^*) \frac{\gamma(u)}{\int\limits_{u \in \Delta_j} \gamma(u) du} \right| = \\
& = \left| \frac{\gamma(u)}{\int\limits_{u \in \Delta_j} \gamma(u) du} \left(\sum_{i \in I} \pi_2(y_i, u_j^*) - \sum_{i \in I} p_i \sum_{i=1}^{n} \pi_2(y_i, u_j^*) \right) \right| \overset{(D4),(\Pi_1)}{\leq} \\
& \overset{(D4),(\Pi_1)}{\leq} \left| \alpha \sum_{i \in I} p_i \sum_{i=1}^{n} \pi_1(y_i, u) \right| = \alpha \sum_{i \in I} p_i \sum_{i=1}^{n} \pi_1(y_i, u), \\
\forall i, \forall j, \qquad & \forall u \in \Delta_j : \pi_1(y_i, u) \overset{(\Pi_1)}{=} \pi_2(y_i, u_j^*) \frac{\gamma(u)}{\int\limits_{u \in \Delta_j} \gamma(u) du} \overset{(D5)}{\geq} 0.
\end{aligned}$$

□

1.7 Technical Details on the Presented Examples

1.7.1 Example 1

Chesher's Approach

In order to present the identification result from Chesher (2009), we first introduce the basic definitions. The notation used differs from that in GH that is employed in the present study.

- A **model** $\mathcal{M}$ is defined as (1.10) with $U \sim Unif(0,1)$ and $U \perp\!\!\!\perp Z$ for all $Z \in \mathcal{Z}$.
- A **structure** $S \equiv \{t, F_{UX|Z}\}$ is a pair of a threshold-crossing function t and a cumulative distribution function of the conditional distribution of U and X given Z.
- A structure S is said to be **admitted by a model** $\mathcal{M}$ if $F_{UX|Z}$ respects the independence property, that is $F_U(u|z) \equiv F_{UX|Z}(u, \bar{x}|z) = u$ for all $u \in (0,1)$ and all $z \in \mathcal{Z}$, where $\bar{x}$ is the upper bound of X.

- A structure S **generates** the joint distribution of Y and X given Z if $F_{YX|Z}(0,x|z) = F_{UX|Z}(t(x),x|z)$.
- Two structures $S^* \equiv \{t^*, F^*_{UX|Z}\}$ and $S^0 \equiv \{t^0, F^0_{UX|Z}\}$ are said to be **observationally equivalent** if they generate the same distribution of Y and X given Z for all $z \in \mathcal{Z}$, that is if $F^*_{YX|Z}(0,x|z) \equiv F^*_{UX|Z}(t^*(x),x|z) = F^0_{YX|Z}(0,x|z) \equiv F^0_{UX|Z}(t^0(x),x|z)$ for all $z \in \mathcal{Z}$ and for all $x \in \mathcal{X}$.

Theorem 1 from Chesher (2009) states that having a structure S_0 admitted by the model $\mathcal{M}$ that generates the conditional distribution of Y and X given Z with cumulative distribution function $F^0_{YX|Z}$ and if this threshold-crossing function t is in structure S admitted by model $\mathcal{M}$ that is observationally equivalent to S^0, then t satisfies:

$$\begin{aligned} c_{0l}(u,z;p) &= Pr_0[Y=0 \cap t(X) < u|Z=z] < u, \forall u \in (0,1),\ \forall z \in \mathcal{Z} \qquad (1.19)\\ c_{0u}(u,z;p) &= 1 - Pr_0[Y=1 \cap u \leq t(X)|Z=z] \geq u, \forall u \in (0,1),\ \forall z \in \mathcal{Z}, (1.20) \end{aligned}$$

where Pr_0 states that probabilities were calculated using the measure that was generated by S^0, that is using $F^0_{YX|Z}$ and l and u stand for the lower and upper bound, respectively.

Given the continuity of X, the converse is also true. This is equivalent to saying that the set of all functions p satisfying the above set of inequalities is a *sharply defined identified set*. In Chesher (2010), this theorem is proven, even for a more general setup. It is important to note that the proof is constructive, so that for a given threshold-crossing function t, a suitable distribution function $F_{UX|Z}$ is constructed such that $\{t, F_{UX|Z}\}$ is admitted by the model $\mathcal{M}$ and generates the $F_{YX|Z}$ observed in the data. This highlights the link to the GH setup, as the aim there is to find the joint probability distribution that satisfies the independence restriction, has correct marginals, and places all the probability on those combinations of variables that are compatible with the data.

Illustration: Discrete Endogenous Variable

Construction of True Data-generating Process

The following example is taken from Chesher (2010). Suppose that both Y and X are binary; $Y \equiv 1(Y^* \geq 0)$ and $X \equiv 1(X^* \geq 0)$, where Y^* and X^* were generated in the following way:

$$Y^* = \theta_0 + \theta_1 X + W, \quad X^* = b_0 + b_1 Z + V$$
$$\begin{bmatrix} W \\ V \end{bmatrix} \perp\!\!\!\perp Z, \quad \begin{bmatrix} W \\ V \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & r \\ r & 1 \end{bmatrix} \right) \tag{1.21}$$

with parameters:

$$(\theta_0, \theta_1, b_0, b_1, r) = (0, 0.5, 0, 1, -0.25). \tag{1.22}$$

and the instrument Z takes values in $\mathcal{Z} = \{-0.75, 1, 0.75\}$.

However, the econometrician does not know how the data were generated. She only assumes (1.10) and $U \perp\!\!\!\perp Z$, $U \sim Unif(0,1)$, $t(X) = \Phi(-\theta_0 - \theta_1 X)$, and observes the distribution of the observable variables p_{ijk}.[22] Even though it is impossible to recover the true value of $\theta = (0, 0.5)$ exactly, it is possible to at least create informative bounds for it.

As the X threshold-crossing function t attains only two values, $t(0) = \Phi(-\theta_0) = 0.5$ and $t(1) = \Phi(-\theta_0 - \theta_1) = 0.308$.

Illustration: Continuous Endogenous Variable

Construction of the True Data-generating Process

Suppose that the economic model is described by (1.10) and the data-generating process by (1.21) with the parameters:

$$(\theta_0, \theta_1, b_0, b_1, s_{wv}, s_{vv}) = (0, -1, 0, 0.3, 0.5, 1). \tag{1.23}$$

as before, the only difference being that X is no longer binary ($X = X^*$).

The distribution of the observable variables $(Y^*, X|Z = z)$ (Y^* and X given $Z = z$) is given by $N(\mu(z), \Sigma)$, where:

$$\mu(z) = \begin{bmatrix} \theta_0 + \theta_1 b_0 + \theta_1 b_1 z \\ b_0 + b_1 z \end{bmatrix} \quad \Sigma = \begin{bmatrix} 1 + 2\theta_1 s_{wv} + \theta_1^2 s_{vv} & s_{wv} + \theta_1 s_{vv} \\ s + wv + \theta_1 s_{vv} & s_{vv}. \end{bmatrix}$$

[22] The observed probabilities p_{ijk} were obtained using the Matlab function `mvtnorm`.

We provide details of the simulations here. Because of the continuity of X, the unobservable U was discretized as the equidistant point masses on $[0,1]$. The distribution of observables is given by:

$$p_{ijk} = Pr(Y = y_i \cap X = x_j \cap Z = z_k) = Pr(Y = y_i \cap X = x_j | Z = z_k) Pr(Z = z_k).$$

It is known that $(Y^*, X|Z) \sim N(\mu(z), \Sigma)$ and a suitable discretization of X is needed. It is easy to show that the density of $(Y^*|X = x, Z = z)$ is:

$$N\left(\mu(z)_1 + \frac{\Sigma_{21}}{\Sigma_{22}}(x - \mu(z)_2), \left(1 - \sqrt{\frac{\Sigma_{21}^2}{\Sigma_{11}\Sigma_{22}}}\right)\Sigma_{11}\right).$$

Integrating the corresponding probability density function at $(-\infty$,0) gives us $Pr(Y = 0|X = x, Z = z)$. The distribution of X given $Z = z$ is $N(b_0 + b_1 z, s_{vv})$, but now the question is how to discretize the support of X, which is $\mathbb{R}$. If the number of nodes is n_x, then one suggestion would be to set the z to its mean value, that is 0, and set the values of the discretized support of X to n_x equidistant quantiles.[23] Even though this discretization appears natural, it brings some degree of arbitrariness to the problem.

Finally, taking all the pieces together yields:

$$p_{ijk} = Pr(Y = y_i | X = x_j, Z = z_k) Pr(X = x_j | Z = z_k) Pr(Z = z_k),$$

where all quantities on the right-hand side are known.

1.7.2 Example 2

Illustration

True Data-generating Process

For the illustration, (ϵ_1, ϵ_2) are assumed to be $N(0, I_2)$. This assumption, together with (1.13) and (1.14), generates the distribution of Y and D given X and Z. The support of Z is assumed to be $\{-1, 1\}$ and the support of X is either $\{0\}$ or $\{-2, -1, 0, 1, 2\}$. (X, Z) are assumed to be uniformly distributed.[24]

[23] Excluding the 0% and 100% quantiles.

[24] As in Example 1, the distribution of exogenous variables per se does not have any identifying power. It is included purely for the simplicity of the exposition.

1.7.3 Example 3

Balke and Pearl's Approach

Balke and Pearl (1997) made use of the fact that these restrictions impose the following decomposition on the joint distribution of (Y, D, Z, U):

$$Pr(Y, D, Z, U) = Pr(Y|D, U)Pr(D|Z, U)Pr(Z)Pr(U). \tag{1.24}$$

There exist four different functions from Z to D and four different functions from D to Y, hence 16 different *types* of individuals that we can consider. Hence, one can think of U as having a discrete support with 16 points, each point representing a pair of functions, one from Z to D and the second from D to Y. For instance, one type u may be persons who always accept treatment and who do not display a positive outcome irrespective of treatment. The bounds on (1.15) are found using linear program searching through the space of distributions of the types (U) subject to the joint distribution to be compatible with observed data $Pr(y, d|z)$. The full setup, together with discussion, is in Balke and Pearl (1997, 1994).

1.7.4 Example 4

Komarova's Approach

Following Manski and Thompson (1986):

$$Pr(Y = 1|X = x) = 1 - Pr(U < -x\beta|X = x),$$

together with the zero-median restriction (1.17), implies:

$$Pr(Y = 1|X = x) \geq 0.5 \;\Leftrightarrow\; x\beta \geq 0. \tag{1.25}$$

Therefore, the bounds on the parameter vector β are obtained as an intersection of linear half spaces. In Komarova (2013), a recursive procedure is proposed that translates this set of linear inequalities into bounds on the parameters.

1.8 Implementation Issues

1.8.1 Extended GH Framework

The following routines were used and compared in order to solve the linear program (1.12).

- `linprog`[25] – Matlab built-in function from the Optimization Toolbox. We found the interior point method superior to the simplex method because of the computational time involved. As the objective value is not minimized to exact zeros, a certain threshold had to be employed. The natural choice was the tolerance level of the optimization routine (10^{-8} for $n_x = n_u = 40$). The results for the two approaches were identical.
- `GNU Linear Programming Kit (GLPK)` – Modified simplex method from Matlab MEX interface for the GLPK library[26]. We found this to be significantly faster than `linprog`, but with similar results.

The linear program is a long-standing and well-understood problem; however, if the discretization of X and U is large, then the matrix that encodes the restrictions for the joint distribution[27] can reach the limits of the largest array that can be created by Matlab. For instance, if the sizes of the supports are $n_x = n_u = 40$ together with $n_y = 2$ and $n_z = 10$, then the joint probability π_{ijkl} has 32000 elements. The matrix that carries the information about restrictions on π_{ijkl} will then have 32000 columns.

[25] http://www.mathworks.com/help/toolbox/optim/ug/linprog.html
[26] http://glpkmex.sourceforge.net/
[27] this is a four-dimensional array π_{ijkl} stacked into a vector

Chapter 2

INFERENCE IN PARTIALLY IDENTIFIED MODELS WITH DISCRETE VARIABLES

Abstract

This paper compares different ways of conducting statistical inference in models with discrete variables when a scalar parameter of interest is partially identified. A Monte Carlo simulation study compares the finite sample properties of the confidence sets obtained by different methods and leads to a list of practical recommendations.

JEL: C01, C12, C15.

Keywords: Bounds, Average treatment effects, Confidence sets, Simulation study, Bootstrap, Linear programming.

2.1 Introduction and Motivation

This paper studies the finite sample properties of various methods of statistical inference in models with discrete variables, when the scalar parameter of interest is partially identified. In many economically interesting situations, the assumptions of the economic model together with the data do not uniquely determine the parameter but only restrict it to lie in a set.

Consideration of the data and economic assumptions involve the following two steps. First is the question of identification. In an ideal case where perfect knowledge of the distribution of observable variables is assumed, it is important to identify the object of interest, which is the collection of models compatible with the data and identifying assumptions and this set is called the identified set. Second, once the problem of identification is resolved, it is important to study how the data limitations affect the conclusions.

This paper addresses the second issue; that is, how to conduct statistical inference in partially identified models with discrete observable variables for a broad class of economic problems. It is important for practitioners to know the properties of different inferential methods. This paper also attempts to provide some practical advice on how to choose the right method in different situations and discusses its pros and cons. More specifically, this paper:

- shows that a bias correction is helpful for diminishing small sample bias,
- points at the problem of a possible empty identified set in small samples when bootstrapping,
- uses a histogram of vertices corresponding to the optimal solutions of the linear program to detect possible bootstrap failure, and
- presents a method of obtaining a confidence set with guaranteed (yet possibly conservative) asymptotic coverage probability in situations when the regular percentile bootstrap fails.

We use simulation analysis to determine the performance of different inferential schemes in situations likely to be faced by an empirical researcher. Different methods are compared, highlighting their theoretical and practical advantages and disadvantages. Furthermore, the analysis discusses some computationally more costly methods. The practical advice provided in this paper should be considered by empirical researchers working on bounds analysis in models with discrete variables.

This paper also complements the work of Laffers (2013b), where an identification framework was introduced that nests a large class of partially identified problems with discrete observable variables. Developing an inferential scheme was left as an open problem. This paper attempts to fill this gap for the case of a scalar parameter.

There are different ways of expressing the concept of statistical uncertainty in partially identified models. This paper focuses on confidence sets that have a frequentist interpretation. It is assumed that there exists a true parameter that is fixed but unknown to the researcher. Having a random data sample of finite length, the aim is to draw inference on the true parameter. This is different from the Bayesian way of thinking, and Bayesian credible sets and frequentist confidence sets do not even coincide asymptotically (Moon and Schorfheide, 2012). This disagreement is explained in Kitagawa (2012).

Inference in partially identified models is an area of ongoing research, and our analysis contributes to this growing stream of literature. Among the first papers dealing with inference in partially identified setting is Imbens and Manski (2004) for a scalar parameter, which was further extended to account for nuisance parameters by (Stoye, 2009). The difference between the confidence region that includes the whole identified set asymptotically with a fixed probability and the one that covers the true parameter with this probability asymptotically is also discussed, and Imbens and Manski (2004) argued why the latter is preferred. Henry and Onatski (2012) found that policy makers concerned with robust decision making may actually prefer the confidence region that covers the entire identified set with a prescribed probability.

The first study that considered vector-valued parameters is Chernozhukov et al. (2007), which is based on the subsampling of a criterion function, which is a function of the data and of the parameter. Zeros of the criterion function define the identified set, which is the set of parameters that are compatible with the data and with the assumptions that constitute the economic model. Romano and Shaikh (2010) builds upon these results and proposes an iterative scheme that increases the statistical power and does not require an initial consistent estimate of the identified set. Rosen (2008) considers the case in which the identified set is defined by a finite collection of moment inequalities and derives the asymptotic distribution of an associated statistic. Galichon and Henry (2009a) construct the confidence region by inverting the Kolmogorov–Smirnov statistic for Choquet capacity functionals. Beresteanu and Molinari (2008), Beresteanu et al. (2011, 2012) make use of the tools of the Random Sets Theory (Molchanov, 2005). The recent work of Chernozhukov, Lee, and Rosen (2013) considers inference on the bounds that are defined as the infimum or supremum of a parametric or a nonparametric function and is the only method that can handle the continuum of the moment inequalities. Bugni (2010) proposes a corrected bootstrap procedure with favorable finite sample properties for the moment inequalities case. The work of Andrews and Shi (2013) is also concerned with problems defined by conditional moment inequalities, and it shows how these can be transformed into unconditional ones using instruments. The work of Henry et al. (2011) uses a combinatorial bootstrap that significantly elevates the computational cost related to creating the identified set. Performance of different criterion functions that define the identified set for the moment inequalities case is studied in Canay (2010) and Bugni (2010, 2011).

This paper is organized as follows. Section 2.2 describes the setup and the notation used in this paper. Different inferential methods are presented in Section 2.3. Section 2.4 presents the simulation results that compare the performance of the different methods. Section 2.5 presents concluding remarks and some practical recommendations.

2.2 Setup and Notation

Laffers (2013b) builds upon the results of Galichon and Henry (2009a) and discusses how an identified set can be determined using a linear programming technique in a large class of partially identified models. The set of assumptions that define the economic model and the probability distribution of the observed variables together often translate into linear restrictions on the joint probability distribution of the observed and unobserved variables. The joint distribution need not be uniquely determined by these restrictions, and the parameter of interest is only partially identified. If the scalar parameter of interest is a linear function of the joint distribution function, then the identified set corresponds to an interval between the minimum and the maximum of the linear program. The leading example is the identification of the bounds on average treatment effects under different assumptions. The linear programming identification formulation is rich, and many problems can be formulated within this framework (Laffers, 2013b). There are other studies that consider partially identified parameters using linear programming; most notably, Balke and Pearl (1997, 1994), Manski (2007), Honoré and Tamer (2006), Chiburis (2010) and Freyberger and Horowitz (2012). There is also an early literature in stochastic programming that studied a *distribution problem*, where the object of interest was the distribution of the optimal solution of a random linear program (Babbar, 1955; Wagner, 1958; Tintner, 1960; Prekopa, 1966). In order to derive the asymptotic distribution of the optimal solution and of the optimum, these papers make strong assumptions that are not always appropriate in our setup. This paper will look more closely at the method in Freyberger and Horowitz (2012), which is closest to this stream of literature.

The inference schemes considered in this paper are mainly based on the bootstrap (Efron, 1979). Horowitz (2001) presents an expository overview of the topic for econometricians.

Most empirical studies focus on inference for a scalar parameter that is a linear function of the joint probability distribution where the linear programming formulation is convenient. Instead of running a linear program for every single parameter value as in Laffers (2013b), only two optimizations are required, one for the minimum and one for the maximum. Hence our object of interest, the identified set $(L_{\min}, L_{\max})$, is an interval between the minimum and the maximum of the objective function of a random linear program in a standard form:

$$\begin{gathered} \max(\min) c^T \pi \\ s.t. \\ A\pi = b, \\ \pi \geq 0, \end{gathered} \tag{2.1}$$

where c is a fixed vector, and matrix A and vector b are data dependent and need to be estimated.[1] Strongly consistent estimators of A and b are available and are denoted as $\hat{A}_n$ and $\hat{b}_n$ based on a sample size n. $\hat{L}_{\max}$ estimates the optimized objective function $L_{\max}$ by solving the sample analog problem of (2.1):

$$\begin{gathered} \max(\min) c^T \pi \\ s.t. \\ \hat{A}_n \pi = \hat{b}_n, \\ \pi \geq 0. \end{gathered} \tag{2.2}$$

We introduce some further notation.

- Let $p_0 = (p_0^1, p_0^2, \ldots, p_0^k)$ be the true vector of probabilities of an observable discrete random variable X, which can take k different values $\{X^1, \cdots, X^k\}$, let X_n denote a random sample of size n, and $\hat{p}_n = \left(\hat{p}_n^1, \hat{p}_n^2, \ldots, \hat{p}_n^k\right)$ is a sample analog of p_0 based on X_n ($\hat{p}_n^j = \sum_{i=1}^n 1(X_n^i = X^j)/n$).
- Let $\Theta_I(p)$ be the identified set if the probability of observables is p; $\Theta_0 \equiv \Theta_I(p_0)$ is the true identified set and $\hat{\Theta}_n \equiv \Theta_I(\hat{p}_n)$ is the estimator of the identified set Θ_0 based on the i.i.d. data sample X_n. $L_{\min} = \min \Theta_I(p)$, $L_{\max} = \max \Theta_I(p)$, $\hat{L}_{\min} = \min \hat{\Theta}_n$ and $\hat{L}_{\max} = \max \hat{\Theta}_n$.

The goal is to create a confidence region, $\hat{C}_n$, with the following properties.

- Asymptotic $(1-\alpha)$-coverage of the unknown parameter from the identified set: $\inf_{\theta_0 \in \Theta_0} \lim_{n\to\infty} \Pr(\theta_0 \in \hat{C}_n) \geq 1 - \alpha$.
- Convergence at the fastest possible rate.[2]
- It works robustly under different scenarios.
- It is easy to implement.

The data enter the analysis only through the vector of probabilities p.

[1] Note that the identified set is an interval. This was proven in Laffers (2013a), and the reasoning is very similar to that presented in Freyberger and Horowitz (2012); convex combinations of the optimal solutions that correspond to the lower and upper bound trace the whole interval between them and satisfy the linear constraints in (2.1).

[2] e.g., $d_H(\Theta_0, \hat{C}_n) = O_p(n^{-\gamma})$ with γ as large as possible, where $d_H(A,B) = \max\left\{\sup_{a\in A} \inf_{b\in B} d(a,b), \sup_{b\in B} \inf_{a\in A} d(a,b)\right\}$ with $d(.,.)$ being the Euclidean distance.

2.2.1 Example

Consider the problem of the bounding average treatment effect of a mother's schooling on her child's schooling as considered in de Haan (2011) and Laffers (2013d). These are the variables in the model for an individual j.

- An outcome: $y_j \in Y = \{0, 1\}$ - child's college (0 - no college, 1 - college), $y_j(0)$ and $y_j(1)$ are counterfactual outcomes and y_j is the observed outcome.
- A treatment: $z_j \in Z = \{0, 1\}$ - mother's college (0 - no college, 1 - college).
- A monotone instrument: $v_j \in V = \{1, 2, 3, 4\}$ - father's schooling level (high school or less ($\leq$ 12 years), some college (13–15 years), bachelor's degree (16 years), master's degree or higher ($\geq$ 17 years)).

Furthermore, consider these identifying assumptions.

- The *monotone treatment response* (MTR) assumption: $\forall j, t_2 \geq t_1 : y_j(t_2) \geq y_j(t_1)$ ensures that the outcome function for each individual j is weakly increasing in the treatment.
- The *monotone treatment selection* (MTS) assumption: $\forall t, t_2 \geq t_1 : E[y(t)|z = t_2] \geq E[y(t)|z = t_1]$ states that individuals with higher observed treatment have either a greater or equal potential mean outcome.
- The *monotone instrumental variable* (MIV) assumption: $\forall t, v_2 \geq v_1 : E[y(t)|v = v_2] \geq E[y(t)|v = v_1]$ assumes that the mean outcome is weakly increasing in the instrument value.
- The *monotone selection bias* (MSB) assumption: $\forall t, t_2 \geq t_1, v_2 \geq v_1 : E[y(t)|z = t_2, v = v_2] - E[y(t)|z = t_1, v = v_2] \geq E[y(t)|z = t_2, v = v_1] - E[y(t)|z = t_1, v = v_1]$ states that the size of the selection bias is increasing in the value of the instrument value.[3]

Finding the upper bound of the average treatment effect of a mother's schooling on her child's schooling ($E[y(1)] - E[y(0)]$) under the MTR+MTS+MIV assumption is equivalent to solving the linear program shown in Figure 2.1 as presented in Laffers (2013a,b). Observed are frequencies of the 16 different categories of (y, z, v). The linear program can be transformed to the standard form as shown in Figure 2.2.

$$\max_{\pi} \overbrace{\left[0\,1\,0\,1\,0\,1\,0\,1\,0\,0\,0\,0\,0\,0\,0\,0\,1\,0\,1\,0\,1\,0\,1\,0\right] \times \pi}^{\text{Average Treatment Effect}}$$

subject to

$$\text{DATA}\left\{\begin{bmatrix}
1\,1\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0 \\
0\,0\,1\,1\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0 \\
0\,0\,0\,0\,1\,1\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0 \\
0\,0\,0\,0\,0\,0\,1\,1\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0 \\
0\,0\,0\,0\,0\,0\,0\,0\,1\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0 \\
0\,0\,0\,0\,0\,0\,0\,0\,0\,1\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0 \\
0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,1\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0 \\
0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,1\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0 \\
0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,1\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0 \\
0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,1\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0 \\
0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,1\,0\,0\,0\,0\,0\,0\,0\,0\,0 \\
0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,1\,0\,0\,0\,0\,0\,0\,0\,0 \\
0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,1\,1\,0\,0\,0\,0\,0\,0 \\
0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,1\,1\,0\,0\,0\,0 \\
0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,1\,1\,0\,0 \\
0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,1\,1
\end{bmatrix}\right. \times \pi = \left.\begin{bmatrix}
0.397 \\ 0.055 \\ 0.029 \\ 0.017 \\ 0.013 \\ 0.01 \\ 0.013 \\ 0.012 \\ 0.155 \\ 0.055 \\ 0.054 \\ 0.047 \\ 0.017 \\ 0.018 \\ 0.043 \\ 0.065
\end{bmatrix}\right\} \begin{array}{c}\text{Observed}\\ \text{probabilities}\end{array}$$

$$\begin{array}{l}
\text{MTS}\left\{\begin{bmatrix}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & .19 & .19 & .19 & .19 & 0 & -.80 & 0 & -.80 & 0 & -.80 & 0 & -.80 \\
0 & .19 & 0 & .19 & 0 & .19 & 0 & .19 & 0 & 0 & 0 & 0 & .19 & .19 & .19 & .19 & -.80 & -.80 & -.80 & -.80 & -.80 & -.80 & -.80 & -.80
\end{bmatrix}\right. \\
\text{MIV}\left\{\begin{bmatrix}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & .13 & -.58 & 0 & 0 & 0 & .13 & 0 & -.58 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & .13 & -.13 & 0 & 0 & 0 & 0 & .13 & 0 & -.13 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & .14 & -.13 & 0 & 0 & 0 & 0 & 0 & .14 & 0 & -.13 \\
0 & .13 & 0 & -.58 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & .13 & -.58 & 0 & 0 & .13 & .13 & -.58 & -.58 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & .13 & 0 & -.13 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & .13 & -.13 & 0 & 0 & 0 & .13 & .13 & -.13 & -.13 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & .14 & 0 & -.13 & 0 & 0 & 0 & 0 & 0 & 0 & .14 & -.13 & 0 & 0 & 0 & 0 & .14 & .14 & -.13 & -.13
\end{bmatrix}\right.
\end{array} \times \pi \le \begin{array}{c}
\begin{bmatrix} 0 \\ 0 \end{bmatrix} \\
\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}
\end{array}$$

$$\pi \geq \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix},$$

$$\pi^* = \begin{array}{l}
[0.2 \;\; 0.2 \;\; 0.003 \; 0.052 \quad 0 \quad 0.029 \quad 0 \quad 0.017 \; 0.013 \;\; 0.01 \;\; 0.013 \; 0.012 \quad \ldots \\
\ldots \; 0.16 \; 0.055 \; 0.054 \; 0.047 \quad 0 \quad 0.017 \; 0.018 \quad 0 \quad 0.042 \; 0.001 \;\; 0.01 \;\; 0.055]'.
\end{array}$$

Figure 2.1: This linear program searches in the space of the joint probability distributions assigned to all combinations of the observed component (y, z, v) and the unobserved component $(y(0), y(1))$ that are compatible $(\forall i, t : z_i = t \rightarrow y_i = y_i(t))$ and that satisfy the MTR assumption. The space of the joint distributions is further restricted to satisfy the MTS assumption and the MIV assumption, and to be compatible with the observed probabilities. The optimal solution π^* maximizes the average treatment effect.

Original Form	Added Slack Variables	Standard Form
$(\min)\max_\pi c^T\pi$ $s.t$ $A_d\pi = p$ $A_s\pi \leq b_s$ $\pi \geq 0$	$(\min)\max_\pi [c^T 0^T]\begin{bmatrix}\pi \\ v\end{bmatrix}$ $s.t$ $\begin{bmatrix}A_d & 0 \\ A_s & -I\end{bmatrix}\begin{bmatrix}\pi \\ v\end{bmatrix} = \begin{bmatrix}p \\ b_s\end{bmatrix}$ $\begin{bmatrix}\pi \\ v\end{bmatrix} \geq 0$	$(\min)\max_\pi \bar{c}^T\pi$ $s.t$ $A\pi = b$ $\pi \geq 0$

Depends on data Fixed

Figure 2.2: Converting the original linear program to its standard form. Blue elements are fixed, and red elements depend on the data.

The following section will present different inferential methods and will discuss their theoretical and practical advantages and disadvantages. This list of methods is not exhaustive, and some computationally more expensive methods are not included in the simulation study. Some of them will be discussed briefly.

2.3 Methods for Statistical Inference

2.3.1 Freyberger and Horowitz

This method is presented in Freyberger and Horowitz (2012) and is specially designed for random linear programs of the form (2.1). The bound is estimated by the sample analog, solving (2.2), and then we determine the vertex (a basic solution) of the set of feasible solutions that corresponds to the optimal solution.[4] We create a large number of bootstrapped pseudo-samples and obtain values of the objective function at the (previously) optimal vertex. The coordinates of this vertex have asymptotically multivariate normal distribution. If there is only one optimal solution, as the sample size increases, the probability that the correct vertex is chosen approaches one, and the optimized objective value is normally distributed by the virtue of the Cramer–

[3] This assumption was not considered in de Haan (2011).

[4] The basic solution is a nonzero subvector, and the corresponding nonzero variables are called basic variables.

Wold theorem. If there are multiple optimal solutions, the error goes in the direction of overcoverage.

The full procedure is as follows.

1. Generate a bootstrap sample of data with replacement, and let A^* and b^* denote the estimates of A and b based on the bootstrap sample.

2. Let $\hat{k}_{\max}$ ($\hat{k}_{\min}$) be the index of the optimal basic solution of the problem (2.2), and define $\Delta^*_{1\hat{k}_{\max}} = n^{1/2}(c^T_{\hat{k}_{\max}} \hat{A}^{*-1}_{\hat{k}_{\max}})\hat{b}^*$ ($\Delta^*_{2\hat{k}_{\min}} = n^{1/2}(c^T_{\hat{k}_{\min}} \hat{A}^{*-1}_{\hat{k}_{\min}})\hat{b}^*$), where the subscript indicates the columns corresponding to the basic variables.

3. Repeat steps (1) and (2) many times, to obtain the empirical distribution of $\Delta^*_{1\hat{k}_{\max}}$ ($\Delta^*_{2\hat{k}_{\min}}$). Find $c^*_{\alpha,\max}$ ($c^*_{\alpha,\min}$) that satisfy $P^*(\Delta^*_{1\hat{k}_{\max}} \geq -c^*_{\alpha,\max}) = 1 - \alpha$ (for lower bound $P^*(-\Delta^*_{2\hat{k}_{\min}} \leq c^*_{\alpha,\max}) = 1 - \alpha$).

4. The upper (lower) bound of the confidence set is $\hat{L}_{\max} + n^{-1/2}c^*_{\alpha,\max}$ ($\hat{L}_{\min} - n^{-1/2}c^*_{\alpha,\min}$).

When it comes to practical implementation, it is essential to find the columns that correspond to the basic variables; i.e., those that have nonzero elements of the optimal solution of (2.1). When using linear programming software, as a numerical artefact, the nonbasic variables will not be exactly zero, so it is necessary to set a threshold value to determine the zeros. This can be set to some number smaller than the precision of the optimization routine used to solve the linear program. Furthermore, especially in small samples, the matrix $\hat{A}^*_{\hat{k}_{\max}}$ need not be invertible; however, pseudo-inversion can be used.[5] This may happen if: (1) a simulated probability vector p has some mass point of zero probability or (2) if an optimal solution of the linear program is degenerate. The inversion or the pseudo-inversion is the most costly step in the computation. Freyberger and Horowitz (2012) makes an assumption that rules out degenerate optimal solutions of the linear program, which implies that the $A_{k_{\max}}$ matrix is invertible. This is not an appropriate assumption in the context of our problem. If economic assumptions that translate into the linear restrictions in our linear program do not bind, they may give rise to the degenerate optimal solutions. However, these nonbinding constraints are not irrelevant. They do not bind for our vector

[5] The Moore–Penrose pseudo-inverse of a real matrix A is a matrix A_+ that satisfies $AA_+A = A$, $A_+AA_+ = A_+$

of probabilities p but may bind for some other vector $\bar{p}$, which can easily occur in the bootstrap simulations. The distribution of the optimal vertex of the simplex of feasible solutions changes with increases in the sample size. If the optimal solution is unique, the larger the sample size, the smaller the number of different optimal vertices and the more concentrated the distribution.

2.3.2 Percentile Bootstrap

The (approximate) percentile bootstrap confidence interval was proposed in Efron (1979). The procedure is as follows.

1. Generate a bootstrap sample of data with replacement, and let A^* and b^* denote the estimates of A and b based on the bootstrap sample.

2. Calculate the optimal value of (2.2) for the bootstrap sample, and denote it as $L^*_{\max}$ ($L^*_{\min}$).

3. Repeat steps (1) and (2) many times, to obtain the empirical distribution of $L^*_{\max}$ ($L^*_{\min}$).

4. The upper (lower) bound of the confidence set is the $(1-\alpha)$-quantile of the distribution of $L^*_{\max}$, so that the number c solves $P^*(L^*_{\max} \leq c_{\max}) = 1-\alpha$ (or α-quantile of the distribution of $L^*_{\min}$, so that the number $c_{\min}$ solves $P^*(L^*_{\min} \leq c_{\min}) = \alpha$).

The percentile bootstrap is straightforward to implement and is transformation respecting.[6] However, it is not justified if the parameter of interest is on the boundary of the parameter space (Andrews, 2000) or if the estimator is a nonsmooth function of the parameter Horowitz (2001). This is often the case if the estimator is a result of minimization or maximization of a discontinuous function, and asymptotic theory based on a Taylor series expansion does not apply.

[6] The transformation-respecting property means that the percentile confidence interval for any monotone transformation $m(\theta)$ of a parameter of interest θ is the transformed percentile interval for θ.

2.3.3 Normalized and Centered Percentile Bootstrap

Horowitz (2001) recommends that the bootstrap should be used on asymptotically pivotal statistics; therefore, this method normalizes and centers the bounds before calculating the quantiles. This was also used in Bugni (2010) in the context of models defined by moment inequalities. The procedure is described below.

1. Generate a bootstrap sample of data with replacement and let A^* and b^* denote the estimates of A and b based on the bootstrap sample.

2. Calculate the optimal value of (2.2) for the bootstrap sample, and denote it as $L^*_{\max}$ ($L^*_{\min}$).

3. Repeat steps (1) and (2) many times, to obtain the empirical distribution of $L^*_{\max}$ ($L^*_{\min}$). Find $c^*_{\alpha,\max}$ ($c^*_{\alpha,\min}$) that satisfy $P^*(n^{1/2}(L^*_{\max} - \hat{L}_{\max}) \geq -c^*_{\alpha,\max}) = 1 - \alpha$ ($P^*(n^{1/2}(L^*_{\min} - \hat{L}_{\min}) \leq c^*_{\alpha,\max}) = 1 - \alpha$).

4. The upper (lower) bound of the confidence set is $\hat{L}_{\max} + n^{-1/2}c^*_{\alpha,\max}$ ($\hat{L}_{\min} - n^{-1/2}c^*_{\alpha,\min}$).

2.3.4 Bias Corrected Percentile Bootstrap

Although consistent, the percentile bootstrap may give misleading results in finite samples. The bias correction in the context of the percentile bootstrap intervals was presented in Efron (1981). The heuristic bias adjustment in the context of bounds analysis was proposed by Kreider and Pepper (2007), and its finite sample properties were studied in Manski and Pepper (2009).

1. Generate a bootstrap sample of data with replacement, and let A^* and b^* denote the estimates of A and b based on the bootstrap sample.

2. Calculate the optimal value of (2.2) for the bootstrap sample, and denote it as $L^*_{\max}$ ($L^*_{\min}$).

3. Repeat steps (1) and (2) many times, to estimate the bias correction term $z_{0,max}$ ($z_{0,min}$), the proportion of bootstrapped upper (lower) bounds that are lower than the estimated upper (lower) bound; that is, $z_{0,max} = \Phi^{-1}(P^*(L^*_{\max} \leq \hat{L}_{\max}))$

($z_{0,min} = \Phi^{-1}(P^*(L^*_{\min} \leq \hat{L}_{\min}))$) and then to get $p_{0,max} = \Phi(2z_{0,max} + z_{1-\alpha})$ ($p_{0,min} = \Phi(2z_{0,min} + z_{\alpha})$) , where $z_{1-\alpha}$ is the $(1-\alpha)$-quantile of the standard normal distribution.

4. The upper (lower) bound of the confidence set is the p_2-quantile of the distribution of the bias corrected $L^*_{\max}$ ($L^*_{\min}$), so the number $c_{\max}$ ($c_{\min}$) that solves $P^*(L^*_{\max} \leq c_{\max}) = p_{0,max}$ ($P^*(L^*_{\min} \leq c_{\min}) = p_{0,min}$).

There also exists a bias-corrected and *accelerated* (BC_a) bootstrap confidence interval that corrects for the fact that the standard error for $\hat{L}_{\max}$ is not the same for all $L_{\max}$ as it would be under the standard normal approximation (Efron, 1987). However, the estimation of the parameter of acceleration is computationally costly: it requires n calculations of the optimal value. The BC_a bootstrap is more precise and is second-order accurate, so that the coverage error is of order $1/n$ instead of $1/\sqrt{n}$ as for the percentile CIs.

2.3.5 Imbens and Manski

This method was developed in Imbens and Manski (2004) (IM) and was the first inferential scheme for the partially identified scalar parameter. Unlike the percentile bootstrap methods, which were developed for a point-identified parameter, this method covers the true parameter with a prescribed probability, no matter what this true parameter happens to be.

1. Generate a bootstrap sample of data with replacement, and let A^* and b^* denote the estimates of A and b based on the bootstrap sample.

2. Calculate the optimal value of (2.2) for the bootstrap sample, and denote it as $L^*_{\max}$.

3. Repeat steps (1) and (2) many times, to get an estimate of the standard deviation $\hat{\sigma}_{\max}$ ($\hat{\sigma}_{\min}$) of $L_{\max}$ ($L_{\min}$). Find c_α that solves $\Phi(c_\alpha + \frac{L^{\max} - L^{\min}}{max\{\hat{\sigma}_{\max}, \hat{\sigma}_{\min}\}}) - \Phi(-c_\alpha) = 1 - \alpha$.

4. The upper (lower) bound of the confidence set is $\hat{L}_{\max} + c_\alpha \hat{\sigma}_{\max}$ ($\hat{L}_{\min} + c_\alpha \hat{\sigma}_{\min}$).

The implementation is straightforward, and finding c_α is simple. The asymptotic properties for this interval were derived under the assumption that the upper and lower bounds asymptotically behave like means. Stoye (2009) observes that one of the assumptions that IM impose implies superefficient estimation of the length of the identified set and proposed a weakened version of the assumption. It was also recognized that inconsistency of the bootstrap comes from the boundary problem (Andrews, 2000) and can be similarly resolved by pretesting whether the nuisance parameter and the length of the identified set is zero or not. A similar idea was used in Bugni (2010, 2011) in the context of models defined by moment inequalities.

2.3.6 Imbens and Manski Bias Corrected

This method is similar to the previous method with the difference being that the confidence region is adjusted for the finite sample bias.

1. Generate a bootstrap sample of data with replacement, and let A^* and b^* denote the estimates of A and b based on the bootstrap sample.

2. Calculate the optimal value of (2.2) for the bootstrap sample, and denote it as L^*_{max} (L^*_{min}).

3. Repeat steps (1) and (2) many times, to obtain an estimate of the standard deviation $\hat{\sigma}_{\text{max}}$ ($\hat{\sigma}_{\text{min}}$) of L_{max} (L_{min}) and to estimate the finite sample bias by $bias^*_{\text{max}} = E^*(L^*_{\text{max}}) - \hat{L}_{\text{max}}$ ($bias^*_{\text{min}} = E^*(L^*_{\text{min}}) - \hat{L}_{\text{min}}$). Find c_α that solves $\Phi(c_\alpha + \frac{L^{\text{max}} - L^{\text{min}}}{max\{\hat{\sigma}_{\text{max}}, \hat{\sigma}_{\text{min}}\}}) - \Phi(-c_\alpha) = 1 - \alpha$.

4. The upper bound of the confidence set is $\hat{L}_{\text{max}} - bias^*_{\text{max}} + c_\alpha \hat{\sigma}_{\text{max}}$ ($\hat{L}_{\text{min}} - bias^*_{\text{min}} + c_\alpha \hat{\sigma}_{\text{min}}$).

2.3.7 Projection

The linear programming formulation of the identified set allows us to introduce a new method for constructing the confidence set.

Given that the only way that the data enters the analysis is via the vector of mass probabilities p, we can consider the set of all probability vectors that do not differ much from the observed $\hat{p}$; more precisely, the set of all vectors p so that $\hat{p}$ would not have been rejected as being equal to p by some test of equality of the vectors. Every vector p in this set gives us an estimate of the identified interval. Taking a union of these identified intervals across the set of vectors so that $\hat{p}$ is not statistically different from them gives a confidence set that meets the coverage requirement.

Therefore we propose the following confidence region:

$$\hat{C}_n = \bigcup_{p \in \hat{\mathbf{P}}_n} \Theta_I(p), \tag{2.3}$$

$$\hat{\mathbf{P}}_n = \left\{ p : \sum_{i=1}^{k} p^i = 1; \forall i = 1, \ldots, k : p^i \geq 0; \sum_{i=1}^{k} n \frac{(\hat{p}_n^i - p^i)^2}{p^i} \leq q_{k-1}^{(1-\alpha)} \right\}, \tag{2.4}$$

where $q_{k-1}^{(1-\alpha)}$ is $(1-\alpha)$-quantile of chi-squared distribution with $k-1$ degrees of freedom and $\hat{\mathbf{P}}_n$ is a set of all probability vectors p so that a Pearson's chi-squared test would not have rejected the null hypothesis "H_0: data come from p".[7] We *project* the confidence interval for p into the one for the identified set; therefore, we call it a projection. An idea similar to this one is presented in Woutersen and Ham (2013).

It is easy to show that the confidence region given by (2.3) guarantees the asymptotic coverage requirement:

$$\lim_{n\to\infty} \Pr(\Theta_0 \subseteq \hat{C}_n) = \lim_{n\to\infty} \Pr\left(\Theta(p_0) \subseteq \bigcup_{p \in \hat{\mathbf{P}}_n} \Theta_I(p) \right) \geq \lim_{n\to\infty} \Pr(p_0 \in \hat{\mathbf{P}}_n) \geq 1 - \alpha, \tag{2.5}$$

where the first equality is maintained because $p_0 \in \hat{\mathbf{P}}_n \implies \left(\Theta(p_0) \subseteq \bigcup_{p\in\hat{\mathbf{P}}_n} \Theta_I(p)\right)$, and the second is the coverage of a standard chi-squared test with fewer degrees of freedom because $\sum_{i=1}^{k} p^i = 1$. This procedure is schematically depicted in Figure 2.3.

The downside of this method is that it can be too conservative and the confidence bounds may be too wide and uninformative, because the first inequality in 2.5 is strict. There is no way that we can guarantee an exact coverage of this procedure. In problems where it is suspected that the bootstrap procedures will fail, the projection method may be preferable. If the confidence set based on the projection does not differ much from the one based on the bootstrap, this suggests that the confidence

[7] There are many tests that can be used, and Pearson's chi-squared test was chosen because it is computationally convenient.

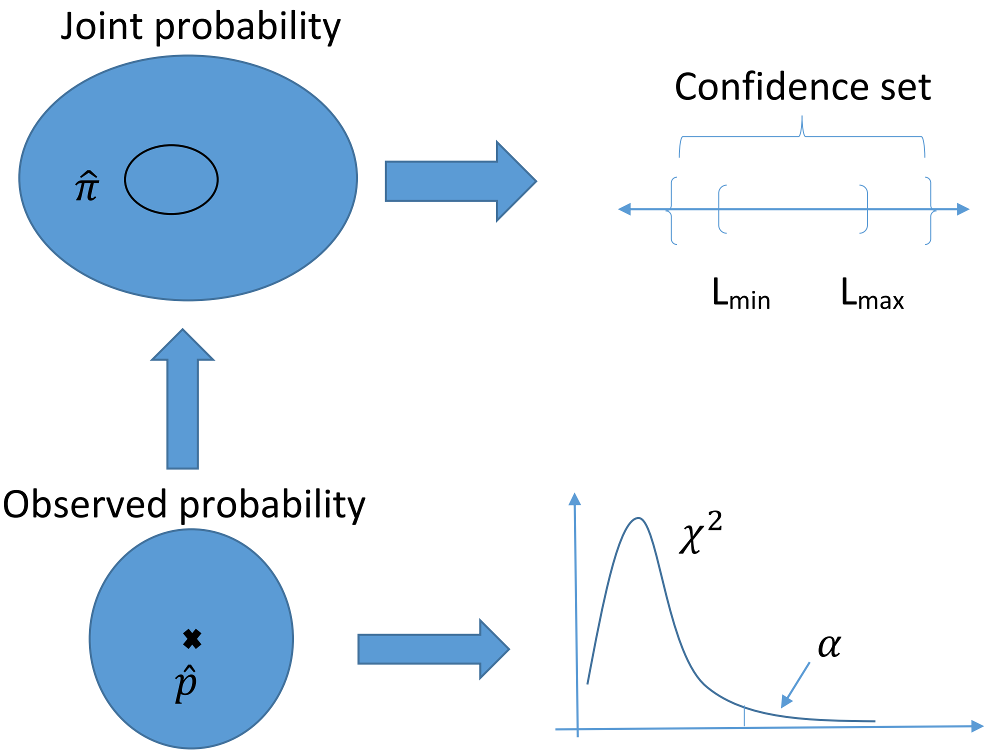


Figure 2.3: This scheme shows how the projection method works. For all observed probability distributions that are not statistically different from $\hat{p}$ according to a chi-squared test, we find the joint distributions that maximize and minimize the average treatment effect. The union of these intervals constitutes the confidence set with the desired asymptotic coverage.

set has correct coverage. In contrast, if the two confidence sets differ, it should not be concluded that the bootstrap failed to provide an accurate approximation.

Implementation requires an optimization over the set of probability vectors in the outer loop. The projection method is therefore only feasible for small-scale problems.

2.3.8 Alternative methods

This subsection briefly discusses some alternative confidence sets that are not included in the simulation study. These methods are omitted because they are impractical or computationally expensive.

- Bias-corrected and accelerated percentile bootstrap
- Calibration
- Subsampling
- Freyberger and Horowitz 2

Bias-corrected and accelerated percentile bootstrap

Efron (1981) introduces the bias-corrected and accelerated percentile bootstrap interval. The *bias-correction* part was presented in Subsection 2.3.4. The *acceleration part* corrects for the fact that the variance of the estimator is nonconstant. More precisely, suppose that there exists a monotone transformation $\phi = m(\theta)$ of the parameter of interest θ, so that $\hat{\phi} = m(\hat{\theta})$ is normally distributed with a bias and a nonconstant standard deviation $\hat{\phi} \sim N(\theta - z_0\sigma_\phi, \sigma_\phi^2)$, $\sigma_\phi = 1 + a\phi$. The parameter z_0 is called bias correction, and a is the acceleration parameter. Full derivation and discussion of the acceleration parameter is in DiCiccio and Efron (1996).[8] The calculation of the acceleration parameter requires n evaluations of the statistic of interest and makes it impractical for our purposes.

Calibration

Calibration as a technique to improve the coverage accuracy is presented in Efron and Tibshirani (1993) and DiCiccio and Efron (1996). Suppose we are interested in the $100\alpha\%$ upper confidence bound $\hat{L}^{\alpha}_{\max}$, and let $\beta(\alpha) = Pr(L_{\max} < \hat{L}^{\alpha}_{\max})$ denote the true coverage probability. It is possible to use the bootstrap to estimate the calibration curve $\beta(\alpha)$. We fix the estimator $\hat{L}_{\max}$ and create a large number of bootstrap samples, and for every pseudo-sample, we calculate the upper confidence bound $\hat{L}^{*\alpha}_{\max}$. This requires many evaluations of $\hat{L}_{\max}$, as the upper confidence bounds are also obtained by the bootstrap. The estimated calibration curve is equal to $\hat{\beta}(\alpha) = Pr^*(\hat{L}_{\max} < \hat{L}^{*\alpha}_{\max}) = Pr^*(\hat{\alpha}^* < \alpha)$, where $\hat{\alpha}^*$ is the value for which the upper confidence bound is equal to the estimated upper bound: $\hat{\alpha}^* : \hat{L}^{*\alpha}_{\max} = \hat{L}_{\max}$. If we are interested in the 90% upper confidence bound and $\hat{\beta}(0.93) = 0.9$, then $\hat{L}^{*0.93}_{\max}$ is the 90% upper confidence bound.

Subsampling

Subsampling (Politis et al., 1999) provides a theoretically interesting alternative to the bootstrap, because it correctly approximates the asymptotic distribution of a statistic even in the cases when the bootstrap fails. It is based on drawing samples of smaller

[8] MATLAB's `bootci` function uses the bias-corrected and accelerated percentile method as the default option.

size *without* replacement. The main difference is that whereas the bootstrap draws samples from the *estimated* model, the subsampling samples are drawn from the *original* model. The subsampling theory is based on the U-statistic, and the requirements for obtaining the correct asymptotic distribution are easily satisfied as follows.

- The subsample size b_n must go to infinity at a rate lower than n: $b \to \infty$, $b_n/n \to 0$.
- $n^{\beta}(\hat{\theta}_n - \theta)$ converges to some nondegenerate distribution J (with β known).[9]
- The limiting distribution J is continuous at the approximated quantile.

The biggest practical problem is the choice of b_n. While the general asymptotic theory requirements for the subsample size are satisfied easily, in finite samples, the choice of b_n greatly affects the size of the confidence sets. Furthermore, the subsampling tends to be less precise than the bootstrap in situations where the bootstrap works. The recent advances in the theory of subsampling provide conditions for uniform asymptotic validity (Romano and Shaikh, 2012).

Freyberger and Horowitz 2

This method is similar to the method presented in Subsection 2.3.1, with the only difference being that instead of calculating the objective value corresponding to the optimal vertex of the feasible solutions of problem (2.1), we calculate the optimum across all the vertices that give objective function values in the c_n neighborhood of $\hat{L}_{\max}$. This is a matter of pretesting, where it is not known whether the basic solution that is optimal in our data sample is also optimal for the true data-generating process. However, as the data sample increases in size, it must not become too large. The sequence of constants c_n must converge to zero but not too fast ($c_n[n/(\log\log n)]^{\frac{1}{2}} \to \infty$), so that by the Law of Iterated Logarithm, the optimal vertex (or vertices) will not be missed, with the probability going to one as the sample size n approaches infinity.

1. Generate a bootstrap sample of data with replacement, and let A^* and b^* denote the estimates of A and b based on the bootstrap sample.

[9] Usually, β is either 0.5 or 1, but β can be estimated following the arguments in Bertail et al. (1999).

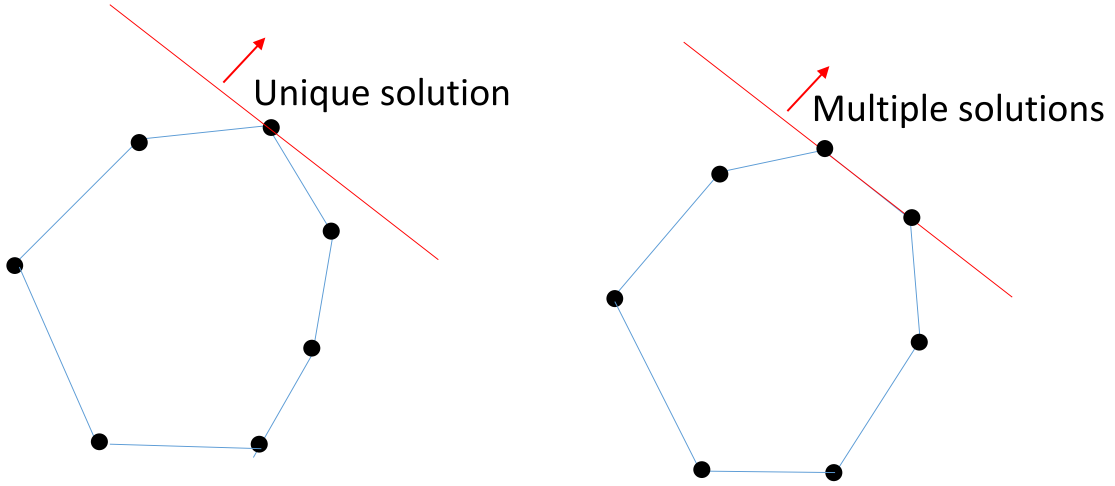


Figure 2.4: This figure illustrates the simplex of feasible solutions for Problem 2.1. There is a unique solution in the left pane. The objective function may happen to be collinear with one of the active inequalities as depicted in the right pane.

2. Let $\hat{\beta}_{\max}$ be the set of indices of basic solutions of the problem (2.2) that satisfies $|\hat{L}_{\hat{k}_{\max}} - \hat{L}_k| \leq c_n$, and define $\Delta^* = \max_{k \in \hat{\beta}_{\max}} n^{1/2}(c^T_{\hat{k}} \hat{A}^{*-1}_k)\hat{b}^*$, where the subscript indicates the columns corresponding to the basic variables.

3. Repeat steps (1) and (2) many times, to obtain the empirical distribution of $\Delta^*_{\hat{k}_{\max}}$. Find $c^*_{\alpha,\max}$ that satisfy $P^*(\Delta^*_{\hat{k}_{\max}} \geq -c^*_{\alpha,\max}) = 1 - \alpha$.

4. The upper bound of the confidence set is $\hat{L}_{\max} + n^{-1/2}c^*_{\alpha,\max}$.

The implementation challenges are similar to the method in Subsection 2.3.1 with extra effort required to identify the set of indices $\hat{\beta}_{\max}$; i.e., vertices of the simplex of the feasible solutions that cannot be rejected as nonoptimal. This is in general a challenging problem, especially if the solutions are degenerate. One solution is to create a large number of bootstrap datasets and to consider all basic solutions that are optimal in at least one generated dataset. This method should theoretically be preferred to the previous method in case when the true optimal solution is not unique as schematically shown in Figure 2.4, and the previous procedure may produce conservative confidence sets that are too wide.

	$V=1$		$V=2$		$V=3$		$V=4$	
	$Z=0$	$Z=1$	$Z=0$	$Z=1$	$Z=0$	$Z=1$	$Z=0$	$Z=1$
$Y=0$	0.3965	0.0134	0.0550	0.0099	0.0291	0.0126	0.0173	0.0121
$Y=1$	0.1553	0.0169	0.0550	0.0182	0.0539	0.0433	0.0467	0.0648

Table 2.1: Cell probabilities of observed variables for Scenarios 1 and 3.

2.4 Monte Carlo Simulation Study

This simulation study considers three different scenarios that mimic real situations.

1. Upper bound under the MTR+MIV assumption.

2. Upper bound under the MIV assumption that has no identifying power.

3. Identified set under the relaxed MTR+MIV+cMTS assumption.

2.4.1 Scenario 1 – Empirical Application

The first scenario considers the inference on the upper bound under the MTR+MIV assumption. The MTR assumption sets the lower bound of the average treatment effect to be fixed at zero; hence, we are only interested in the upper bound. The setup considered is that of Example 2.2.1. This scenario is relevant, as it was used in the empirical paper de Haan (2011). It is assumed that the true vector of probabilities p is the one constructed from the data from de Haan (2011), and Table 2.1 lists these probabilities. This probability vector was used to generate 1000 pseudo-samples. For each pseudo-sample, we calculated the confidence sets based on different methods. For methods based on the bootstrap, we used 200 bootstrap replications from each pseudo-sample to approximate the relevant quantiles and to construct the confidence sets.[10]

The identified set in this situation is $(0, 0.58822)$. An analytic expression for the upper bound is available.[11]

Figure 2.5 reports empirical coverages from the simulation. All methods show undercoverage in small samples. There is not a clear winner, but the bias-corrected

[10] For confidence regions, Efron (1987) recommends 2000 bootstrap replications. However, the differences between using 200 and 2000 bootstrap replications were negligible.

[11] In our case, with the binary outcome $y(y)$ and binary treatment z, the expression for the upper bound on the average treatment effect under the MTR-MIV assumption reduces to

methods and the normalized percentile bootstrap together are closest to the nominal coverage. In contrast, the canonical percentile bootstrap performs worst. The Freyburger and Horowitz method is very similar to the Imbens and Manski method. For very large sample sizes, the differences diminish, and bias-corrected methods show slight overcoverage.

Figure 2.6 shows the empirical distributions of the upper 90% confidence bounds. All the methods are similar in terms of the variance and the shape of this distribution. This suggests that the gains from using the computationally more expensive bias-corrected accelerated percentile method to correct for shape or skewness are likely to be low.

Figure 2.7 depicts the evolution of the distribution of the upper 90% confidence bound as the sample size increases. All methods show a very similar pattern. The bias and the difference between the bias-corrected methods and their uncorrected counterparts diminish as the sample size increases. For a very large sample size, the empirical distributions of the different confidence regions almost coincide.

The histogram of different optimal basic solutions of the linear program is shown in Figure 2.8. For a very large sample size, the optimum of the linear program (2.2) is always realized on one particular vertex. This is because the benchmark linear program has a unique optimal solution, and this solution is selected with probability approaching one as the sample size grows.

The projection method for the upper 90% confidence bound is too conservative for $n = 16912$ as shown in Figure 2.9. Its empirical coverage is 100%, and the difference between the 10% quantile of the distribution of the confidence sets based on the projection and the true upper bound is 0.0149.

2.4.2 Scenario 2 – Assumption with No Identifying Power

This scenario will focus on inference on the upper bound under the MIV assumption with an artificially generated probability vector (shown in Table 2.2) so that the monotone instrument has no identifying power, and the MIV assumption is independent of

$$\sum_{m=1}^{4} P(v=m)\cdot\left[\min_{m_2\geq m}\left(P(z=0|v=m_2)(1-E[y|z=0,v=m_2])+P(z=1|v=m_2)E[y|z=1,v=m_2]\right)\right]$$

	$V=1$		$V=2$		$V=3$		$V=4$	
	$Z=0$	$Z=1$	$Z=0$	$Z=1$	$Z=0$	$Z=1$	$Z=0$	$Z=1$
$Y=0$	0.1703	0.0058	0.1703	0.0058	0.1703	0.0058	0.1703	0.0058
$Y=1$	0.0667	0.0073	0.0667	0.0073	0.0667	0.0073	0.0667	0.0073

Table 2.2: Cell probabilities of observed variables for Scenario 2.

the other observable variables. This scenario is interesting, as it highlights the fragility of the studied inference methods. A case similar to this one was considered in Manski and Pepper (2009) to illustrate the finite sample bias of the MIV bounds. In this case, a naive bootstrap procedure will approximate poorly the limiting distribution, and the associated confidence regions may suffer from low coverage. Chernozhukov et al. (2013) also consider a similar scenario to motivate a half median unbiased estimator of the bounds.

The identified set is $(-0.28979, 0.71021)$.

Figure 2.10 shows that the probability coverage does not improve as the sample size increases. With a small sample, some simulated probability vectors had zero mass probability and the identified sets were empty, even though the true data-generating process leads to a nonempty identified set. These cases were omitted and make the results seem better for smaller samples. The canonical percentile bootstrap performs worst, the Imbens and Manski method and the Freyberger and Horowitz method are similar and the two bias-corrected methods and the normalized and scaled bootstrap are closest to the desired nominal coverage.

Figure 2.11 shows that the distributions of the upper 90% confidence bound for the different methods do not coincide even for very large sample sizes. The variance decreases, but coverage does not improve. Among all the methods, only the canonical bootstrap differs from the other competing methods and has the lowest coverage. The shapes of the distributions are similar.

Figure 2.12 depicts how the distributions of the upper 90% quantile change with increased sample size for different methods. For small sample sizes, the distributions tend to be slightly skewed to the left. All the upper confidence bounds show the same pattern.

Figure 2.13 sheds some light on why the bootstrap methods fail to result in correct nominal coverage in the second scenario. Unlike in the first scenario, the histogram of the indices of the optimal basic solutions is similar for different sample sizes and does not change even for a very large sample size. The optimal solution, therefore, is not unique even asymptotically.

The upper 90% confidence bound based on the projection method led to a coverage of 97% and may actually be preferred over the bootstrap methods in this case for $n = 16912$. Figure 2.14 shows its distribution in comparison with the distributions of the other confidence bounds.

2.4.3 Scenario 3 – Sensitivity Analysis

In the third scenario, we consider inference on the whole identified set under a relaxed version of the cMTS+MIV+MSB assumption. The identifying assumptions are relaxed in the following ways.

- $P[z_i = t \implies y_i = y_i(t)] \geq 0.999$ - 0.1% of the outcomes may be mismeasured).
- $\forall z_2 \geq z_1 : E[y(t)|z = z_1] - E[y(t)|z = z_2] \leq 0.01$ - the difference between the potential probability of getting into college for children with college-educated mothers cannot be greater than that of children with mothers without college education.
- $\forall v_2 \geq v_1 : E[y(t)|v = v_1] - E[y(t)|v = v_2] \leq 0.01$ - the difference between the potential probability of getting into college for children with college-educated fathers cannot be greater than that of children with fathers without college education.

This scenario is interesting because relaxing assumptions can help researchers to conduct sensitivity analysis or to determine the source of the identifying power. In this case, there is no analytical solution available, and the linear programming formulation is essential. Laffers (2013a) provides a detailed discussion on the method and motivation for different relaxations of the identifying assumptions. Linear programming widens the scope of the usefulness of bounds analysis to different sets of identifying assumptions and sensitivity analyses. The vector of probabilities is the same as that in scenario 1 shown in Table 2.1. The identified set is $(-0.336, 0.244)$.

Figure 2.15 highlights the fact that the differences between the methods arise particularly in small samples. The confidence bounds show slight overcoverage even for large sample sizes, and the bias-corrected percentile bootstrap is closest to the nominal coverage. In small samples, the (canonical) percentile bootstrap shows severe undercoverage for both the upper and lower bounds. The coverage of the Freyburger and Horowitz method is slightly smaller than desired, whereas the coverage of both Imbens and Manski methods are above the nominal values. The normalized and the bias-corrected bootstrap seem to perform best in the current setup.

Figure 2.16 presents the coverage probability of the whole identified set, even though this was not an objective of these confidence sets. Differences are only visible in small samples, and here the bias-corrected bootstrap and the Freyburger and Horowitz method outperform the other methods. The Imbens and Manski methods together with the normalized bootstrap are slightly conservative, whereas the percentile bootstrap shows severe undercoverage.

When comparing the empirical distributions of the lower and upper confidence bounds (Figure 2.17), the distributions with smaller variance are preferred. The methods based on the bootstrap show the smallest variance with the exception of the normalized bootstrap methods. Furthermore, the shape is similar except for the distribution of the upper confidence bound based on the normalized bootstrap, which is bimodal.

Figure 2.18 shows that the optimal solution is not unique even if the sample size is very large. The distribution of the optimal basic solution settles down to two values.

Figure 2.19 shows that the 90% upper confidence bound is too conservative with an empirical coverage of 99.2%.

2.5 Conclusion

This paper has considered statistical inference for a partially identified scalar parameter in models with discrete observable variables. Furthermore, the methods considered in this paper are inferential alternatives that complement the identification results of Laffers (2013b), where the upper and lower bounds of a parameter are found by means of linear programming. The simulation study used three different simulation designs to compare the finite sample performance of different methods for constructing confidence bounds on the partially identified parameter. The three scenarios include: an actual empirical application, a case where one of the assumptions has no identifying power and relaxed assumptions that are relevant for studying sensitivity analysis. The results from the simulation study suggest the following implementation recommendations.

- In small samples, bootstrapping may lead to problems with (1) zero mass probability and (2) possible empty sets. Both of these problems should be addressed in any simulation study that considers performance of inference schemes on bounds in small samples.
- Possible degenerate optimal solutions lead to failure of assumption 4 of FH, which invalidates its asymptotic correct coverage.
- Bias correction is useful; bootstrap estimators are prone to bias, especially in small samples.
- Centering and normalization when using a bootstrap improves its finite sample properties.
- A histogram of the bootstrapped optimal basic solutions of a linear program can be useful for detecting bootstrap failure. If there exists only one optimal basic solution, the bootstrap yields the correct asymptotic distribution.
- The confidence set obtained by the projection method presented in this paper can be helpful, especially if there are reasons to believe that the bootstrap will fail. However, the projection method cannot detect the bootstrap failure.

As for all simulation studies, the results and recommendations presented in this paper are not general but design specific. Nevertheless, they shed some light on the behavior of different confidence sets for researchers conducting empirical studies using bounds analysis in models with discrete variables.

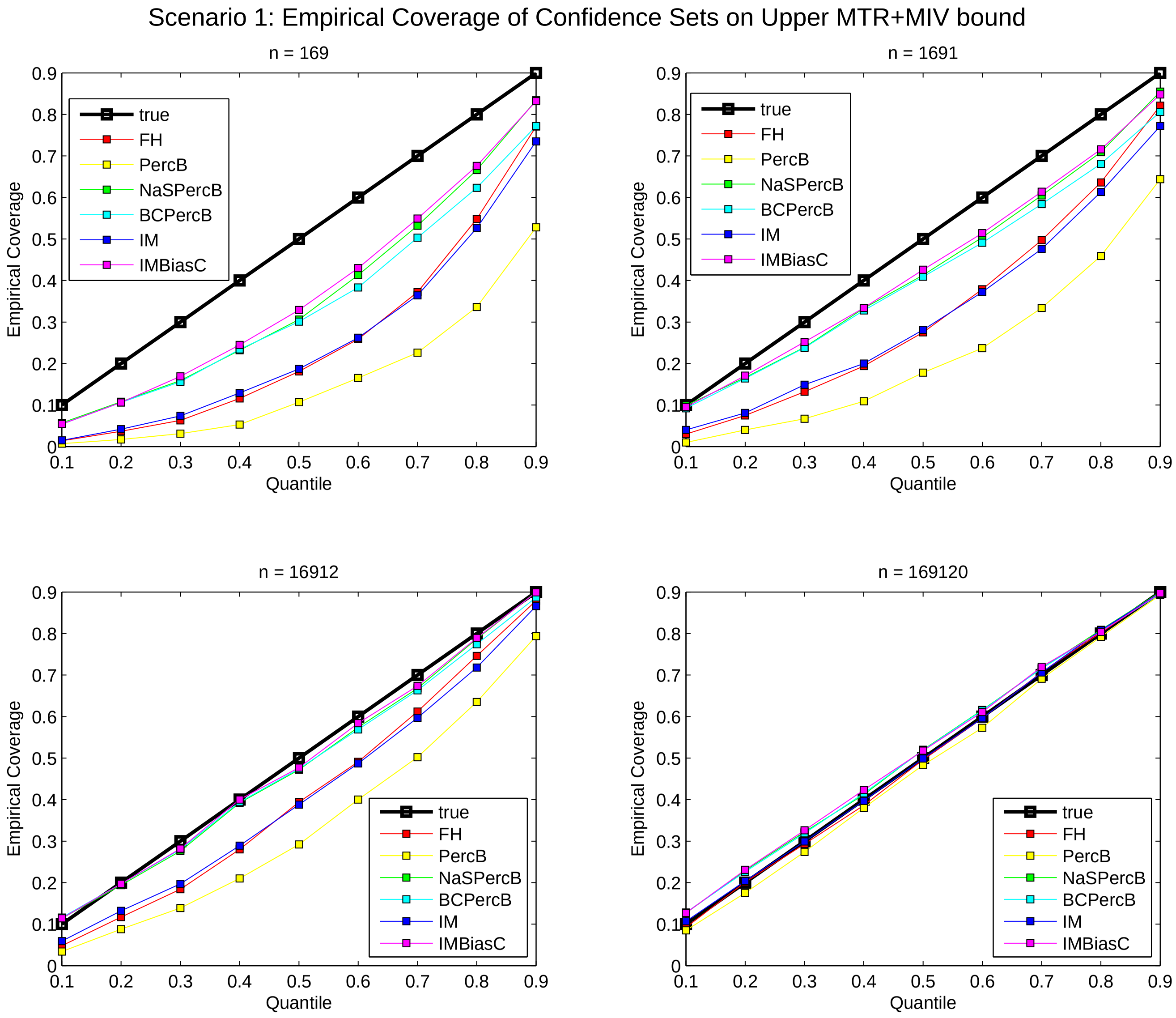


Figure 2.5: Empirical coverage of the upper MTR+MIV confidence bound.

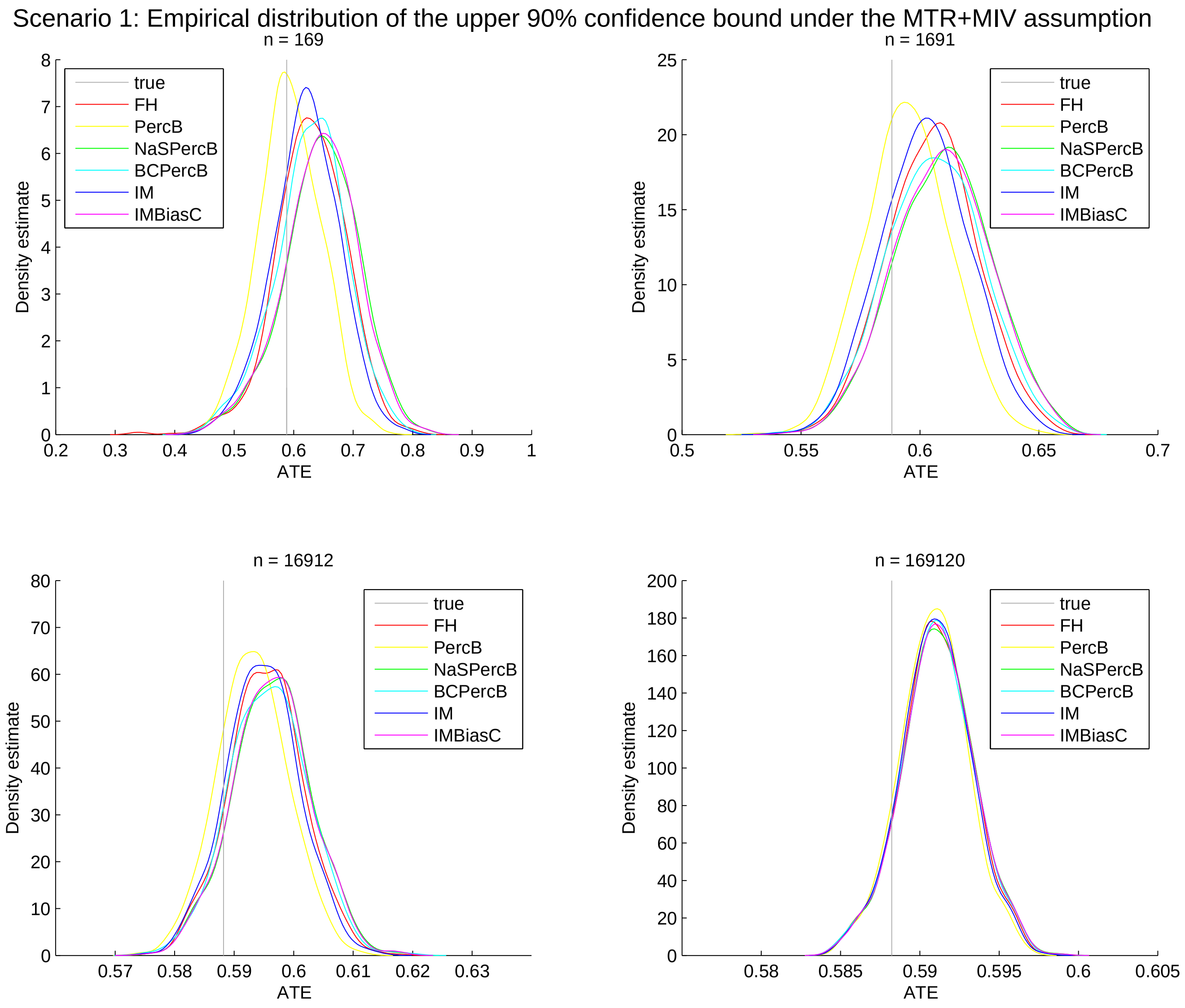


Figure 2.6: Empirical distributions of the upper MTR+MIV 90% confidence bound for different inferential methods.

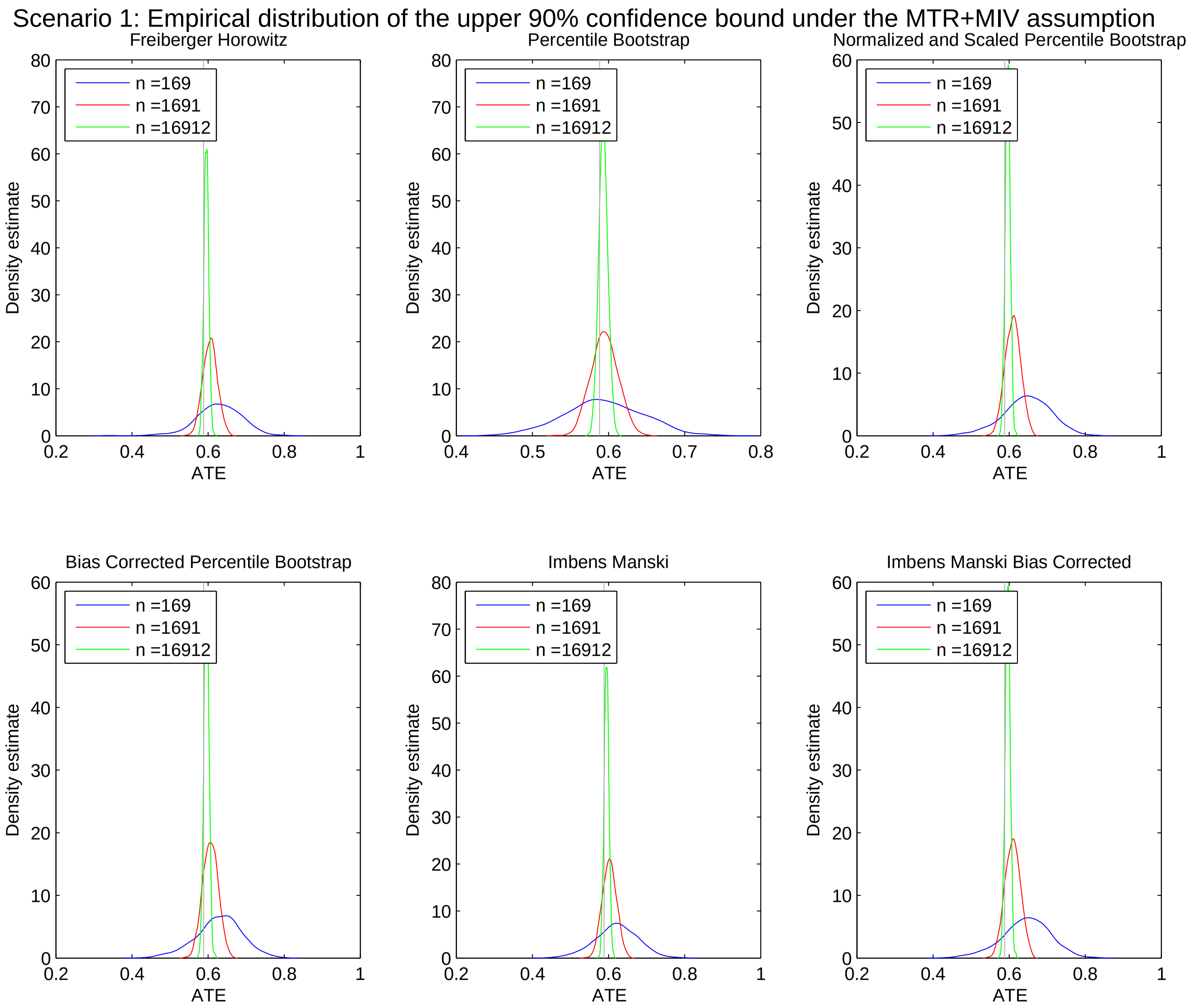


Figure 2.7: Evolution of the distribution of the upper MTR+MIV 90% confidence bound with a sample size increase.

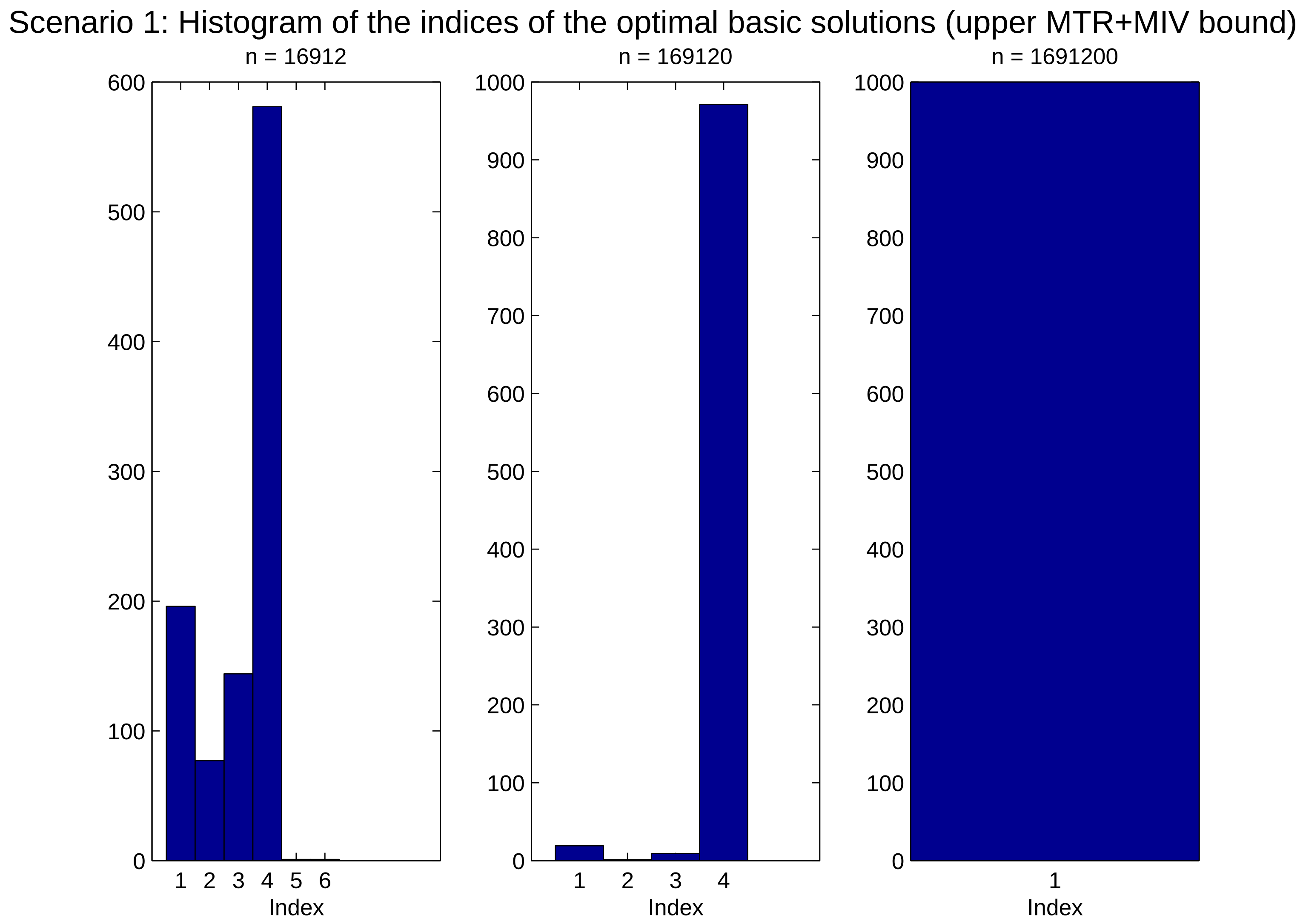


Figure 2.8: Histogram of different optimal basic solutions in the linear program (upper MTR+MIV bound).

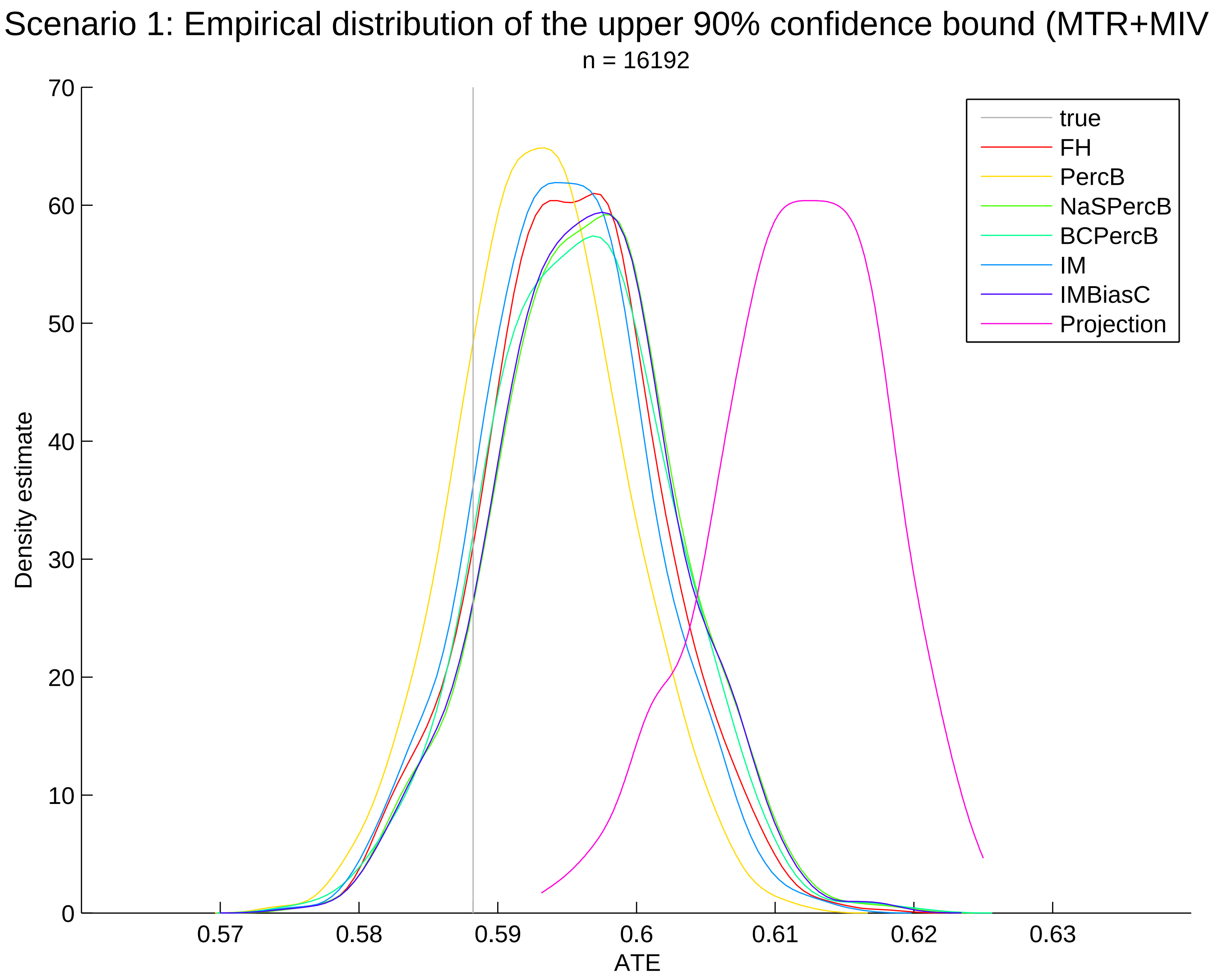


Figure 2.9: Projection method compared with other inferential schemes. The distribution of the 90% upper MTR+MIV confidence bound.

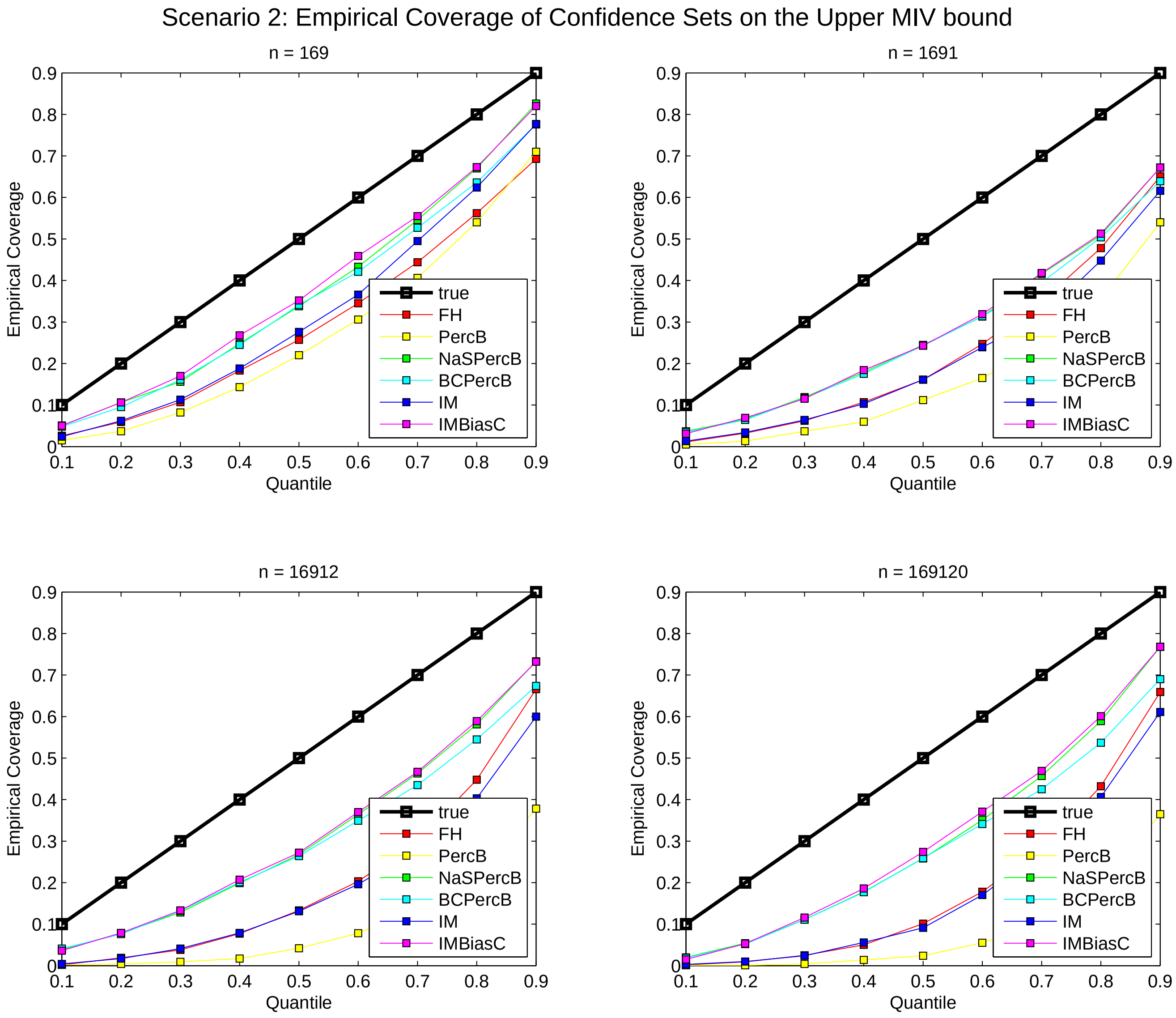


Figure 2.10: Empirical coverage of the upper MIV confidence bound.

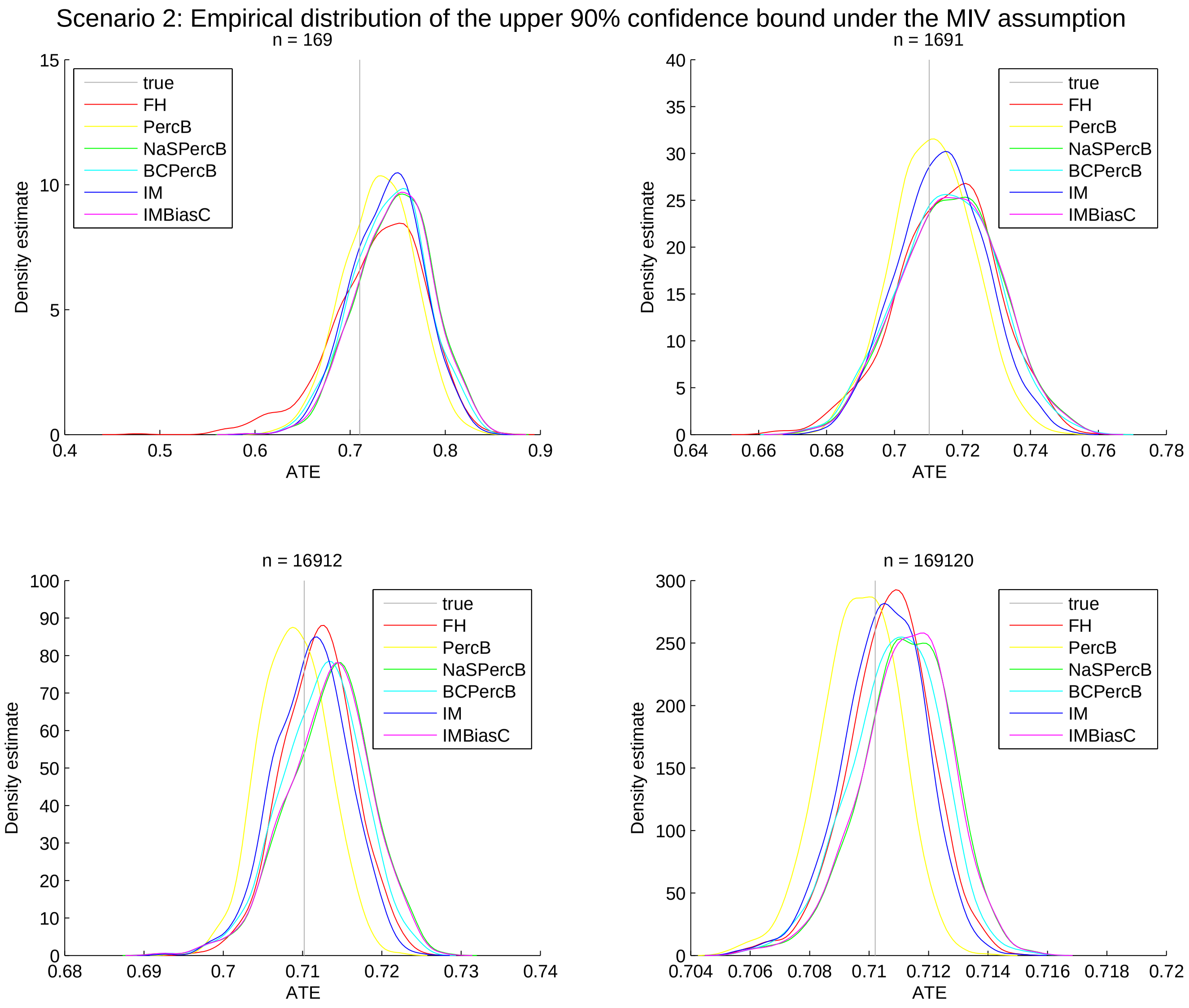


Figure 2.11: Empirical distributions of the upper MIV 90% confidence bound for different inferential methods.

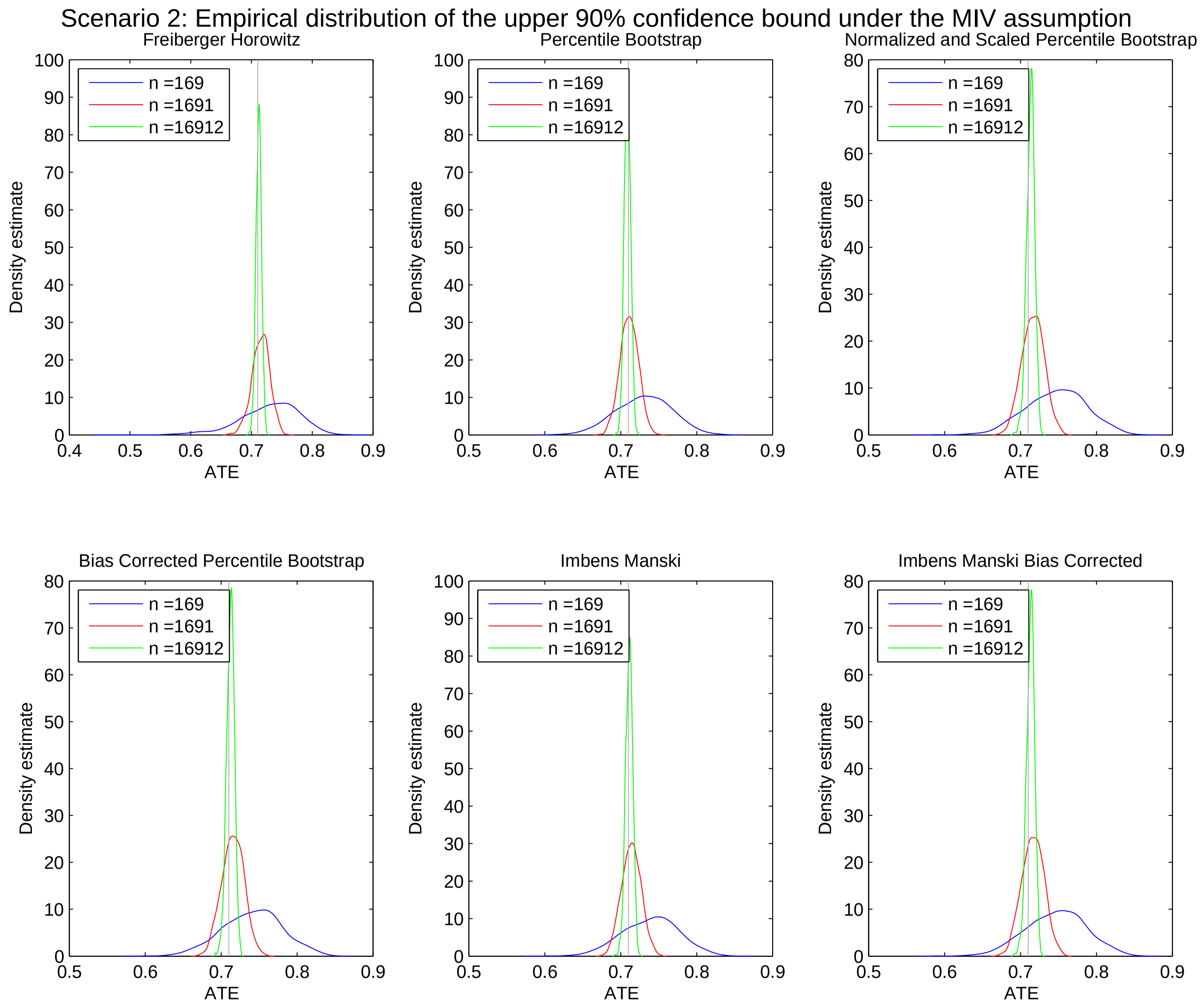


Figure 2.12: Evolution of the distribution of the upper MIV 90% confidence bound with a sample size increase.

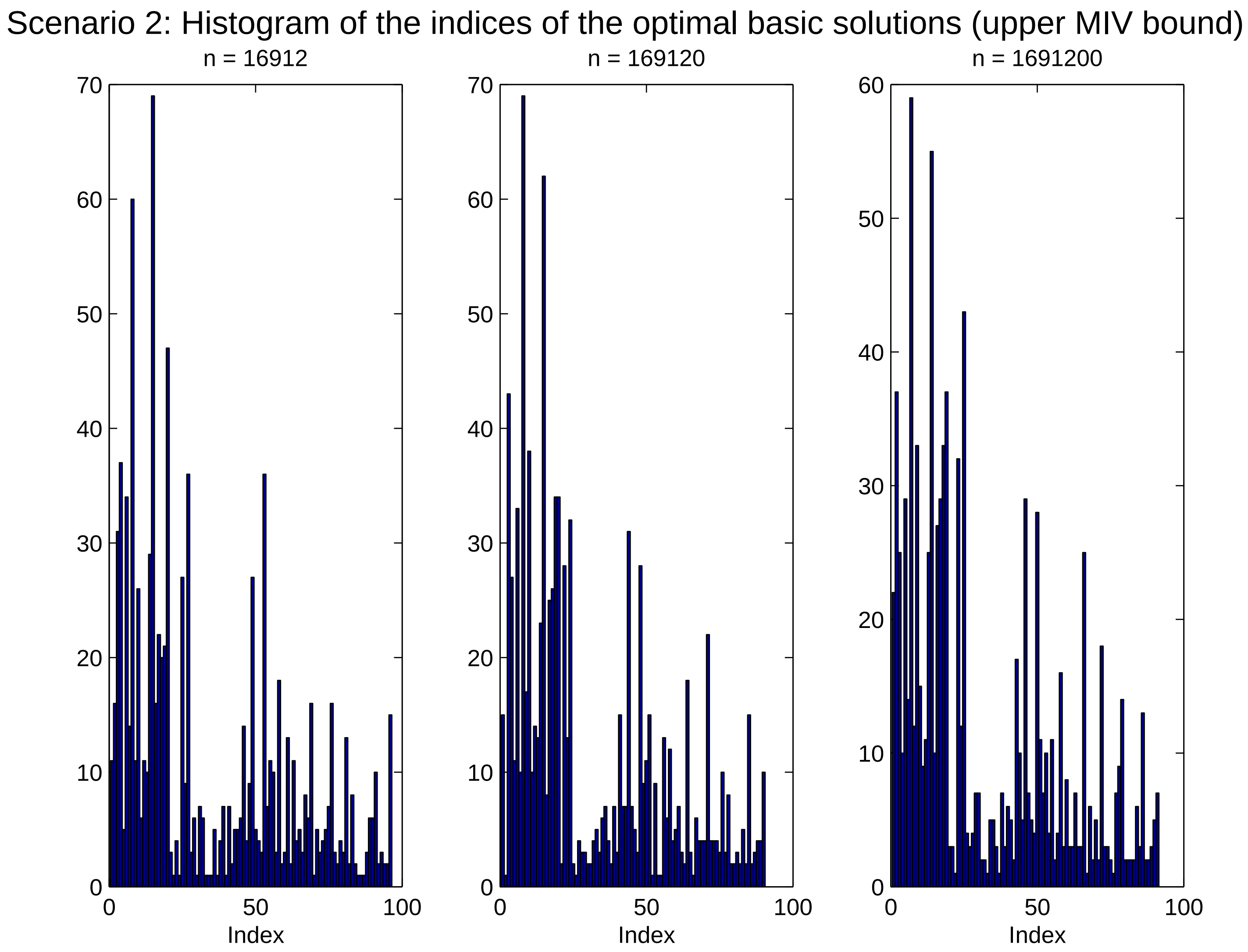


Figure 2.13: Histogram of different optimal basic solutions in the linear program (upper MIV bound).

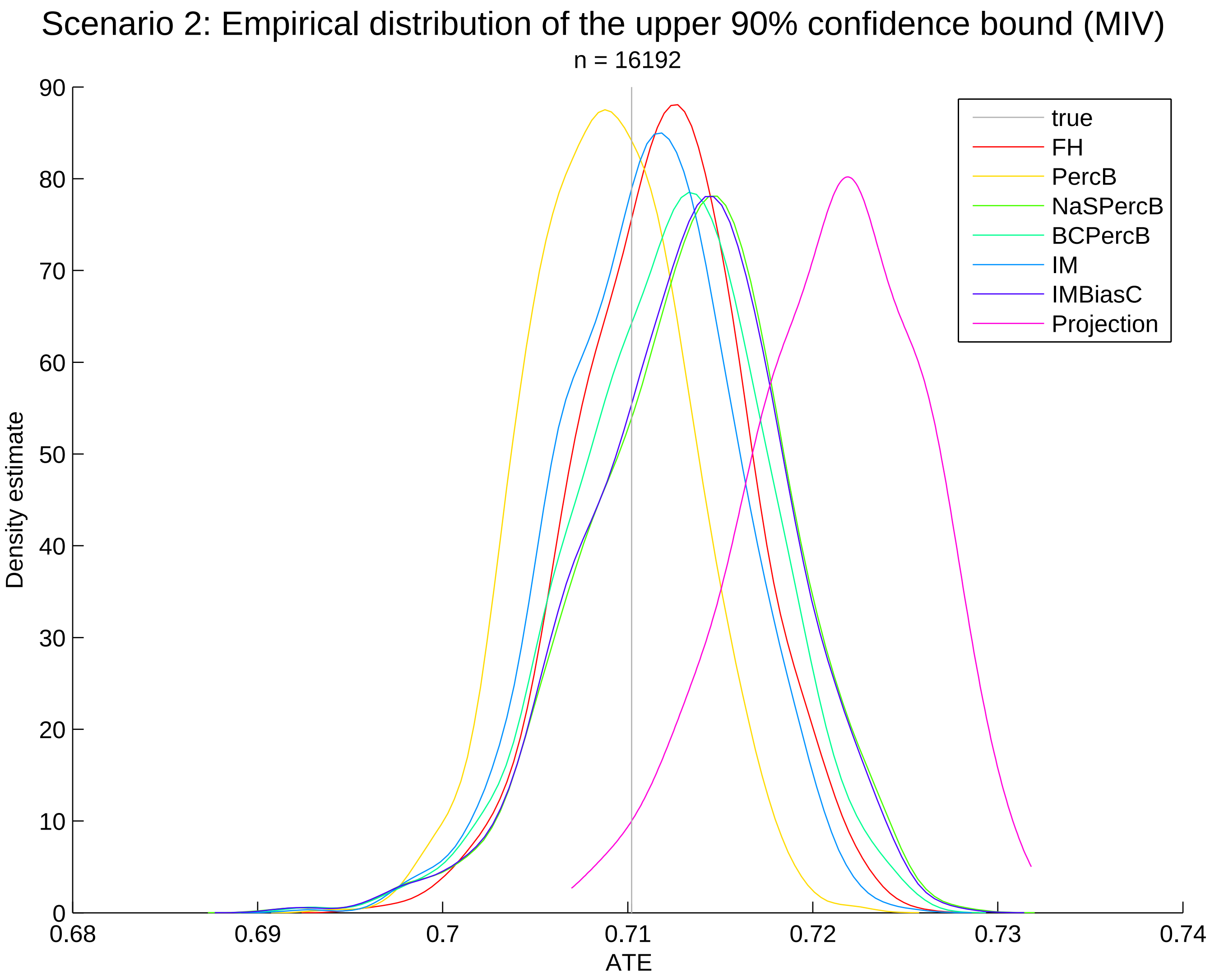


Figure 2.14: Projection method compared with other inferential schemes. The distribution of the 90% upper MIV confidence bound.

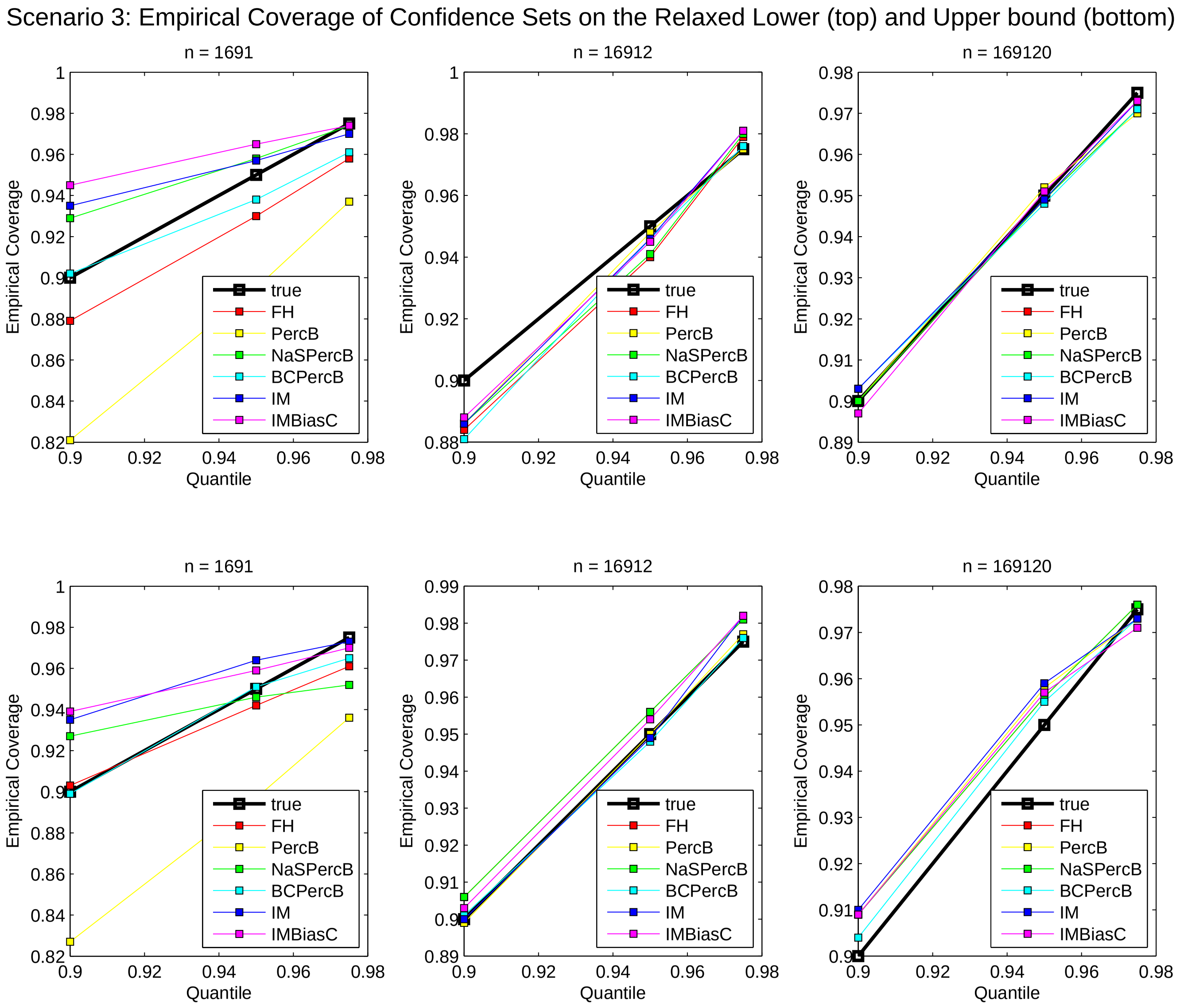


Figure 2.15: Empirical coverage of the lower (top) and upper (bottom) confidence bound.

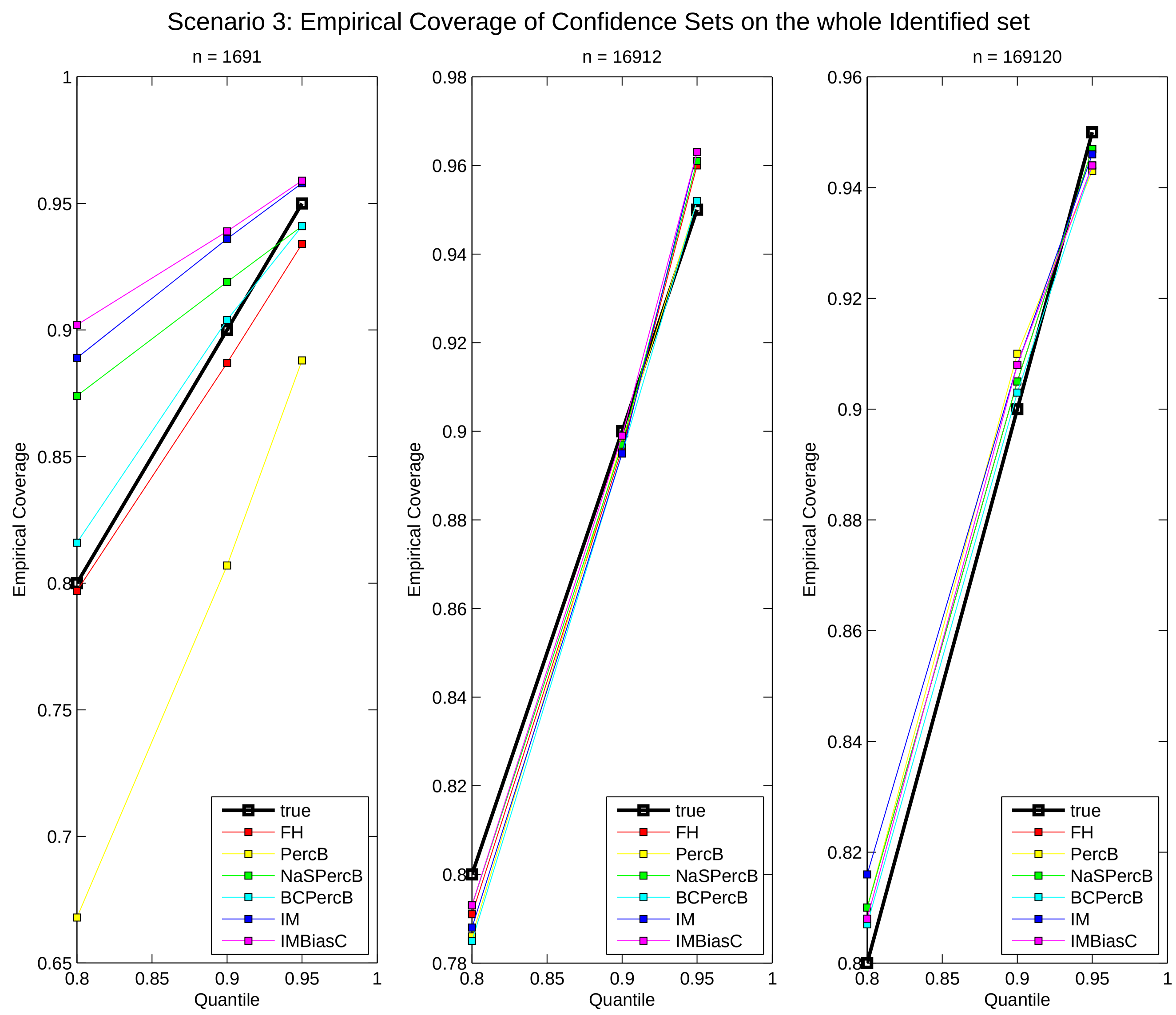


Figure 2.16: Empirical coverage of the whole identified set.

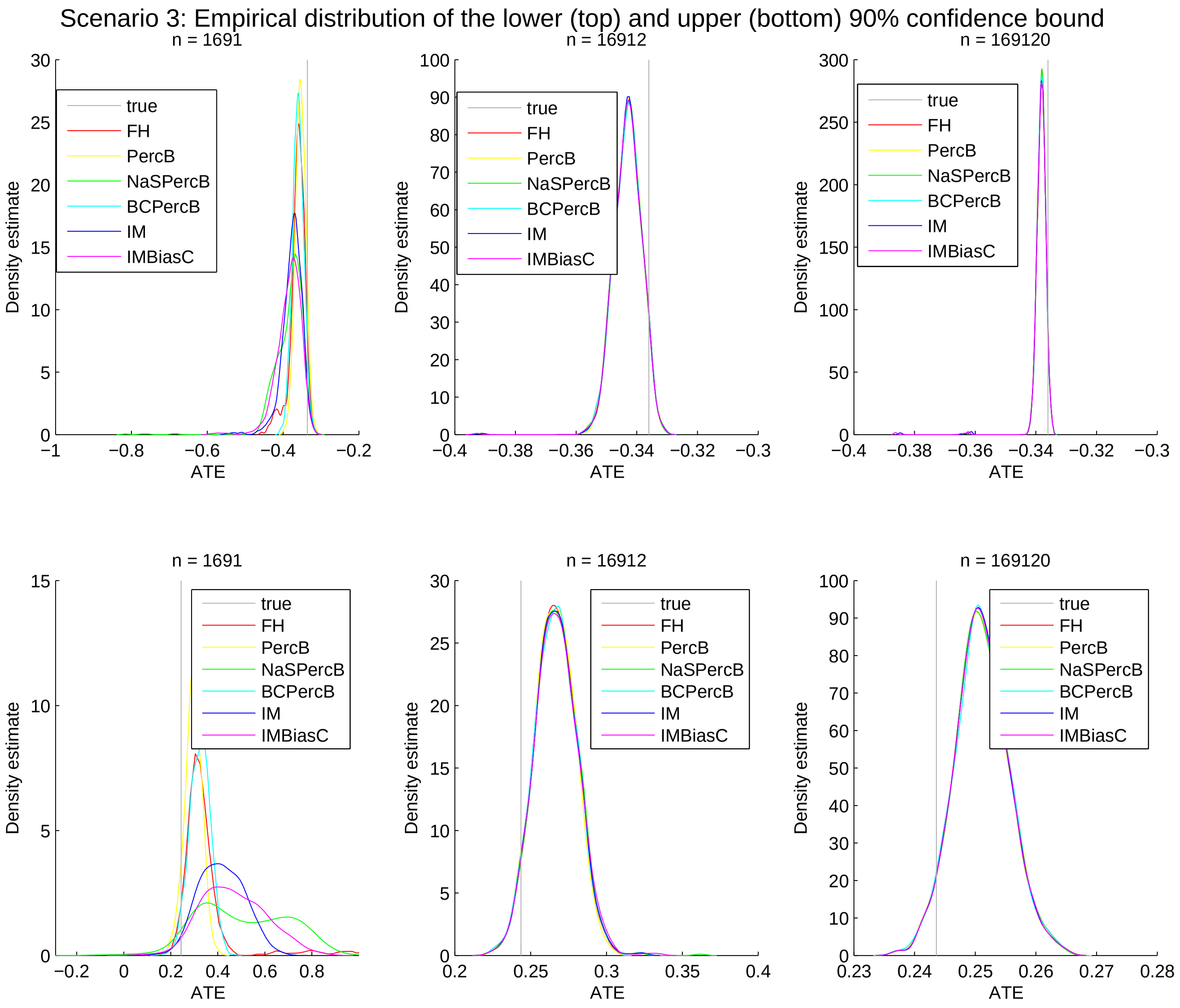


Figure 2.17: Empirical distributions of the lower (top) and upper (bottom) confidence bounds.

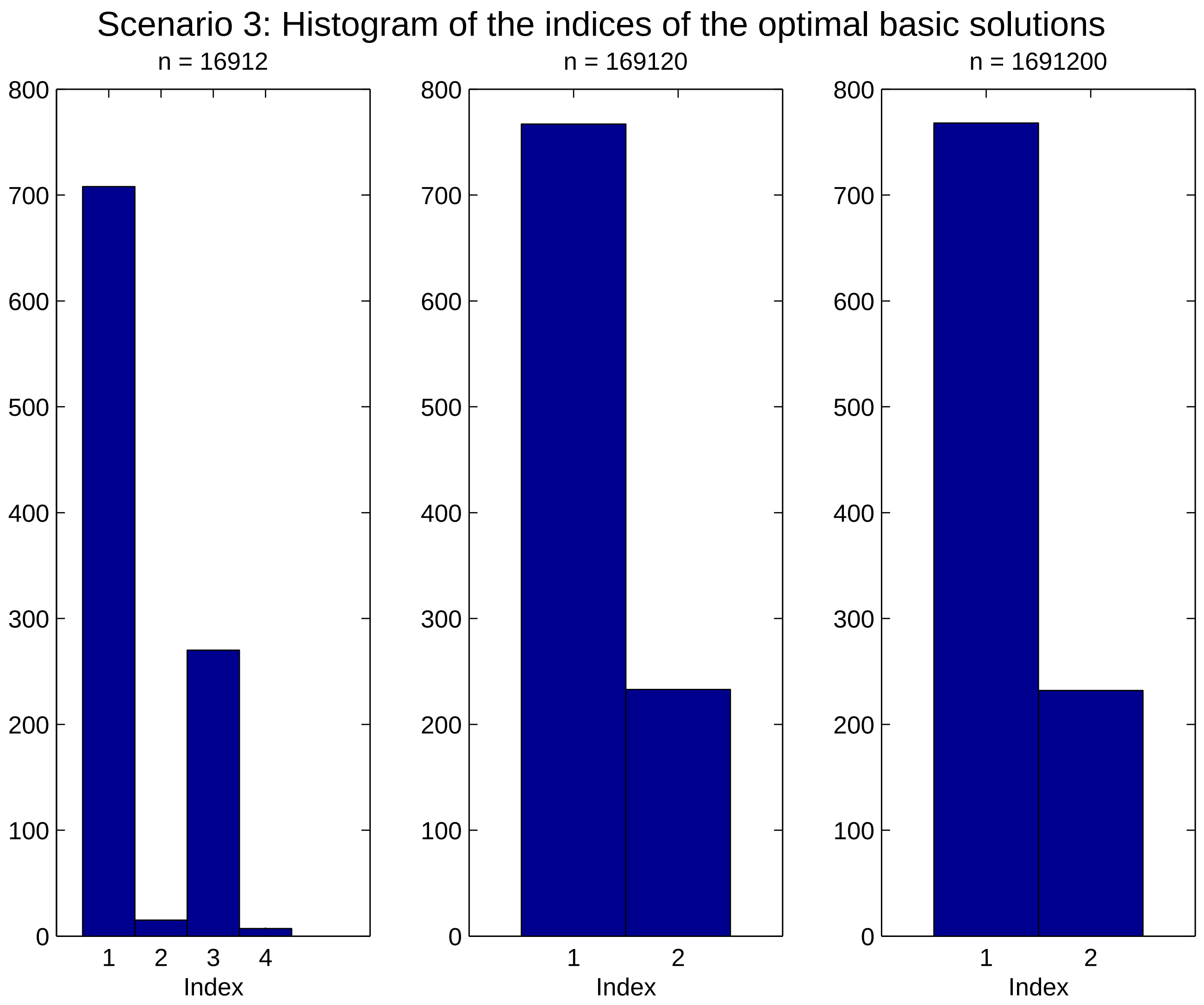


Figure 2.18: Histogram of different optimal basic solutions in the linear program.

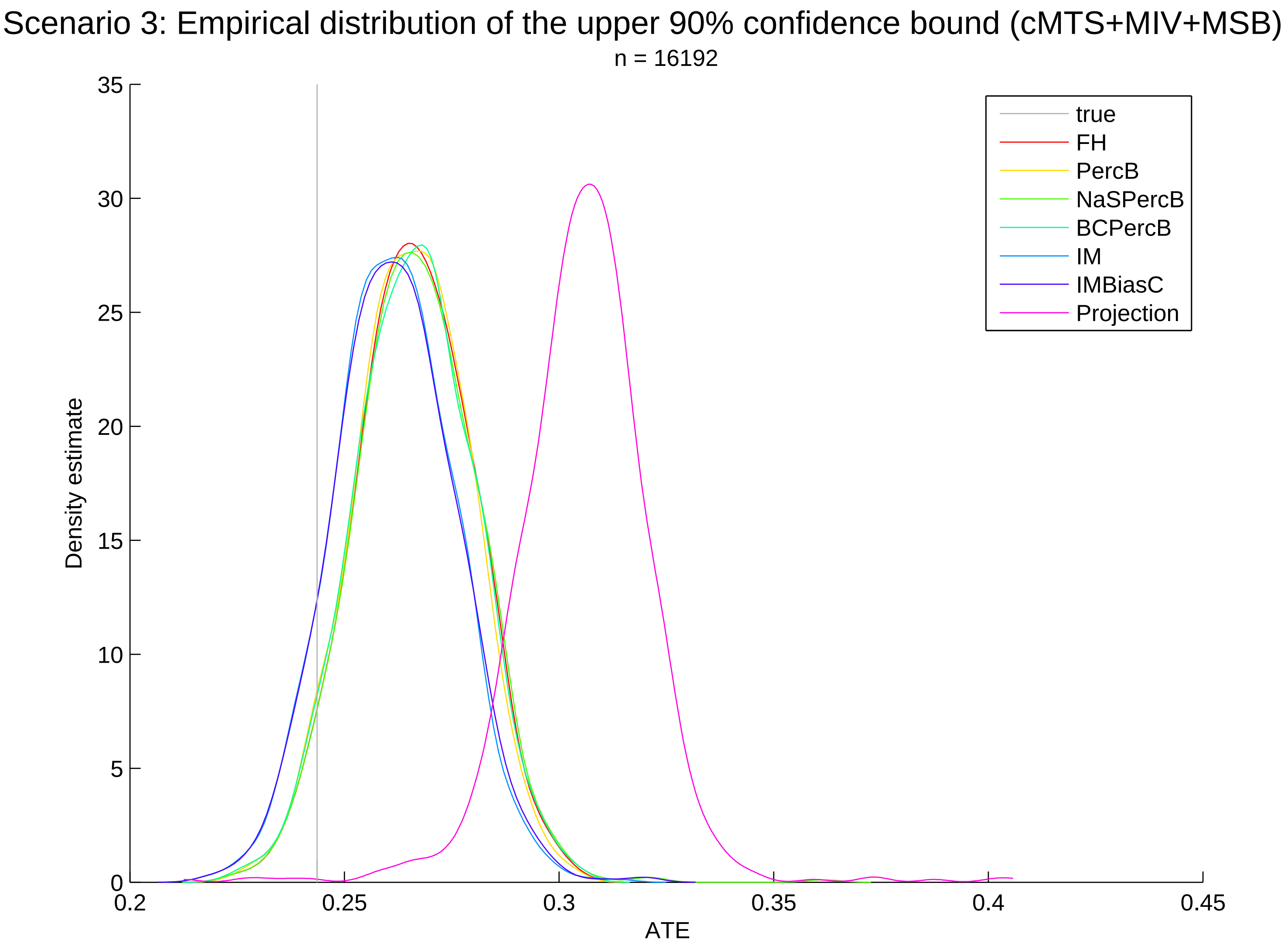


Figure 2.19: Projection method compared with other inferential schemes. The distribution of the 90% upper cMTS+MIV+MSB confidence bound.

Chapter 3

BOUNDING AVERAGE TREATMENT EFFECTS USING LINEAR PROGRAMMING

Abstract

This paper presents a method of calculating sharp bounds on the average treatment effect using linear programming under identifying assumptions commonly used in the literature. This new method provides a sensitivity analysis of the identifying assumptions and missing data in an application regarding the effect of parent's schooling on children's schooling. Even a mild departure from identifying assumptions may substantially widen the bounds on average treatment effects. Allowing for a small fraction of the data to be missing also has a large impact on the results.

JEL: C4, C6, I2.

Keywords: Partial identification, Bounds, Linear Programming, Average treatment effect, Sensitivity analysis.

3.1 Introduction and Literature Review

The recent literature on the average effect of parent's schooling on children's schooling appears inconclusive. Identification strategies based on twins, adoptees or instrumental variables lead to results that differ in size and statistical significance in terms of the average treatment effect and that lead to conflicting policy recommendations on educational reform. An attempt to address this problem was made in de Haan (2011), who studied the nonparametric bounds on the average treatment effect and relied on weaker nonparametric assumptions that have clear economic interpretations. Nevertheless, these assumptions may and should be challenged. This study discusses the validity and the importance of these assumptions. Moreover, this paper presents a method that allows some assumptions to be relaxed and an examination of how fragile or robust the reported bounds are to some mild violations of these assumptions. This paper also looks at how missing data may affect the results, and it imposes no structure on the missingness mechanism. Knowing what drives the results, and which assumptions are important, may sharpen the discussion about the underlying identifying assumptions, and also that about the economic problem at hand.

The contribution of this paper is twofold. First, this paper presents a flexible way of calculating the sharp bounds on the average treatment effect using a linear program. If all the variables are discrete, it is often practical to achieve identification by conducting a search of the set of joint probability distributions of the observed and unobserved variables. Second, this paper uses the linear programming method to compute the bounds on the average treatment effect when some or all of the identifying assumptions are relaxed, also allowing for the presence of missing data, in the context of the effect of parents' schooling on children's schooling. The linear programming formulation helps to clarify why one presumably irrelevant identifying assumption becomes important once another assumption is relaxed, and therefore, the two assumptions work as substitutes for each other.

There are two opposing explanations of how a parent's schooling affects a child's schooling. One relates to causation and the other to selection. Either the parents change during their education process (and this changes the way that they approach the education of their children) or the child's education merely reflects the transmission of the high-ability genes from his or her parents. An understanding of the intergenerational transmission of education has very important policy consequences. First, policy makers care about the return on investment to schooling. If the link between parents' schooling and children's schooling is causal, the beneficial spillover effect has to be taken into account when devising an educational policy. Second, if the effect is purely related to the transmission of genes, then the inequality in opportunities may simply be a consequence of the distribution of high-ability genes, and inequality-reducing policy is unlikely to be beneficial.

There are three main identifying strategies in the literature for estimating the *causal* effect of parents' education on children's education, as presented in a comprehensive overview in de Haan (2011).

The first approach is based on twins data in Behrman and Rosenzweig (2002, 2005) and Antonovics and Goldberger (2005). Children of identical twins should not differ much in the unobservable genetic endowments that they inherit from their parents, and this helps to remove an important source of correlation between parents' and children's schooling. This approach assumes that any differences other than genetic between the schooling levels of identical twins are exogenous.

The second method is based on adopted children (Björklund et al., 2006), where there clearly is no genetic link between the parents and the adopted children. This method assumes that the way the parents raise their children is unrelated to their schooling level.

The last approach is based on an instrumental variable. The strategy is to find a variable that provides a source of variation in parents' schooling that is unrelated to children's schooling. Black et al. (2005) use a school reform in Norway that changed the number of compulsory years of education from seven to nine. Chevalier (2004) use a law that changed the minimum school leaving age in the 1970s in Britain. Oreopoulos et al. (2006) also use the timing of the compulsory-schooling law changes as an instrument for completed parents' education. College availability is used as an instrument for maternal education in Currie and Moretti (2003) for US data. Carneiro et al. (2013) instruments maternal educational attainment with schooling cost during the mother's adolescence. Maurin and McNally (2008) is based on the series of events in May 1968 that led to the lowering of thresholds in the education system and enabled students to remain longer in the higher education system. Validity of the results from these papers hinges upon the validity and relevance of the instruments in use and may be challenged. It is also known that instrumental variable models only estimate the average treatment effect for a subpopulation of individuals (LATE of Imbens and Angrist (1994)).

The results from all these analyses are mixed. They differ in the size and statistical significance of the potential effect of the intergenerational transmission of human capital. The analysis in Holmlund et al. (2011) compares the three different identification strategies using Swedish data and finds similar patterns to the previous literature. They conclude that the differences follow from the identification, not from the different data sources. These findings stress the importance of the careful inspection of the identification strategy. As a solution to the diverging results, the analysis in de Haan (2011) studies the bounds on the average treatment effect rather than a point-identified model, and the analysis is based on weaker identifying assumptions.[1] This paper will discuss the validity and the importance of these assumptions. The analysis will consider the sensitivity of the results to some mild deviations from the identifying assumptions and to the missing data, and why the sensitivity analysis is relevant.

This paper also contributes to the literature on bounds analysis advocated by Manski (1990, 1995, 1997, 2003, 2007, 2008) by providing a way to conduct a sensitivity analysis. This paper uses the linear programming identification framework presented in Laffers (2013b), which is based on Galichon and Henry (2009a). Not only is it possible to determine which assumptions are important and drive the results but also the linear programming formulation helps to *quantify* how sensitive the results are. Note that there are other papers that consider partially identified models using linear programming; most notably, Balke and Pearl (1997, 1994), Honoré and Tamer (2006), Manski (2007), Chiburis (2010) and Freyberger and Horowitz (2012).

Section 3.2 introduces the setup and notation, and how an identification problem can be captured within a linear programming framework. Section 3.3 presents data and results, and a sensitivity analysis on the effect of mothers' schooling on children's schooling follows in Section 3.4. Section 3.5 concludes.

[1] One may argue that these assumptions are not weaker, they are just different. "Weaker" means that these assumptions are not strong enough to deliver point identification.

3.2 Method and Identifying Assumptions

3.2.1 Notation

Following the notation of Manski (1990), child j from population J has a specific response function $y_j(.)$ that maps the schooling of parent $t \in T$ (a *treatment*) to the child's schooling $y_i(t) \in Y$ (an *outcome*). For every child, we observe the schooling of his or her parent z_j (a *realized treatment*), schooling $y_j \equiv y_j(z_j)$ (a *realized outcome*), and other parent's (or grandparent's) schooling $v_j \in V$ (a *monotone instrument*), but we do not observe the child's schooling $y_j(t)$ for parents' schooling $t \neq z_j$ (a *counterfactual outcome*). The data reveal the probability distribution $P(y, z, v)$ (realized outcomes, realized treatments and instruments), yet the probability distribution of the counterfactual potential outcomes $P(y(t_1), \ldots, y(t_k))$ remains unknown.[2] The goal of the analysis is to uncover some features of the unobserved probability distribution of counterfactual outcomes $P(y(t_1), \ldots, y(t_k))$. The feature of interest may be an expectation of the child's schooling if his or her parents' schooling is equal to t ($E[y(t)]$), or it may be the average treatment effect of the change of parents' schooling from s to t on the child's schooling ($\Delta(s, t) = E[y(t)] - E[y(s)]$).

Under exogenous selection, the average treatment response to treatment t ($E[y(t)]$) is point identified, but this assumption is often not plausible, as discussed later. Depending on the strength of the maintained identifying assumptions, the expectation of children's schooling with parents' education equal to t may be set rather than point identified. There may exist an interval of values for $E[y(t)]$ so that all the values in this interval are compatible with the observed probability distribution $P(y, z, v)$ and with the identifying assumptions.

3.2.2 Method

The method of obtaining the bounds for average treatment effects follows in this subsection. For a given set of assumptions, instead of analytically deriving the bounds, we translate all the assumptions into restrictions on the joint probability distribution of the unobserved component $(y(t_1), y(t_2), ..., y(t_m))$ and the observed compo-

[2] Formally, the population forms a probability space $(I, \mathcal{F}, \mathcal{P})$, where the population of individuals I is the sample space, $\mathcal{F}$ is a set of events and P is a probability measure. Hence, the only source of randomness is the choice of individual. The individual's behavior is deterministic.

nent (y, z, v). The joint probability distribution carries complete information about the probabilistic behavior of all the variables in the model; there is nothing more that could possibly be learned.

If the outcome is a child's college attendance, and the treatment is the college attendance of a parent, so that it takes two different values (0 - no college, 1 - college), and we are interested in identifying the probability that a child will obtain a college degree if his or her parent has a college education ($E[y(1)]$), we will search in the space of probability distributions of $(y(0), y(1), y, z, v)$ that are compatible with the observed probabilities of (y, z, v), that satisfy all the identifying assumptions, and that minimize (maximize) $E[y(1)]$, which would give the lower (upper) bound. If both the assumptions and the feature of interest are linear in the joint probability distribution $(y(0), y(1), y, z, v)$, then finding a lower or upper bound corresponds to solving one linear program.

The approach of the presented identification scheme follows that of Galichon and Henry (2009a) and Ekeland et al. (2010), which was further extended in Laffers (2013b).

The linear programming method presented in this paper offers flexible identification. It is easy to add, remove or change assumptions. This paper will use this method to explore how sensitive the bounds are to some mild violations of the identifying assumptions.

The following subsections discuss how different identifying assumptions translate into restrictions on the joint probability distribution $(y(0), y(1), y, z, v)$ in the light of the following specific example from de Haan (2011).

- $y_i \in Y = \{0, 1\}$ - child's college (0 - no college, 1 - college).
- $z_i \in Z = \{0, 1\}$ - mother's (father's) college (0 - no college, 1 - college).
- $v_i \in V = \{1, 2, 3, 4\}$ - other parent's (grandparent's) schooling level (high school or less ($\leq$ 12 years), some college (13–15 years), bachelor's degree (16 years), master's degree or more ($\geq$ 17 years)).

The aim is to learn about the average treatment effect of an increase in mother's college attendance on a child's college attendance ($\Delta(0, 1) = E[y(1)] - E[y(0)]$).

3.2.3 Identifying assumptions

This subsection explains how the linear program whose extremes are the bounds on the ATE is created. The presentation of identifying assumptions begins with a discussion of how the unobserved component $(y(0), y(1))$ must be linked to the observed component (y, z, v), and it is called the **correct specification**. The marginal distribution of the joint probability distribution of $(y(0), y(1), y, z, v)$ must be the probability distribution of the observed component, and this is called **compatibility with observed probabilities**. Furthermore, the **monotone treatment response**, the **monotone treatment selection**, the **conditional monotone treatment selection** and the **monotone instrumental variable** assumptions are presented and explained. The figures associated with these assumptions elucidate how they translate into restrictions on the joint probability distribution $(y(0), y(1), y, z, v)$.

Correct Specification

The observed component (y, z, v) has to be compatible with the unobserved component $(y(0), y(1))$; that is, they are linked by $\forall j : z_j = t \implies y_j(t) = y_j$. If this assumption fails, it means that either the child's schooling level or the mother's schooling level is not correctly measured or that child j's schooling is not determined by mother's education.[3]

Figure 3.1 depicts the support of the joint probability distribution of $(y(0), y(1), y, z, v) \in Y^3 \times T \times V$. Every point in the figure represents a subpopulation of individuals. The blue circle denotes children with a college degree ($y = 1$), with a college-educated mother ($z = 1$), and with a grandparent with a high school education ($v = 1$). The unobserved counterfactual outcomes for these children are $y(0) = 1$ and $y(1) = 1$. The observed component $(y = 1, z = 1, v = 1)$ implies that $y(1) = 1$, which is compatible with the counterfactual outcomes. For a child that belongs to the subpopulation denoted by a red triangle, the observed component $(y = 0, z = 1, v = 1)$ implies that $y(1) = 0$. At the same time, the unobserved counterfactual outcomes are $y(0) = 0$ and $y(1) = 1$, and therefore, not compatible with the observed component. There must be

[3] The assumption that outcome is a deterministic function of a treatment is intrinsic in the potential outcome framework of Rubin (1974).

no such children, and the probability of the point $(0, 1, 0, 1, 1)$ must be equal to zero. Figure 3.2 shows all the points that can be assigned nonzero probabilities.

Compatibility with Observed Probabilities

The joint probability distribution of $(y(0), y(1), y, z, v)$ must be compatible with the observed probabilities of (y, z, v). In the data, 39.6% of the children do not have college education ($y = 0$), their mother has a college degree ($z = 1$), and their father has a high school education ($v = 1$) so that the probabilities in the column of $(0, 1, 1)$ sum to 0.396 as depicted in Figure 3.3.

Monotone Treatment Response

There seems to be a consensus that a child's schooling does not decrease with mother's schooling. The monotone treatment response (MTR) assumption (Manski, 1997) interprets this statement such that for every child, the schooling level is an increasing function of mother's schooling, specifically $\forall j, t_2 \geq t_1 : y_j(t_2) \geq y_j(t_1)$. The MTR assumption is a strong assumption and guarantees that the average treatment effect is nonnegative.

The MTR assumption rules out all the rows of unobservables for which $y(1) \leq y(0)$ does not hold (that is if $(y(0), y(1)) = (1, 0)$) as shown in Figure 3.4. Given the MTR assumption, there must exist no children who would obtain a college degree if their mother had not, and who would not finish college if their mother had finished college.

Monotone Treatment Selection

The assumption of monotone treatment selection (MTS) (Manski and Pepper, 2000) provides another interpretation of how a child's schooling increases with mother's schooling. Instead of assuming the selection bias away by imposing exogenous treatment selection ($\forall t_1, t_2 : E[y(t)|z = t_1] = E[y(t)|z = t_2]$) that delivers point identification, the MTS assumption restricts the direction of the selection bias.[4] The MTS assumption states that for a fixed potential mother's college attendance, children with observed college-educated parents have a weakly higher probability of graduating

[4] Ordinary least squares regression analysis assumes ETS, and it point identifies the average potential outcome: $E[y(t)] = E[y(t)|z = t]P(z = t) + E[y(t)|z \neq t]P(z \neq t) = E[y(t)|z = t]$.

from college. That is, the probability that a child with a college-educated mother obtains a college degree ($E[y(1)|z = 1]$) is higher than the *potential* probability of a child with a mother without a college degree if (counterfactually) this mother had a college education ($E[y(1)|z = 0]$). Moreover, the probability that a child with a less-educated mother finishes college ($E[y(0)|z = 0]$) is not as high as it would be for a child with a more-educated mother if (counterfactually) this mother does not have a college education ($E[y(0)|z = 1]$). The differences in these probabilities may stem from fact that higher-educated parents tend to have higher abilities, and these can be transmitted to their children, and that these parents with higher abilities create a more stimulating environment for their children. Formally, the MTS assumption is $\forall t_2 \geq t_1 : E[y(t)|z = t_2] \geq E[y(t)|z = t_1]$.

The MTS assumption restricts the space of the joint probability distribution functions of $(y(0), y(1), y, z, v)$ to those that are compatible with the corresponding set of linear constraints. Figure 3.5 shows that the probability of graduating from a college if the mother's school attainment is equal to t conditional on her having a college degree ($E[y(t)|z = 1]$, which is calculated using the probabilities in blue rectangles) is greater than or equal to the probability conditional on her not having a college degree ($E[y(t)|z = 0]$, which is calculated using the probabilities in red (dashed) rectangles). The MTS assumption states that given a mother's schooling, any difference in unobserved characteristics between college-educated and non-college-educated mothers does not make a child's probability of graduating from a college lower than that of children with higher-educated mothers.

Conditional Monotone Treatment Selection

The conditional monotone treatment selection (cMTS) assumption, formally $\forall i, t_2 \geq t_1 : E[y(t)|z = t_2, v = i] \geq E[y(t)|z = t_1, v = i]$, also states that a child's potential probability of getting into college increases with the mother's education but conditional on (and hence regardless of) the father's (or grandparent's) schooling level. The father's (or grandparent's) education is, therefore, restricted to have no impact on the direction of the selection bias due to mother's education.

Restricting the space of the joint probability distribution functions is similar to the MTS assumption, with the conditioning on events $[z = t, v = i]$ instead of $[z = t]$.

Figure 3.6 illustrates the effect of the cMTS assumption for a subpopulation with $v = 1$.

The difference between the MTS and the cMTS assumption is explained in Laffers (2013d). The distinction is similar to whether or not to include father's education into a regression as an explanatory variable as discussed in Holmlund et al. (2011). The inclusion (similar to the cMTS assumption) would imply that the effect of mother's schooling is net of assortative mating effects. On the other hand, not including father's schooling as an explanatory variable (similar to the MTS assumption) means capturing both direct effects of mother's education and indirect effects of assortative mating. As was pointed out in Laffers (2013d), when considering higher-educated mothers that "married down" to less-educated men, we have to consider any observed or unobserved factors that made these mothers self-select into such marriages. These mothers might have compensated for unobserved low ability, or the cost of finding a partner might have been high, which is true especially for older women (Lichter, 1990), and children of older women have lower cognitive skills on average (Zybert et al., 1978).

Monotone Instrumental Variable

The monotone instrumental variable (MIV) assumption (Manski and Pepper, 2000) is a weakened version of the instrumental variable assumption ($\forall i_1, i_2 : E[y(t)|v = i_1] = E[y(t)|v = i_2]$). It ensures that a child's mean potential schooling is weakly increasing in its grandparent's schooling. The MTS assumption is, in fact, a special case of the MIV assumption.

The restrictions on $(y(0), y(1), y, z, v)$ implied by the MIV assumption work in a similar way as for MTS as depicted in Figure 3.7. Given the mother's college attainment, a child's probability of graduating from college is greater for children with higher-educated grandparents.

The average treatment effect is a linear function of the joint probability distribution of $(y(0), y(1), y, z, v)$. To find the upper and lower bounds on ATE, we conduct a search in the joint probability distributions that maximizes and minimizes the average treatment effect under linear identifying constraints, which is a linear program. The resulting bounds on ATE are sharp by construction, and the identifying assumptions translate one-to-one to restrictions on the joint probability distribution; therefore, there is no information gain or loss. If there is no joint probability distribution that satisfies the constraints imposed by the identifying assumptions and is compatible with the data, then the linear program has no feasible solution, and the model can be refuted. The linear program that leads to the upper bound on ATE of an increase in mother's college education on the probability that the child finishes college is depicted in Figure 3.8, and the joint probability distribution that maximizes the ATE under the MTR+cMTS+MIV assumption is shown in Figure 3.9. Lemma 3 shows that if the identification problem takes the form of a linear program, then the identified set is an interval between the lower and upper bound.[5] The average treatment effect is a linear function of the joint probability distribution of $(y(0), y(1), y, z, v)$.

Lemma 3. *The identified set for the ATE is an interval.*

Proof. Let p denote the probability vector of the observed variables. Let $\Pi(p)$ denote the set of all joint probability distributions of the observed and unobserved components that are compatible with p and with the identifying assumptions, and let $ATE(\pi)$ be the average treatment effect when the joint probability distribution is π. Furthermore, let $ub(p) \equiv \max_{\pi \in \Pi(p)} ATE(\pi)$ be the upper bound, and let $lb(p) \equiv \max_{\pi \in \Pi(p)} ATE(\pi)$ be the lower bound on ATE under the set of identifying assumptions.

Consider a nontrivial case where $lb(p)$ and $ub(p)$ exist and $lb(p) \neq ub(p)$. It is sufficient to show that $\forall a \in (lb(p), ub(p)) : \exists \pi \in \Pi(p) : ATE(\pi) = a$. For every a, there must exist γ so that $a = \gamma lb(p) + (1-\gamma)ub(p)$. Let $\pi_{ub} = \arg\max_{\pi \in \Pi(p)} ATE(\pi)$ and $\pi_{lb} = \arg\min_{\pi \in \Pi(p)} ATE(\pi)$ denote the joint probability distributions that maximize and minimize the ATE, respectively. For $\pi_a = \gamma\pi_{ub} + (1-\gamma)\pi_{lb}$, it must hold that $\pi_a \in \Pi(p)$, because $\Pi(p)$ is defined as a set of vectors that satisfy a finite number of linear equalities and inequalities. Finally, $ATE(\pi_a) = ATE(\gamma\pi_{ub} + (1-\gamma)\pi_{lb}) = \gamma ATE(\pi_{ub}) + (1-\gamma)ATE(\pi_{lb}) = a$ because $ATE(\pi)$ is a linear function, and this completes the proof. □

The identified interval is finite, because the feasible set is bounded.

[5] The proof is very similar to that in Freyberger and Horowitz (2012).

Manski (1990, 1995, 2003) study the bounds on $E[y(t)]$ and $\Delta(s,t)$ under various combinations of these assumptions (apart from cMTS), and de Haan (2011) explains these assumptions in great detail in the context of the presented schooling application.

3.3 Data and Results

3.3.1 Data

The Wisconsin Longitudinal Study (WLS) involves a random sample of 10317 high school graduates in Wisconsin in 1957.[6] WLS also collects information from parents, spouses, and siblings of the original graduates. Similarly to de Haan (2011), this paper uses the data from the most recent surveys (2004: original respondents or their parents, 2005: siblings, 2006: spouses) and restricts the sample to the parents that have children from their first marriage, because spouses are not linked to children. Children that might still be at school (1.5%) are eliminated from the sample. Overall, the data consist of information on 21545 children.

3.3.2 Results

This paper employs the 90% confidence sets based on the bias-corrected bootstrap method of Imbens and Manski (2004), which considers the situation where the aim is to cover the unknown parameter with a fixed probability asymptotically.[7] The confidence sets are based on 500 bootstrap replications. Different statistical inference schemes, when the identified set follows from a linear programming formulation, are compared in Laffers (2013c). There is no clear winner, but the method that is used here performed well in most scenarios.

Our discussion of the results begin with the bounds on the effect of an increase in a mother's (father's) education on the probability that the child has a college degree. Table 3.1 presents the bounds for two different monotone instruments: other parent's and grandparent's schooling level, under different sets of identifying assumptions. The no-assumption bounds are not very informative, and the length of the identified

[6] Available at `http://www.ssc.wisc.edu/wlsresearch/`.

[7] The confidence sets that cover the whole identified set asymptotically are generally larger and may be preferable for a policy maker concerned with robust decisions as is argued in Henry and Onatski (2012).

interval is equal to one. The MTR assumption only affects the lower bound and sets it equal to zero with the exception that the effect of a father's attendance at college increases when the grandparent's schooling level is used as the monotone instrument, but the lower bound is not significantly different from zero under the 90% confidence level. The MTS assumption reduces the upper bound from 64.1% to 36.5% for the mother's college education and from 68.1% to 39.3% for the father's college education. The cMTS is much stronger for an increase in the mother's college attendance than for an increase in the father's, and it reduces the upper bound on the probability that the child obtains a college degree to 21.4% compared with 37.2% for the father's. The MIV only slightly affects the lower bound for the father's college attendance and has no effect on the upper bound. The monotone instrument affects the upper bound indirectly, via conditioning when the cMTS assumption is assumed. None of the sets of assumptions yields a lower bound significantly different from zero.

Table 3.2 shows the bounds on the effect of a parent's college degree on a child's years of finished schooling. No assumption bounds are not informative. The results show a similar pattern for both mother's and father's college as treatments and other parent's or grandparent's schooling level as monotone instruments. The MTR assumption increases the lower bound to zero. The MTS assumption reduces the upper bound from 10.8 years to 1.8 years when mother's college is a treatment and from 11.6 years to 1.9 years for father's college. The MIV assumption affects the upper bound only in connection with the cMTS assumption and if grandparent's schooling is used as the monotone instrument. Other parent's schooling level has greater identifying power than grandparent's schooling, and the resulting bounds are narrower. Finally, under the MTR+cMTS+MIV assumption, the effect of mother's education on child's years of completed schooling is between zero and 1.08 years or 1.52 years, respectively, when father's and grandparent's schooling level is used as the monotone instrument. The effect of father's college degree increases child's schooling by 0 to 1.43 years if mother's education is used as the MIV and 0.008 years (three days) to 1.7 years with grandparent's schooling level as the MIV. The lower bound is not statistically significant.

3.4 Sensitivity Analysis

This section studies the sensitivity of the results to relaxed identifying assumptions. The flexibility of the linear programming identification framework allows this in a straightforward manner. The identifying assumptions are relaxed in the following ways.

- Mismeasurement of Outcomes or Treatments (MOT):
  $P[z_i = t \Rightarrow y_i = y_i(t)] \geq 1 - \alpha_{MOT}$.
- Relaxed monotone treatment response (rMTR):
  $P[t_2 \geq t_1 \Rightarrow y_i(t_2) \geq y_i(t_1)] \geq 1 - \alpha_{MTR}$.
- Relaxed monotone treatment selection (rcMTS):
  $\forall z_2 \geq z_1 : E[y(t)|z = z_1] - E[y(t)|z = z_2] \leq \alpha_{cMTS}$.
- Relaxed monotone instrumental variable (rMIV):
  $\forall v_2 \geq v_1 : E[y(t)|v = v_1] - E[y(t)|v = v_2] \leq \alpha_{MIV}$.
- Missing data (MISS): at most α_{MISS}-fraction of the sample is not observed, and nothing is assumed about the nature of the missingness.

Mismeasurement of outcomes or treatments (MOT) says that for α_{MOT} fraction of the population, observed outcome y_i may not be equal to the outcome of the actual treatment z_i, either because y_i or z_i is mismeasured or because individual i's outcome is not a deterministic function of the treatment. As the data were collected mostly via phone interviews, it is reasonable to expect that some entries were not recorded correctly, although the probability of mismeasurement is likely to be low. The joint distribution that maximizes the upper bound on the ATE of mother's college degree on child's college completion under the MTR+cMTS+MIV assumption with the MOT relaxed by $\alpha_{MOT} = 0.001$ is shown in Figure 3.10.

The assumption of **relaxed monotone treatment response** (rMTR) states that α_{MTR} proportion of the population is allowed to have the outcome function that is not monotone in the treatment. The assumption that children's education is weakly increasing in mother's education is consistent with a wide range of studies. Behrman and Rosenzweig (2002) suggest that one possible channel that works in the other direction is that a more educated woman spends less time with her children.[8] The results in the literature deal with the *average* response to mother's schooling; however, it is not unreasonable to think of a small proportion of children whose schooling would not be increasing in mother's schooling. Figure 3.11 shows how relaxing the MTR assumption by $\alpha_{MTR} = 0.01$ allows up to 1% of children to respond negatively to the treatment: mother's college degree.

Relaxed conditional monotone treatment selection (rcMTS) says that the difference in mean potential outcomes between subpopulations with lower and higher observed treatments cannot be larger than δ_{cMTS} when conditioning on a value of the monotone instrument. An argument that goes against this assumption is that the outcome (child's college degree) only reflects the benefits and does not consider the cost of finishing college for the mother. A mother's college degree is an investment. If the cost of studying is very high, it may be optimal for the future mother to give up college education, and she may eventually earn more and be able to support the child's education better.

Similarly, **relaxed monotone instrumental variable** (rMIV) states that the difference in mean potential outcomes between subpopulations with lower and higher instrument values cannot be larger than δ_{MIV}.

So far, all the relaxed assumptions are straightforward modifications of the original assumptions and still linear in the joint probability distribution. This is not the case when considering the **missing data**. Even though the survey's responsiveness' rates are very good, around 90%, the fact that the data are not missing-at-random may lead to potential problems. Hauser (2005) argues that there is a systematic non-responsiveness in the studied dataset and that the missingness mechanism therefore cannot be ignored. We remain agnostic about the actual process that drives the missingness. Let $\mathcal{P}$ denote the space of all probability distributions of observed variables. If no assumptions are made about the missing data, the probability distribution of the missing component p_{MISS} can be any element in $\mathcal{P}$. The data reveals $\hat{p}_n \in \mathcal{P}$, where n is the sample size. Let α_{MISS} be the fraction of the missing part, and let $\mathcal{P}_{MISS}$ be the

[8] This analysis was challenged by Antonovics and Goldberger (2005), who claim that their results are driven by a specific data coding. In a reply, Behrman and Rosenzweig (2005) argue that their story is supported by an additional data source.

space of all probability vectors that are convex combinations of the data component $\hat{p}_n$ and arbitrary probability vector of the missing component p_{MISS}.

$$\mathcal{P}_{MISS} = \{(1-\alpha_{MISS})\hat{p}_n + \alpha_{MISS} p_{MISS} | p_{MISS} \in \mathcal{P}\} \tag{3.1}$$

To find bounds on ATE under the MISS assumption, it is necessary to calculate the minimum and the maximum ATE across all probability vectors in $\mathcal{P}_{MISS}$. The linear program takes the vector of observables p from $\mathcal{P}_{MISS}$ as fixed. The outer loop is an optimization in $\mathcal{P}_{MISS}$, which is a convex set. Note that there are different ways to model the missing data. Here the interpretation is that α_{MISS} proportion of the data is missing. No assumptions are made about the missing subpopulation separately; the identifying assumptions must hold for the whole population.[9] The following lemma states that the identified set is an interval under the MISS assumption.

Lemma 4. *Suppose that the matrices and the vector that define the equalities and the inequalities in $\Pi(p)$ are continuous in p element-wise. Then the identified set for the ATE under the missing data assumption is an interval.*

Proof. This proof uses the notation from the proof of Lemma 3. Further define $p_{\max} = \arg\max_{p \in \mathcal{P}_{MISS}} ub(p)$ and $p_{\min} = \arg\min_{p \in \mathcal{P}_{MISS}} lb(p)$. It is sufficient to show that $\forall a \in (lb(p_{\min}), ub(p_{\max}))\ \exists p \in \mathcal{P}_{\mathcal{MISS}} : \exists \pi \in \Pi(p) : ATE(p) = a$.

Firstly, note that $\mathcal{P}_{MISS}$ defined in equation 3.1 is a convex set. Consider any $p_1, p_2 \in \mathcal{P}_{MISS}$. From 3.1, there exist p_1^M and p_2^M such that $p_1 = (1-\alpha_{MISS})\hat{p}_n + \alpha_{MISS} p_1^M$ and $p_2 = (1-\alpha_{MISS})\hat{p}_n + \alpha_{MISS} p_2^M$. For any $0 < \lambda < 1$, it must hold that $\lambda p_1 + (1-\lambda) p_2 = (1-\alpha_{MISS})\hat{p}_n + \alpha_{MISS}(\lambda p_1^M + (1-\lambda) p_2^M) \in \mathcal{P}_{MISS}$ as $\lambda p_1^M + (1-\lambda) p_2^M \in \mathcal{P}$.

Secondly, Theorem 1.1 in Martin (1975) shows that $ub(p)$ (and $lb(p)$) is a continuous function of p on $\mathcal{P}_{MISS}$.

Finally, by virtue of the Intermediate Value Theorem (Munkres, 2000), the image set $ub(\mathcal{P}_{MISS})$ must contain the interval $(ub(\hat{p}_n), ub(p_{\max}))$, and the image set $lb(\mathcal{P}_{MISS})$ must contain $(lb(p_{\min}), lb(\hat{p}_n))$, and this, together with Lemma 3, completes the proof.

□

[9] Note that the nature of some identifying assumptions (e.g., the MTR assumption) are such that they must also hold for every subpopulation.

All sets of assumptions used in this paper satisfy the assumption of the continuity of the equalities and of the inequalities that define the set of feasible joint probability distributions of $(y(0), y(1), y, z, v)$.

We will now look more closely at the effect of the increae in mother's college education on the probability that the child has a college degree with father's schooling level as a monotone instrument for the sake of simplicity. The results with child's years of schooling as the treatment are qualitatively similar, and the average treatment effect has an appealing interpretation of a probability increase that a child has a college degree. Figure 3.12 illustrates the sensitivity of the bounds to different deviations from the MOT, MTR, cMTS and MIV assumptions.

Relaxing the (**MOT**) leads to the lower bound under the MTR, and the MTR+cMTS +MIV assumption remains at zero. The lower bound under the MIV assumption is linear in the relaxation parameter α_{MOT}. The upper bound under the benchmark MTR+cMTS+MIV assumption jumps from 21% to 35% when 1% of the outcomes are allowed to be mismeasured. The shape of the upper bound curve is convex. The already large upper bounds under the MTR assumption and under the MIV assumption do not respond to α_{MOT} as steeply. It seems that the stronger assumptions make the results more fragile to mild deviation from MOT.

The **MTR** assumption does not affect the upper bounds on ATE at all. The lower bound shows the same linear pattern for all studied models. This is not surprising because allowing 1% of children to respond negatively to mother's college increase cannot lead to an ATE smaller than 1%. Deviation from the **cMTS** assumption only affects the upper bound on ATE and in an exactly linear way.

The **MIV** assumption itself has weak identifying power and only affects the upper bound. If the potential probability that a child gets a college degree is not greater than 2% for children with less-educated fathers ($\alpha_{MIV} = 0.02$), then this assumption is irrelevant, and the upper bound increases to the no-assumption bound. The upper MTR+cMTS+MIV bound is not affected at all.

Figure 3.13 shows how the results are sensitive to **missing data**. The lower bound stays at zero if the MTR assumption is made. Under the MIV assumption, the lower bound is linear in the proportion of missing observations. The upper bound under the MTR+MIV assumption and under the MIV assumption is similar and is linearly increasing in α_{MISS}. The upper bound under the benchmark MTR+cMTS+MIV assumption gets less sensitive with increases in the amount of missing data, and the shape of this sensitivity curve is convex as it was for the MOT assumption.

So far, this analysis has considered the different relaxations one by one. Two different scenarios illuminate how the identifying assumptions interact. In the first, "optimistic", scenario, the assumptions are relaxed slightly. It is assumed that 1% of children may respond to mother's college negatively, that for up to 0.1% of children, the data on mother's or child's college attendance may be mismeasured and also that the potential probability of a child's getting a college degree cannot be greater by more than 1% for a child with a lower-educated mother (cMTS) and father (MIV). Such relaxations lead to bounds on the effect of mother's college on child's college from -1% to 24.36% as shown in Table 3.3. Adding the assumption that 1% of the data are missing shifts the upper bound to 28.62%. It is apparent that the missing data assumption is the most important determinant of the change in the upper bound. Assuming that 1% of the data are missing, the additional relaxations only change the upper bound from 27.31% to 28.62%.

Considering the more realistic ("pessimistic") scenario with 5% of children potentially responding negatively to mother's college increase, 1% of mismeasured data, 5% relaxation of the cMTS assumption and the MIV assumption, the effect is between -5% and 44.1%, so that the upper bound more than doubles from 21.44%, which is the upper bound for the benchmark specification. Adding that up to 10% of the data may be missing, which is the actual rate of survey responsiveness, the upper bound jumps to 53.27%.

This paragraph looks more closely at the last interesting result that the MIV does not affect the bounds if the cMTS assumption is made. The linear program formulation allows us to inspect which assumptions are most important by examining the values of the Lagrange multipliers corresponding to the identifying assumptions. Figure 3.14 shows the Lagrange multipliers that correspond to the linear restrictions that the cMTS assumption and the MIV assumption induce on the joint probability distribution. The cMTS multipliers sum to one, so these numbers also show the relative importance.[10] The cMTS with $v = 1$—that is, for the subpopulation of children with

[10] Figure 3.12 and Table 3.4 show that relaxation of α_{cMTS} translates to the upper bound one by one.

high-school educated fathers—drives the result most, and it accounts for 58.08% of the change in the upper bound. In the situation where the cMTS assumption holds only for children with fathers that have at least some college education ($v \geq 2$), the MIV assumption actually matters. Table 3.4 shows that the MIV assumption has a big impact on the upper bound by shrinking it from 46.71% to 27.54%. The Lagrange multipliers provide some insight into the source of the identifying power. Figure 3.15 indicates that the MIV restriction, which says that the potential probability of getting a college degree for a father with some college ($v = 2$) is greater than that of a child with a high-school educated father ($v = 1$) if their mother had a college degree ($E[y(1)|v = 2] \geq E[y(1)|v = 1]$), now takes the role of the omitted cMTS for children with less-educated fathers with the value of Lagrange multiplier of 0.582. Therefore, not only is it possible to see that the MIV is now important but also this highlights which part of the MIV assumption is relevant. The reason for this is that once the cMTS is not assumed for children with lower-educated fathers ($v = 1$), nothing is assumed about this large proportion of data, 58.21%, which is exactly the value of the Lagrange multiplier for the part of the MIV assumption that binds. Therefore, in this situation, the cMTS and the MIV assumptions are substitutes for each other.

3.5 Conclusion

de Haan (2011) provides a novel attempt to address an identification problem in the context of intergenerational transmission of education. The minimal identifying assumptions that she imposes do not deliver point identification, yet the bounds on the treatment effects are still informative. This paper has presented a method for finding sharp bounds on the average treatment effect via linear programming and has then used this method to show how sensitive the bounds are to mild violations of the identifying assumptions. The sensitivity analysis provides insights into the determinants of the identification. The bounds on ATE are very sensitive to missing data and possible mismeasurement of treatments or outcomes. Realistic relaxations of identifying assumptions double the upper bound on the effect of mother's college increase on the probability that a child finishes college.

The findings in this paper stress the importance of discussing the identification assumptions in great detail. Special care should be exercised with the assumptions with the greatest identifying power, and this paper has presented a method of identifying and analyzing them.

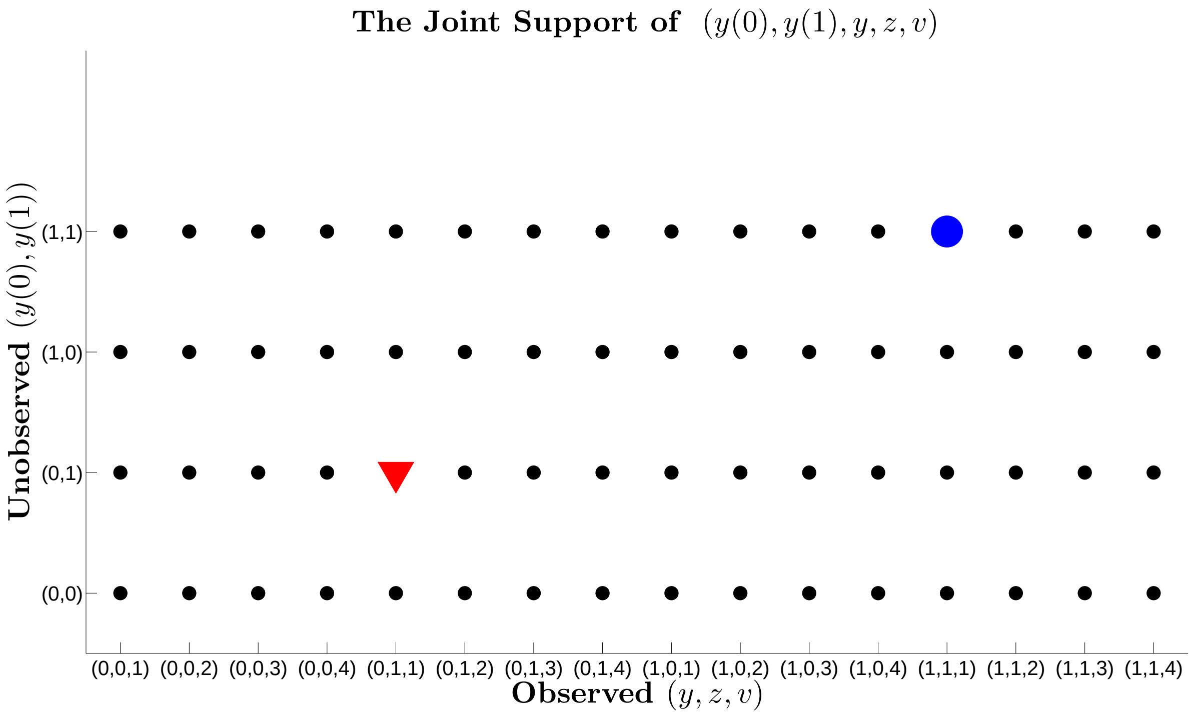


Figure 3.1: The joint support of the observed (y, z, v) and the unobserved component $(y(0), y(1))$. The large blue circle corresponds to the population with the observed $y = 1$, $z = 1$, $v = 1$, and the unobserved outcomes $y(0) = 1$ and $y(1) = 1$. The red triangle stands for the individuals with $y = 0$, $z = 1$, $v = 1$, but $y(0) = 0$ and $y(1) = 1$ so the unobserved component is not compatible with the observed component.

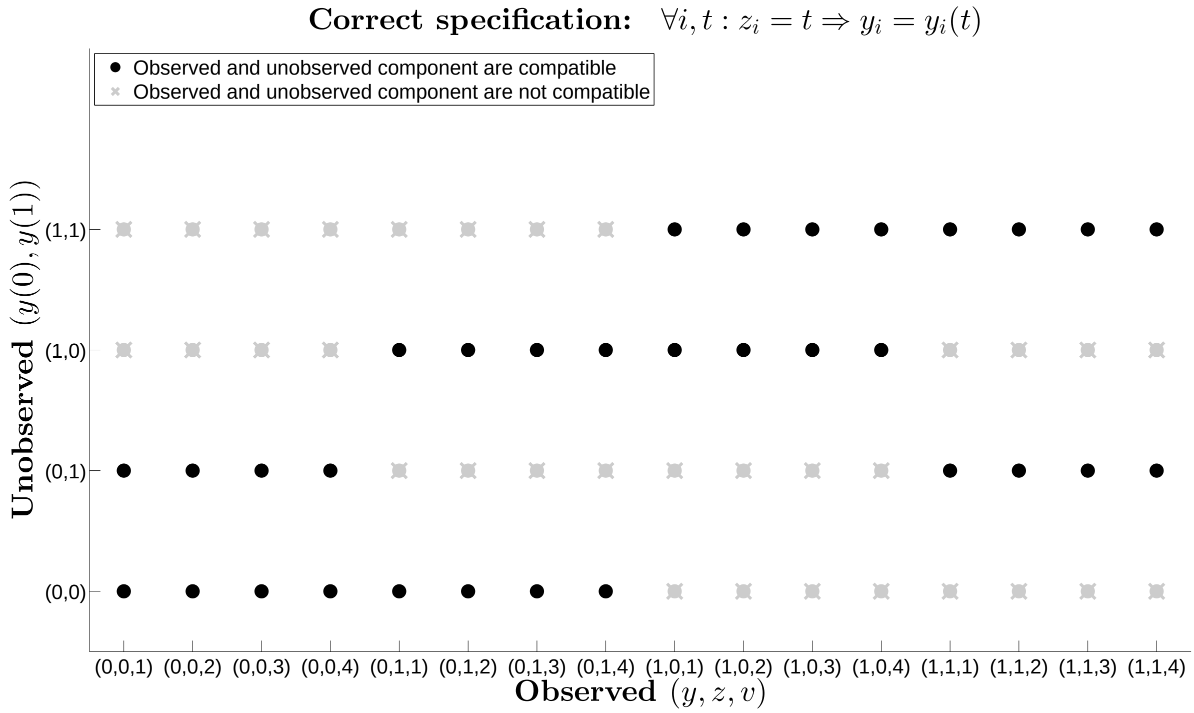


Figure 3.2: The joint support of the observed (y, z, v) and the unobserved component $(y(0), y(1))$. The grey points correspond to populations for which the unobserved component is incompatible with the observed component.

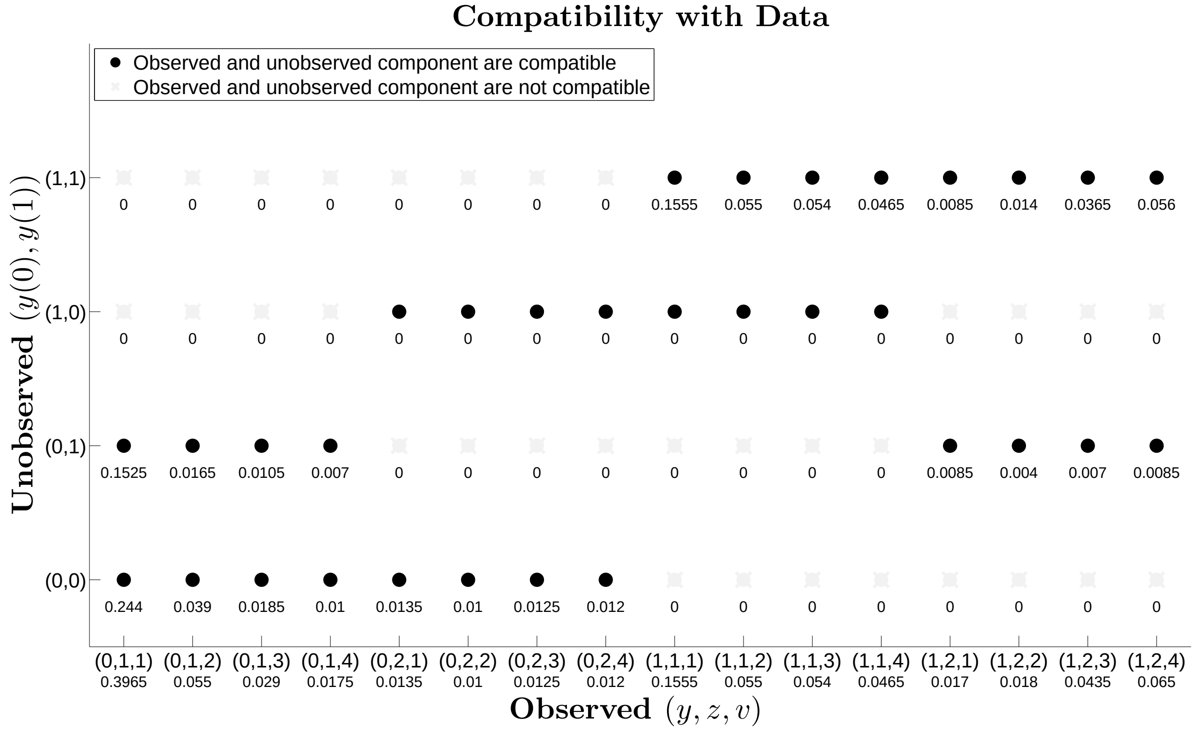


Figure 3.3: An example of the joint probability distribution of $(y(0), y(1), y, z, t)$. We observe one of its marginal distributions from the data (numbers on the horizontal axis).

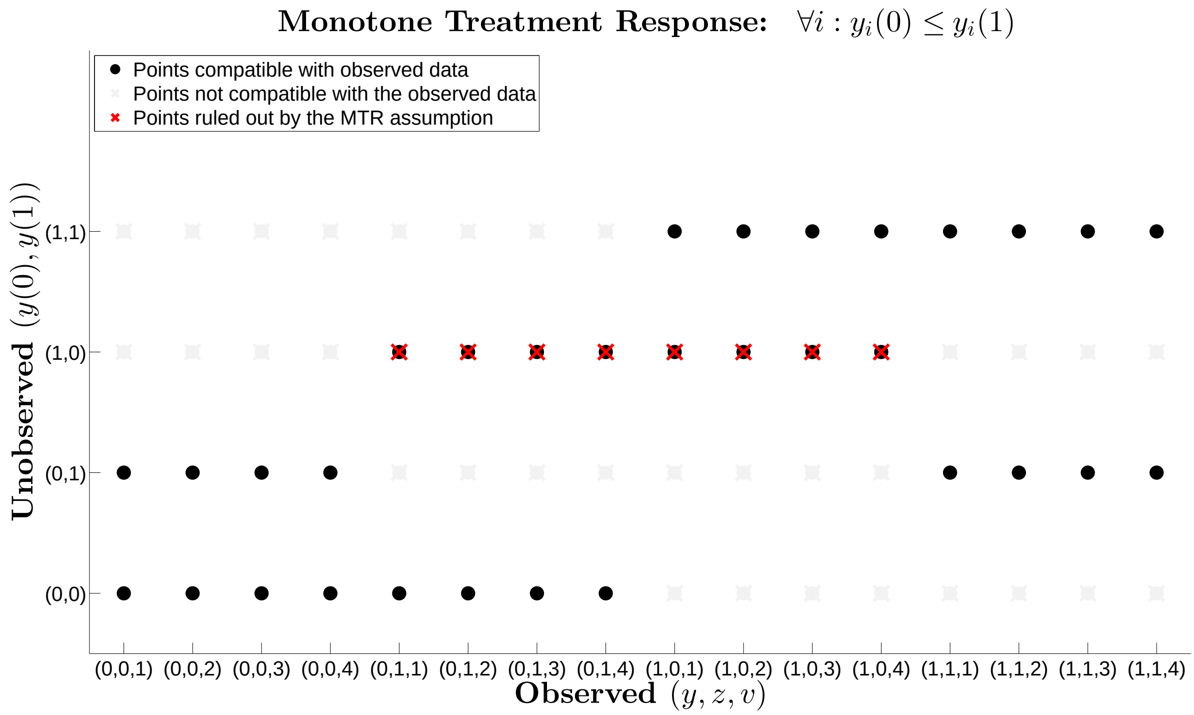


Figure 3.4: The joint support of the observed (y, z, v) and the unobserved component $(y(0), y(1))$. The MTR assumption rules out points for which $y(0) \leq y(1)$ is violated.

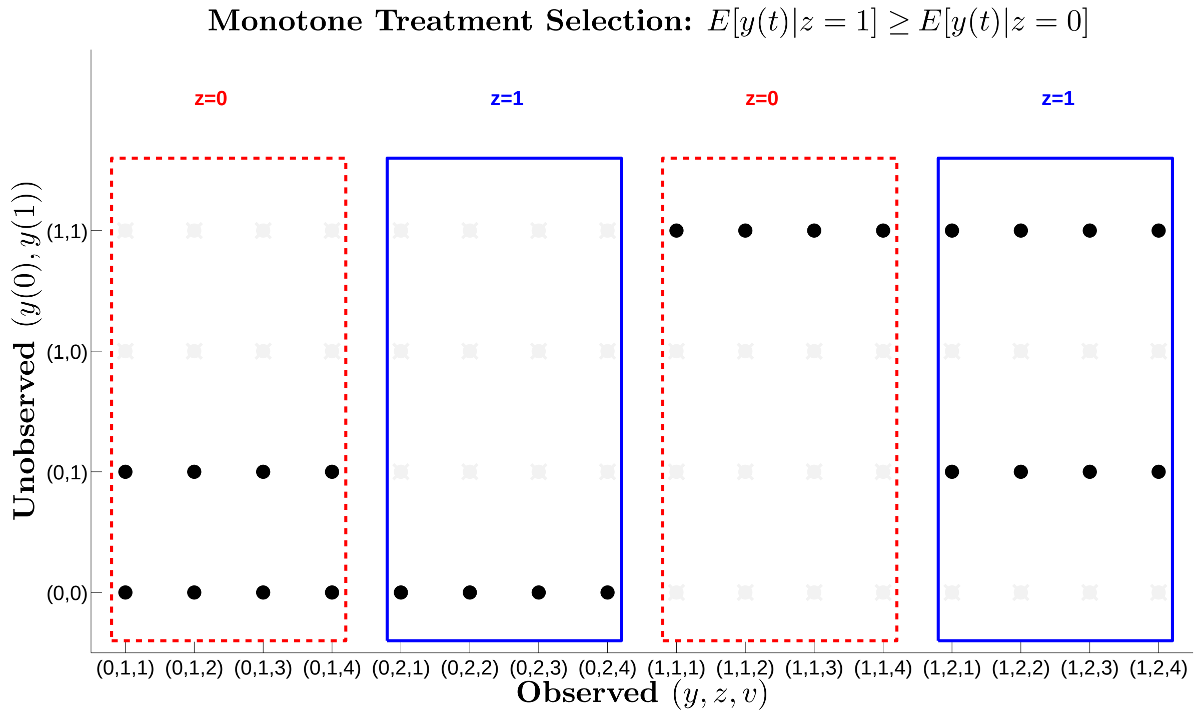


Figure 3.5: The joint support of the observed (y, z, v) and the unobserved component $(y(0), y(1))$. The MTS assumption states that the expectation of $y(t)$ based on the conditional distribution of the blue region is greater than that based on the red (dashed) region.

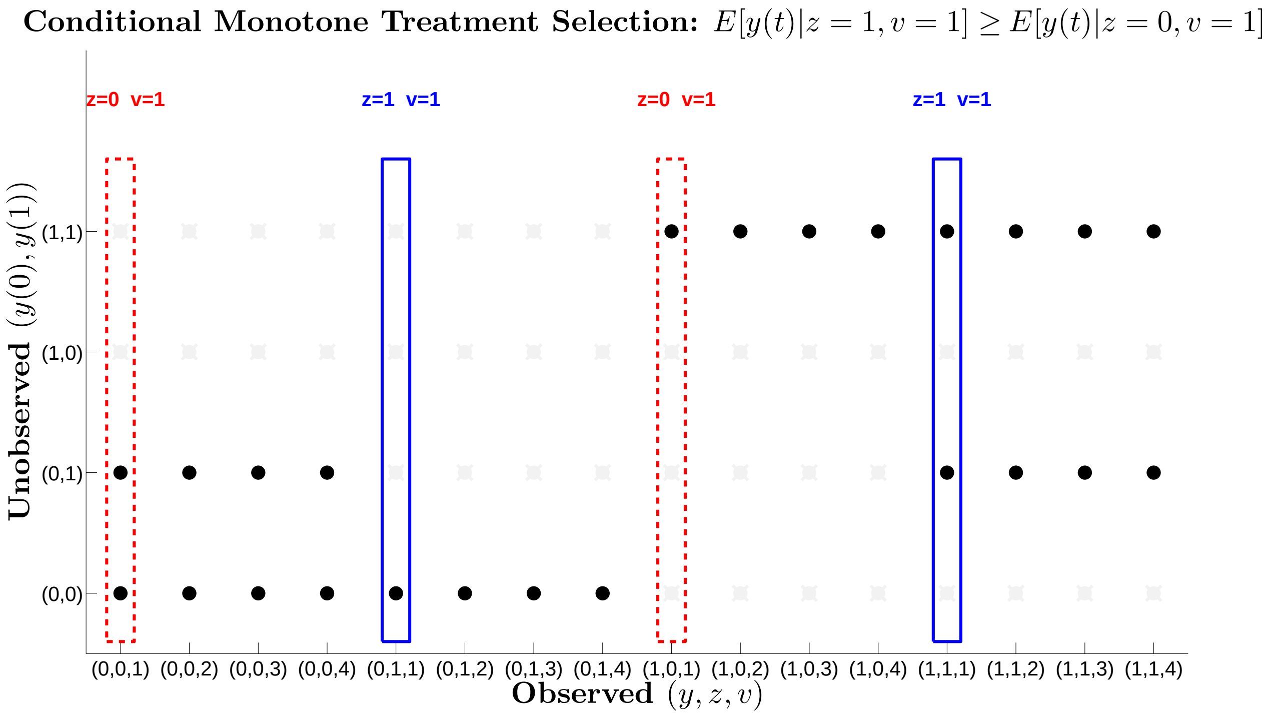


Figure 3.6: The joint support of the observed (y, z, v) and the unobserved component $(y(0), y(1))$. The cMTS assumption states that the expectation of $y(t)$ based on the conditional distribution of the blue region is greater than that based on the red (dashed) region if we condition on $v = 1$.

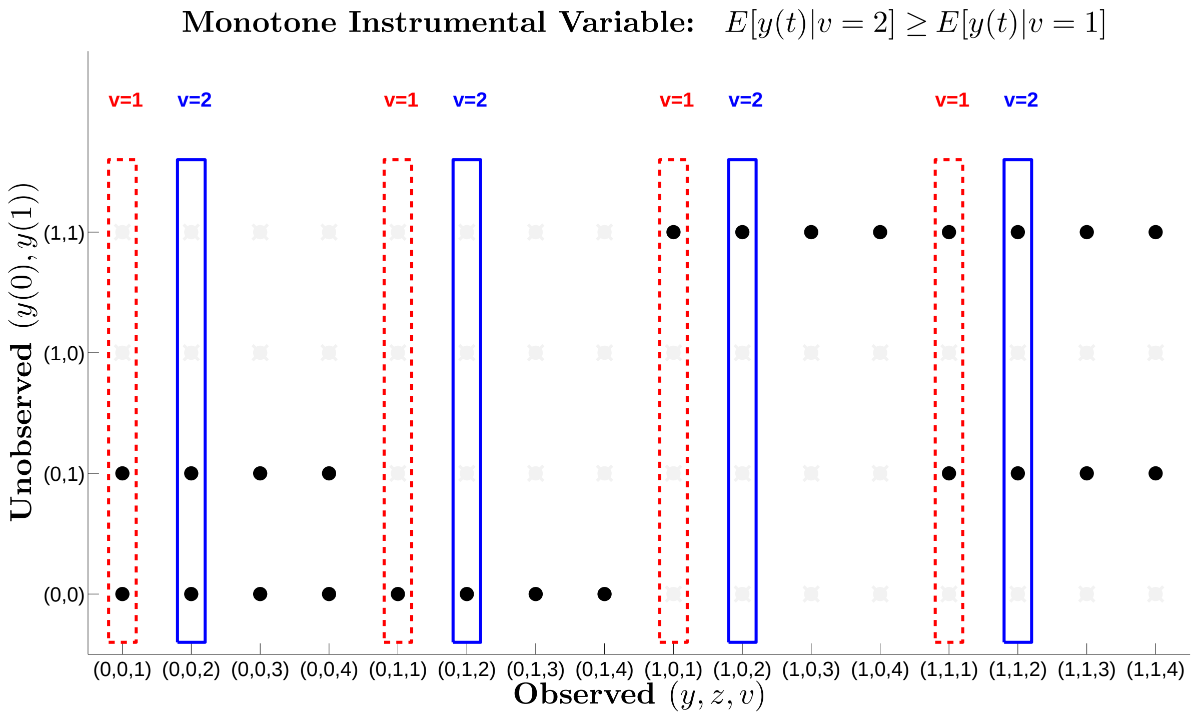


Figure 3.7: The joint support of the observed (y, z, v) and the unobserved component $(y(0), y(1))$. The MIV assumption states that the expectation of $y(t)$ based on the conditional distribution of the blue region is greater than that based on the red (dashed) region.

Bounds on the Effect of an Increase in the Mother′s (Father′s) College Education on the Probability the Child has a College Degree				
Outcome	Child′s college			
Treatment	Mother′s college		Father′s college	
Instrument	Father′s	Grandparent′s	Mother′s	Grandparent′s
No Assumption	[-35.9%, 64.1%]	[-35.9%, 64.1%]	[-31.9%, 68.1%]	[-31.9%, 68.1%]
	(-36.5%, 64.7%)	(-36.5%, 64.7%)	(-32.6%, 68.7%)	(-32.6%, 68.7%)
MTR	[0%, 64.1%]	[0%, 64.1%]	[0%, 68.1%]	[0%, 68.1%]
	(0%, 64.7%)	(0%, 64.7%)	(0%, 68.8%)	(0%, 68.7%)
MTS	[-35.9%, 36.5%]	[-35.9%, 36.5%]	[-31.9%, 39.3%]	[-31.9%, 39.3%]
	(-36.5%, 37.9%)	(-36.6%, 37.9%)	(-32.6%, 40.6%)	(-32.5%, 40.6%)
cMTS	[-35.9%, 21.4%]	[-35.9%, 33.7%]	[-31.9%, 30%]	[-31.9%, 37.2%]
	(-36.5%, 23.7%)	(-36.5%, 35.4%)	(-32.6%, 31.6%)	(-32.5%, 38.7%)
MTR+MTS	[0%, 36.5%]	[0%, 36.5%]	[0%, 39.3%]	[0%, 39.3%]
	(-0%, 37.9%)	(-0%, 37.9%)	(0%, 40.6%)	(-0%, 40.5%)
MTR+cMTS	[0%, 21.4%]	[0%, 33.7%]	[0%, 30%]	[0%, 37.2%]
	(0%, 23.7%)	(0%, 35.3%)	(0%, 31.7%)	(0%, 38.6%)
MTR+MTS+MIV	[0%, 36.5%]	[0%, 36.5%]	[0%, 39.3%]	[0.1%, 39.3%]
	(0%, 37.9%)	(-0.1%, 37.9%)	(-0.1%, 40.7%)	(-0.8%, 40.6%)
MTR+cMTS+MIV	[0%, 21.4%]	[0%, 30.6%]	[0%, 30%]	[0.1%, 34.7%]
	(0%, 23.6%)	(-0.1%, 33.3%)	(-0.1%, 31.7%)	(-0.7%, 37.1%)
Sample size	16912		14614	
90% confidence intervals in parentheses using the method of Imbens and Manski (2004)				

Table 3.1: Bounds on the effect of an increase in the parent's college education on the probability that the child has a college degree under different identifying assumptions.

Bounds on the Effect of an Increase in the Mother's (Father's) College Education
on the Years of Child' schooling

Outcome	Child's years of schooling			
Treatment	Mother's college		Father's college	
Instrument	Father's	Grandparent's	Mother's	Grandparent's
No Assumption	[-12.164, 10.836] (-12.203, 10.874)	[-12.164, 10.836] (-12.204, 10.872)	[-11.387, 11.613] (-11.43, 11.656)	[-11.387, 11.613] (-11.43, 11.652)
MTR	[0, 10.836] (0, 10.872)	[0, 10.836] (0, 10.874)	[0, 11.613] (0, 11.653)	[0, 11.613] (0, 11.655)
MTS	[-12.164, 1.809] (-12.204, 1.881)	[-12.164, 1.809] (-12.203, 1.873)	[-11.387, 1.943] (-11.43, 2.004)	[-11.387, 1.943] (-11.432, 2.002)
cMTS	[-12.164, 1.088] (-12.204, 1.184)	[-12.164, 1.651] (-12.202, 1.723)	[-11.387, 1.437] (-11.429, 1.508)	[-11.387, 1.83] (-11.428, 1.892)
MTR+MTS	[-0, 1.809] (-0, 1.875)	[0, 1.809] (0, 1.87)	[0, 1.943] (-0, 2.002)	[0, 1.943] (-0.151, 2.003)
MTR+cMTS	[0, 1.088] (-0, 1.185)	[-0, 1.651] (-0, 1.72)	[-0, 1.437] (-0, 1.513)	[0, 1.83] (-0, 1.898)
MTR+MTS+MIV	[-0, 1.809] (-0, 1.872)	[-0, 1.809] (-0.139, 1.872)	[-0, 1.943] (-0, 2.005)	[0.008, 1.943] (-0.03, 2.007)
MTR+cMTS+MIV	[0, 1.088] (-0.114, 1.185)	[0, 1.523] (-0.111, 1.658)	[0, 1.437] (-0.147, 1.509)	[0.008, 1.702] (-0.202, 1.815)
Sample size	16912		14614	

90% confidence intervals in parentheses using the method of Imbens and Manski (2004)

Table 3.2: Bounds on the effect of an increase in the parent's college education on the years of the child's schooling under different identifying assumptions.

$$\max_\pi \overbrace{\begin{bmatrix} 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \end{bmatrix}}^{\text{Average Treatment Effect}} \times \pi$$

subject to

$$\text{DATA}\left\{\left(\begin{bmatrix}
1&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&1&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&1&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&1&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&1&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&1&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&1&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&1&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&1
\end{bmatrix} \times \pi = \begin{bmatrix}
0.397\\0.055\\0.029\\0.017\\0.013\\0.01\\0.013\\0.012\\0.155\\0.055\\0.054\\0.047\\0.017\\0.018\\0.043\\0.065
\end{bmatrix}\right)\right\}\begin{matrix}\text{Observed}\\ \text{probabilities}\end{matrix}$$

$$\begin{matrix}
\text{cMTS}\left\{\begin{bmatrix}
0&0&0&0&0&0&0&0&0&0&0&0&.03&0&0&0&0&-.55&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&.03&0&0&0&0&0&-.11&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&.06&0&0&0&0&0&0&-.08&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&.08&0&0&0&0&0&0&0&-.06\\
0&.03&0&0&0&0&0&0&0&0&0&0&.03&0&0&0&-.55&-.55&0&0&0&0&0&0\\
0&0&0&.03&0&0&0&0&0&0&0&0&0&.03&0&0&0&0&-.11&-.11&0&0&0&0\\
0&0&0&0&0&.06&0&0&0&0&0&0&0&0&.06&0&0&0&0&0&-.08&-.08&0&0\\
0&0&0&0&0&0&0&.08&0&0&0&0&0&0&0&.08&0&0&0&0&0&0&-.06&-.06
\end{bmatrix}\right. \\
\text{MIV}\left\{\begin{bmatrix}
0&0&0&0&0&0&0&0&0&0&0&0&.13&-.58&0&0&0&.13&0&-.58&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&.13&-.13&0&0&0&0&.13&0&-.13&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&.14&-.13&0&0&0&0&0&.14&0&-.13\\
0&.13&0&-.58&0&0&0&0&0&0&0&0&.13&-.58&0&0&.13&.13&-.58&-.58&0&0&0&0\\
0&0&0&.13&0&-.13&0&0&0&0&0&0&0&.13&-.13&0&0&0&.13&.13&-.13&-.13&0&0\\
0&0&0&0&0&.14&0&-.13&0&0&0&0&0&0&.14&-.13&0&0&0&0&.14&.14&-.13&-.13
\end{bmatrix}\right.
\end{matrix} \times \pi \leq \begin{matrix}
\begin{bmatrix}0\\0\\0\\0\\0\\0\\0\\0\end{bmatrix}\\
\begin{bmatrix}0\\0\\0\\0\\0\\0\end{bmatrix}
\end{matrix}$$

$$\pi \geq \begin{bmatrix}0\\ \vdots\\ 0\end{bmatrix},$$

$$\pi^* = \begin{matrix}[0.244 \quad 0.152\ 0.039\ 0.016\ 0.019\ 0.010\ 0.010\ 0.007\ 0.013\ 0.010\ 0.013 \quad 0.012 \quad \ldots \\ \ldots 0.155\ 0.055\ 0.054\ 0.047\ 0.008\ 0.009\ 0.004\ 0.014\ 0.007\ 0.036\ 0.009\ 0.056]'.\end{matrix}$$

Figure 3.8: This linear program searches in the space of the joint probability distributions assigned to all combinations of the observed component (y, z, v) and the unobserved component $(y(0), y(1))$ that are compatible $(\forall i, t : z_i = t \rightarrow y_i = y_i(t))$ and satisfy the MTR assumption (as depicted in Figure 3.4). The space of the joint distributions is further restricted to satisfy the cMTS assumption and the MIV assumption, and to be compatible with the observed probabilities. The optimal solution π^* maximizes the average treatment effect.

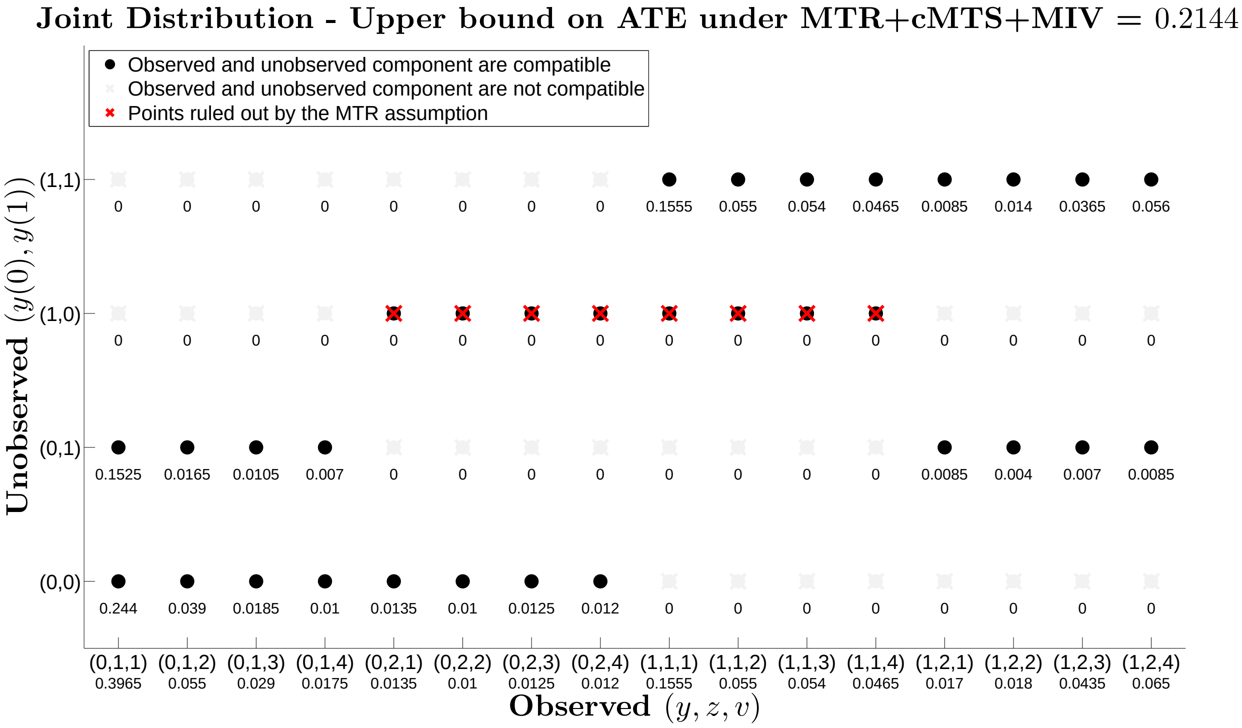


Figure 3.9: The joint probability distribution that maximizes the ATE of mother's college increase on child's probability of getting a college degree using other parent's schooling as a monotone instrumental variable under the MTR+cMTS+MIV assumption. Numbers on the horizontal axis are probabilities of the observed variables.

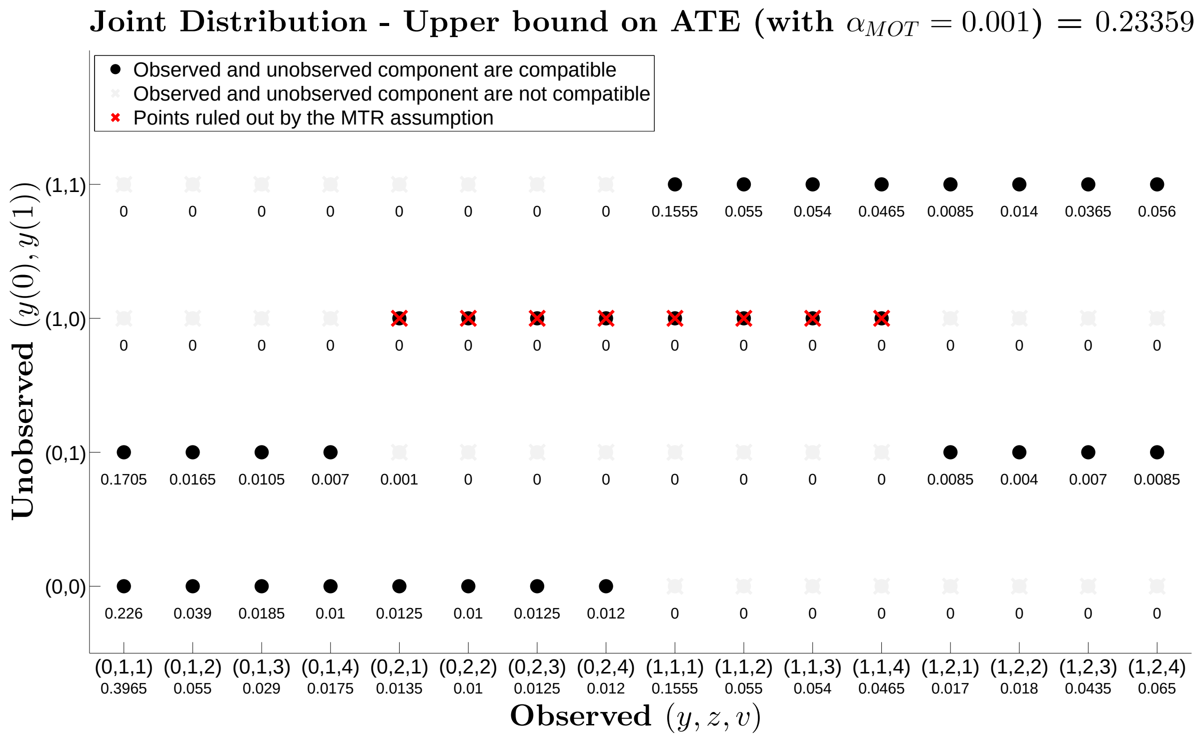


Figure 3.10: The joint probability distribution that maximizes the ATE of mother's college increase on child's probability of getting a college degree using other parent's schooling as a monotone instrumental variable. The MOT assumption is relaxed by $\alpha_{MOT} = 0.001$. We can see that this probability was assigned to the point $(0, 1, 0, 2, 1)$.

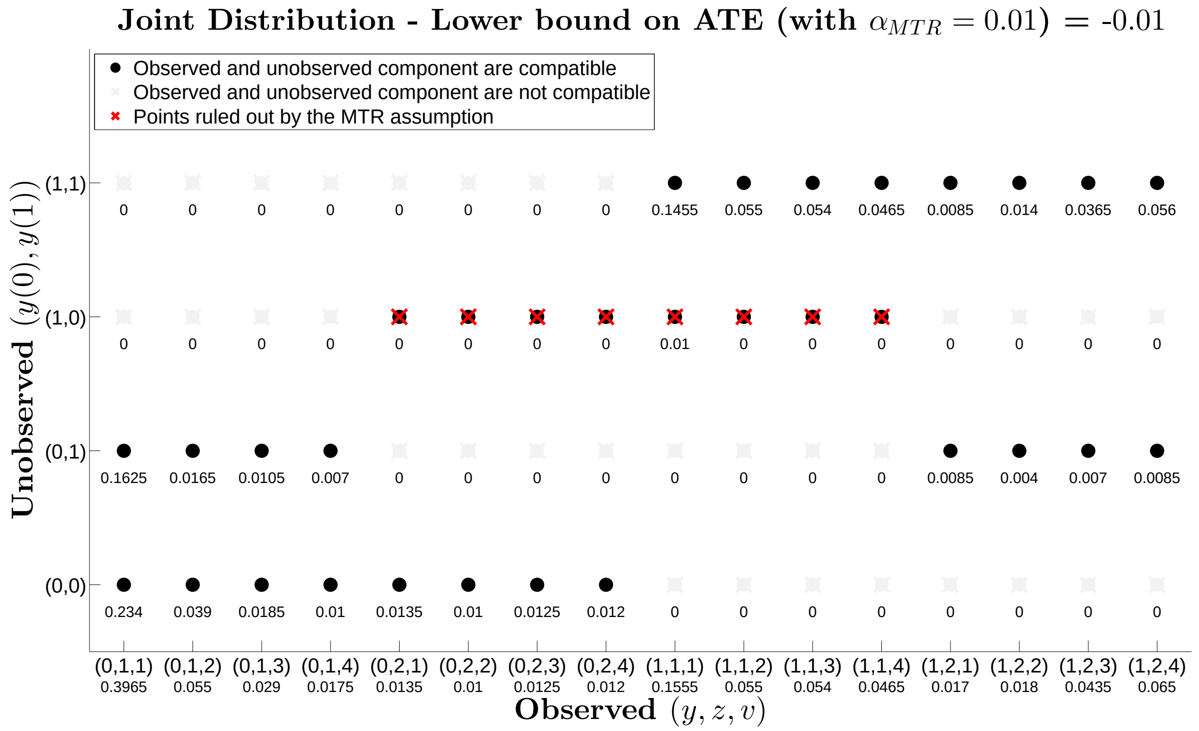


Figure 3.11: The joint probability distribution that minimizes the ATE of mother's college increase on child's probability of getting a college degree using other parent's schooling as a monotone instrumental variable. The MTR assumption does not need to hold for 1% of the children $\alpha_{MTR} = 0.01$. This 1% of children was assigned to the point $(1, 0, 1, 1, 1)$ and decreased the lower bound of ATE accordingly by 0.01.

Bounds on Effect of Mother's College Increase
on the Probability that the Child has a College Degree

MTR+cMTS+MIV

[Lower bound, Upper bound] = **[0, 21.44%]**
Confidence Set = (0, 23.74%)

	Lower bound	Upper bound			
	α_{MTR}	α_{MOT}	α_{cMTS}	α_{MIV}	α_{MISS}
Optimistic	0.01	0.001	0.01	0.01	0.01
	-1%	23.36%	22.44%	21.44%	27.31%
	(-1.46%)	(25.63%)	(24.71%)	(23.71%)	(29.64%)
Pessimistic	0.05	0.01	0.05	0.05	0.10
	-5%	35.66%	26.44%	21.44%	38.15%
	(-5%)	(37.74%)	(28.71%)	(23.71%)	(40.67%)
Optimistic	0.01	0.001	0.01	0.01	0
			$[-1\%, 24.36\%]$		
			$(-1\%, 26.63\%)$		
	0.01	0.001	0.01	0.01	0.01
			$[-1\%, 28.62\%]$		
			$(-1\%, 29.66\%)$		
Pessimistic	0.05	0.01	0.05	0.05	0
			$[-5\%, 41.54\%]$		
			$(-5\%, 43.67\%)$		
	0.05	0.01	0.05	0.05	0.10
			$[-5\%, 53.25\%]$		
			$(-5\%, 55.08\%)$		

Note: Estimates are not bias corrected, $n = 16912$

90% confidence intervals in parentheses using the method of Imbens and Manski (2004)

Table 3.3: Sensitivity analysis of the bounds on the effect of mother's college degree on the probability that the child gets a college degree. Father's education level was used as a monotone instrumental variable.

Bounds on ATE	
$MTR + cMTS$	$[0, 21.44\%]$
$MTR + cMTS + MIV$	$[0, 21.44\%]$

If cMTS holds for $v \in \{2, 3, 4\}$ only:

Bounds on ATE	
$MTR + cMTS$	$[0, 46.71\%]$
$MTR + cMTS + MIV$	$[0, 27.54\%]$

Table 3.4: Bounds on the effect of mother's college increase on the probability that the child has a college degree using father's schooling level as a monotone instrument.

Binding constraints under MTR+cMTS+MIV and Lagrange multipliers:

$$cMTS \left\{ \begin{array}{lcll} E[y(0)|z=1,v=1] & \geq & E[y(0)|z=0,v=1] & 0.0303 \\ E[y(1)|z=1,v=1] & \geq & E[y(1)|z=0,v=1] & 0.5505 \\ E[y(0)|z=1,v=2] & \geq & E[y(0)|z=0,v=2] & 0.0282 \\ E[y(1)|z=1,v=2] & \geq & E[y(1)|z=0,v=2] & 0.1106 \\ E[y(0)|z=1,v=3] & \geq & E[y(0)|z=0,v=3] & 0.0554 \\ E[y(1)|z=1,v=3] & \geq & E[y(1)|z=0,v=3] & 0.0823 \\ E[y(0)|z=1,v=4] & \geq & E[y(0)|z=0,v=4] & 0.0766 \\ E[y(1)|z=1,v=4] & \geq & E[y(1)|z=0,v=4] & 0.0637 \end{array} \right.$$

Nonbinding constraints:

$$MIV \left\{ \begin{array}{lcll} E[y(0)|v=2] & \geq & E[y(0)|v=1] & 0 \\ E[y(1)|v=2] & \geq & E[y(1)|v=1] & 0 \\ E[y(0)|v=3] & \geq & E[y(0)|v=2] & 0 \\ E[y(1)|v=3] & \geq & E[y(1)|v=2] & 0 \\ E[y(0)|v=4] & \geq & E[y(0)|v=3] & 0 \\ E[y(1)|v=4] & \geq & E[y(1)|v=3] & 0 \end{array} \right.$$

Figure 3.14: Binding and nonbinding identifying constraints under the MTR+cMTS+MIV assumption with corresponding Lagrange multipliers.

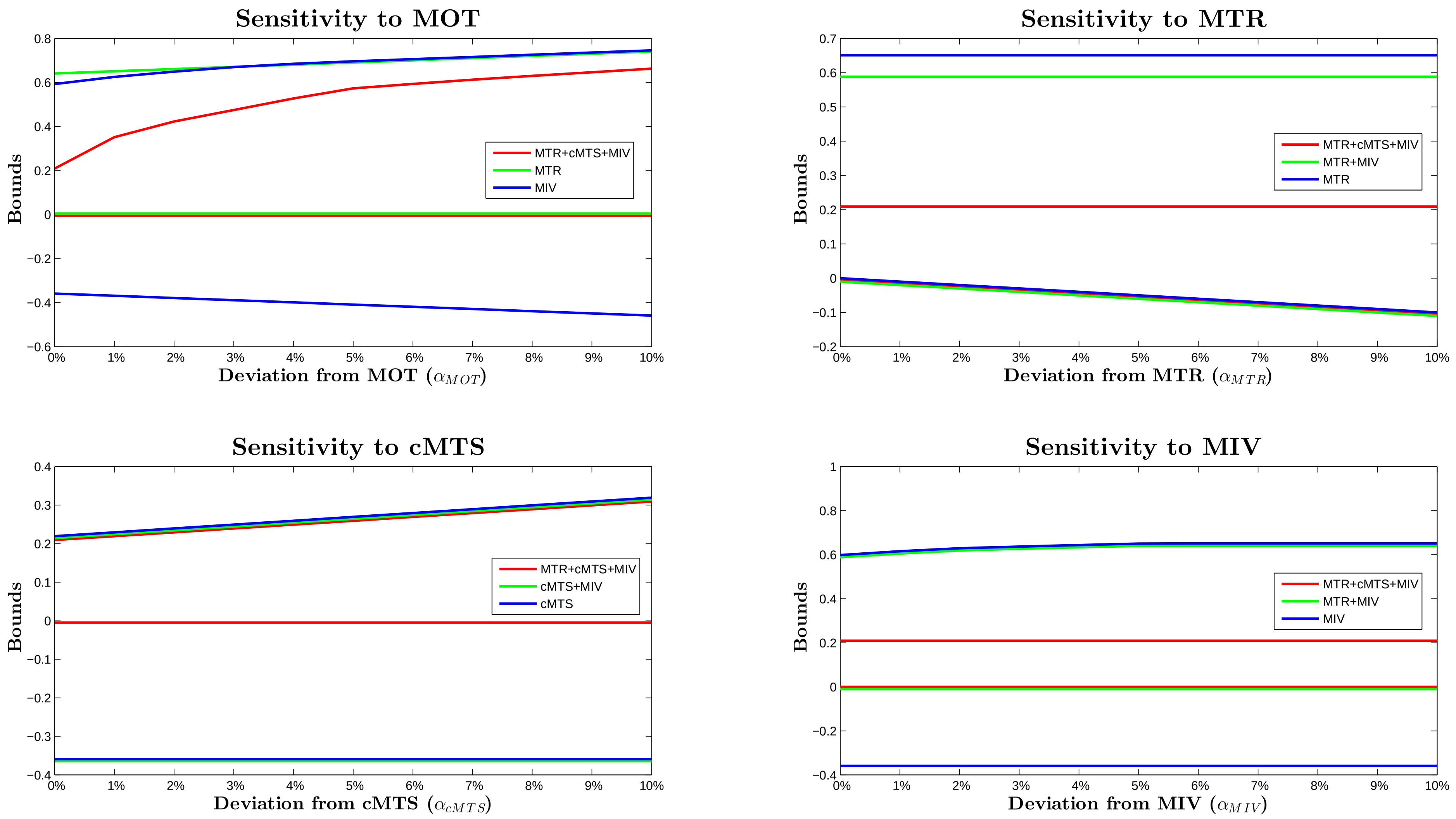


Figure 3.12: Sensitivity of the bounds on the effect of mother's college increase on probability change that child would graduate using father's schooling level as a monotone instrument.

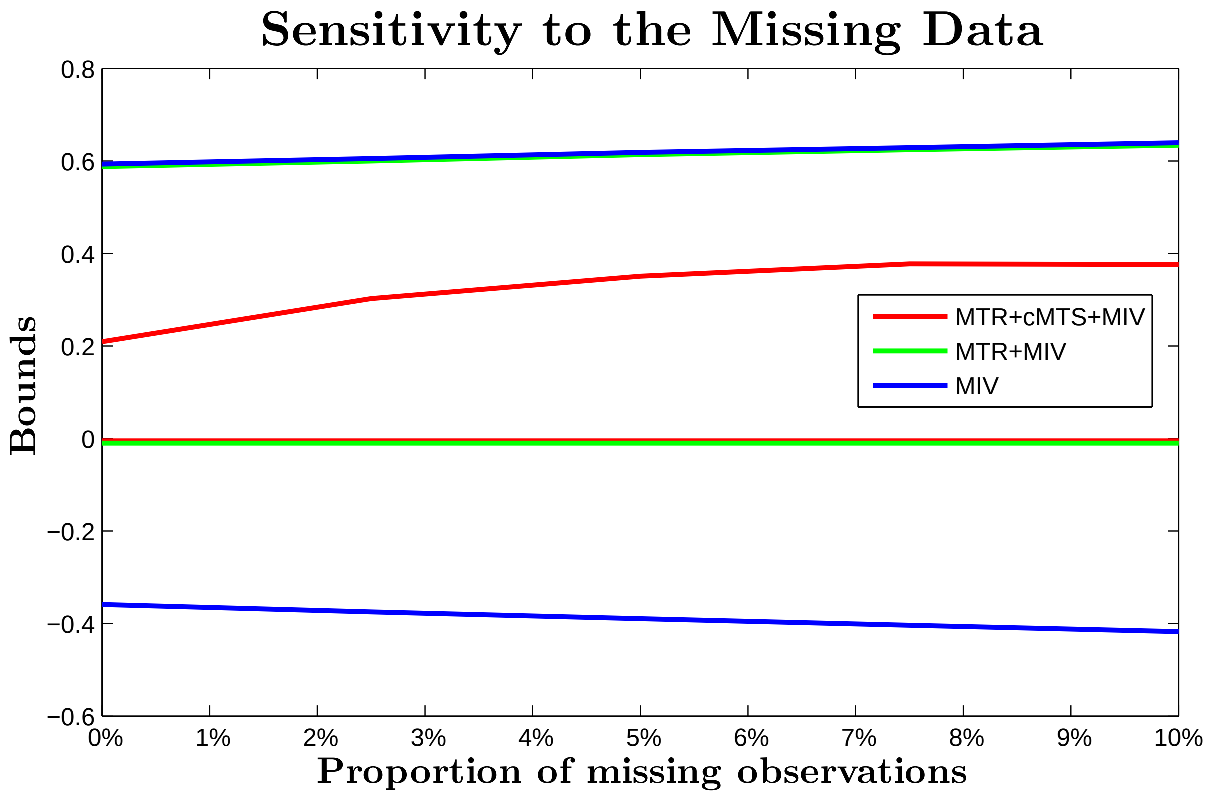


Figure 3.13: Sensitivity of the bounds on the effect of mother's college increase on probability change that child would graduate to missing data under different assumptions.

Binding constraints under MTR+cMTS+MIV:
(cMTS for $v \in \{2,3,4\}$) and Lagrange multipliers

$$\begin{array}{ll} cMTS & \left\{\begin{array}{lcll} E[y(0)|z=1,v=2] & \geq & E[y(0)|z=0,v=2] & 0.0282 \\ E[y(1)|z=1,v=2] & \geq & E[y(1)|z=0,v=2] & 0.5768 \\ E[y(0)|z=1,v=3] & \geq & E[y(0)|z=0,v=3] & 0.0554 \\ E[y(1)|z=1,v=3] & \geq & E[y(1)|z=0,v=3] & 0.0823 \\ E[y(0)|z=1,v=4] & \geq & E[y(0)|z=0,v=4] & 0.0766 \\ E[y(1)|z=1,v=4] & \geq & E[y(1)|z=0,v=4] & 0.0637 \end{array}\right. \\ MIV & \qquad E[y(1)|v=2] \quad \geq \quad E[y(1)|v=1] \qquad 0.5821 \end{array}$$

Nonbinding constraints:

$$MIV \left\{\begin{array}{lcll} E[y(0)|v=2] & \geq & E[y(0)|v=1] & 0 \\ \cancel{E[y(1)|v=2]} & \cancel{\geq} & \cancel{E[y(1)|v=1]} & \\ E[y(0)|v=3] & \geq & E[y(0)|v=2] & 0 \\ E[y(1)|v=3] & \geq & E[y(1)|v=2] & 0 \\ E[y(0)|v=4] & \geq & E[y(0)|v=3] & 0 \\ E[y(1)|v=4] & \geq & E[y(1)|v=3] & 0 \end{array}\right.$$

Figure 3.15: Binding and nonbinding identifying constraints under the MTR+cMTS+MIV assumption (cMTS for $v \in \{2,3,4\}$) with corresponding Lagrange multipliers.

Chapter 4

A NOTE ON BOUNDING AVERAGE TREATMENT EFFECTS

Abstract

The monotone treatment selection (MTS) assumption together with the monotone instrumental variable (MIV) assumption imply bounds on average treatment effects that differ from those commonly reported in the applied literature. Instead, for the bounds to be correct, we should use an MTS assumption *conditional* on the value of a monotone instrument (cMTS). In this paper, we present an empirical example of bounding the effect of mothers education on children's education, in which the MTS and cMTS assumptions lead to considerably different bounds on the treatment effects.

JEL: C4, C6, I2.

Keywords: Partial identification; Bounds; Average treatment effect.

4.1 Introduction

Different identification strategies often lead to differences in the estimated treatment effects. As a result, identifying assumptions are often a matter of some controversy. For instance, Manski (1990, 1995, 2003) suggests more credible inference based on weaker assumptions that lead to set identification. The purpose of this note is to show that monotone instrumental variable (MIV) bounds on average treatment response can only be applied to sharpen the monotone treatment selection (MTS) bounds if the MTS assumption holds conditional on the value of the instrument (cMTS).[1] As the MTS and cMTS assumptions are non-nested, they can lead to different results. In the empirical example, we find substantially different bounds on the effect of an increase in a mother's college education on the probability of a child graduating from college if the father's level of schooling serves as the MIV.

4.2 Notation and Setup

Following the notation in Manski (1990), let individual j from population J have a specific response function $y_j(.)$ that maps an ordered treatment $t \in T$ to an outcome $y \in Y$. For each individual, we observe a realized treatment z_j, a realized outcome $y_j \equiv y_j(z_j)$ and an instrument $v_j \in V$. Data reveals the probability distribution $P(y, z, v)$, yet the distribution of the counterfactual potential outcomes $P(y(.))$ remains unknown. We would then like to discern the average treatment response $E[y(t)]$ and the average treatment effect of changing treatment from s to t ($\Delta(s,t) = E[y(t)] - E[y(s)]$). The sharp lower ($LB_{E[y(t)]}$) and upper ($UB_{E[y(t)]}$) bounds on the average treatment response $E[y(t)]$ for the following identifying assumptions are proven in Manski (1997) and Manski and Pepper (2000).

- The *monotone treatment response* (MTR) assumption: $\forall j, t_2 \geq t_1 : y_j(t_2) \geq y_j(t_1)$ ensures that the outcome function for each individual j is weakly increasing in

[1] For the sake of brevity, we refer to the monotone instrument as the instrument, even though it is not an instrument in a traditional sense. Instead, it is a version of an instrument for which the mean independence restriction is relaxed (Manski and Pepper, 2000).

the treatment. The MTR assumption implies the following bounds on $E[y(t)]$:

$$E[y|z \leq t] \cdot P(z \leq t) + y_{min} \cdot P(z > t) \leq E[y(t)] \leq y_{max} \cdot P(z < t) + E[y|z \geq t] \cdot P(z \geq t). \tag{4.1}$$

- The *monotone treatment selection* (MTS) assumption: $\forall t, t_2 \geq t_1 : E[y(t)|z = t_2] \geq E[y(t)|z = t_1]$ states that individuals with higher observed treatment have either a greater or equal potential mean outcome. The MTS assumption results in

$$y_{min} \cdot P(z < t) + E[y|z = t] \cdot P(z \geq t) \leq E[y(t)] \leq E[y|z = t] \cdot P(z \leq t) + y_{max} \cdot P(z > t). \tag{4.2}$$

- The *monotone instrumental variable* (MIV) assumption: $\forall t, v_2 \geq v_1 : E[y(t)|v = v_2] \geq E[y(t)|v = v_1]$ ensures that the mean outcome is weakly increasing in the instrument value and leads to

$$\textstyle\sum_{i \in V} P(v = i)[\max_{i_1 \leq i} LB_{E[y(t)|v=i_1]}] \leq E[y(t)] \leq \sum_{i \in V} P(v = i)[\min_{i_2 \geq i} UB_{E[y(t)|v=i_2]}]. \tag{4.3}$$

The MTR and MTS assumptions combined yield the following bounds on the mean treatment response

$$E[y|z < t] \cdot P(z < t) + E[y|z = t] \cdot P(z \geq t) \leq E[y(t)] \leq E[y|z = t] \cdot P(z \leq t) + E[y|z > t] \cdot P(z > t). \tag{4.4}$$

Suppose now that we wish to bound $E[y(t)]$ using the MTR, MTS and MIV assumptions. Consider the following procedure.

Step 1: Apply the MTR+MTS bounds (4.4) on the subpopulation for which the instrument takes a specific value i to obtain the upper and lower bounds on $E[y(t)|v = i]$.

Step 2: Apply the MIV bounds (4.3) using the lower and upper bounds of $E[y(t)|v = i]$ from Step 1.

Using this procedure, we obtain the following bounds on the mean treatment response

$$\begin{aligned}&\textstyle\sum_{i\in V} P(v=i)[\max_{i_1\leq i}(E[y|z<t,v=i_1]\cdot P(z<t|v=i_1)+E[y|z=t,v=i_1]\cdot P(z\geq t|v=i_1))]\\ &\qquad\qquad\leq E[y(t)]\leq \qquad\qquad (4.5)\\ &\textstyle\sum_{i\in V} P(v=i)[\min_{i_2\geq i}(E[y|z=t,v=i_2]\cdot P(z\leq t|v=i_2)+E[y|z>t,v=i_2]\cdot P(z>t|v=i_2))].\end{aligned}$$

This note argues that this procedure, which is often used in the applied literature, need not generally yield correct bounds on $E[y(t)]$ under the MTR+MTS+MIV assumption.[2] This is because in the first step we do not obtain correct bounds on $E[y(t)|v=i]$ as the MTS assumption need not hold *conditional* on the instrument v taking the specific value i.

By using the bounds (4.5), we implicitly assume the following assumption in place of the MTS assumption

- The *conditional monotone treatment selection* (cMTS) assumption: $\forall t, i, t_2 \geq t_1$: $E[y(t)|z=t_2, v=i] \geq E[y(t)|z=t_1, v=i]$ - individuals with higher observed treatment have a greater or equal potential mean outcome conditional on the value of the instrument.

The cMTS assumption differs from the MTS assumption and may in general lead to different bounds on the average treatment response and subsequently to different bounds on the average treatment effect. The MTR+cMTS assumption leads to correct bounds on $E[y(t)|v=i]$ in the first step and subsequently correct bounds on $E[y(t)]$ in the second step. Therefore, this note asserts that bounds obtained by the described procedure actually arise from the MTR+cMTS+MIV assumption and not the MTR+MTS+MIV assumption as commonly reported. The following empirical example shows that these bounds can be substantially different.

[2] See e.g. González (2005), Gundersen and Kreider (2009), Gundersen et al. (2012), Kreider and Pepper (2007) or Kreider et al. (2012). A notable exception is Chiburis (2010), which warns that conditioning changes the meaning of the assumptions and so makes the conditioning on the value of the instrument explicit in its definition of the MTS assumption. This note differs from Chiburis (2010) in that it compares the MTR+MTS+MIV and the MTR+cMTS+MIV bounds and discusses the conditions under which the MTS and cMTS assumptions imply each other.

4.3 Empirical Illustration

Consider the problem of bounding the effect of an increase in a mother's college education on the probability of a child's college graduation using the father's level of schooling as the monotone instrument (see de Haan (2011) for a comprehensive study and data description). In the spirit of the potential outcome framework (Rubin, 1974), each child is assumed to have an individual deterministic outcome function for which we assume no interactions.[3,4]

- $y_j \in Y = \{0, 1\}$ - child's college (0 - no college, 1 - college),
- $z_j \in Z = \{0, 1\}$ - mother's college (0 - no college, 1 - college)
- $v_j \in V = \{1, 2, 3, 4\}$ - father's schooling level (high school or less (≤ 12 years), some college (13–15 years), bachelor's degree (16 years), master's degree or higher (≥ 17 years)).[5]

The MTS and cMTS assumptions have different meanings. The MTS assumption states that, for a fixed potential mother's college attendance, children with observed college-educated parents have a weakly higher probability of graduating from college. That is, the probability that a child with a college-educated mother obtains a college degree ($E[y(1)|z = 1]$) is higher than the *potential* probability of a child with a mother without a college degree if (counterfactually) this mother had a college education ($E[y(1)|z = 0]$). Moreover, the probability that a child with a less educated mother finishes college ($E[y(0)|z = 0]$) is not as high as it would be for a child with a more highly educated mother if (counterfactually) this mother had not a college education ($E[y(0)|z = 1]$). The differences in these probabilities may potentially stem from the fact that more highly educated parents tend to have greater abilities that they can transmit to their children and that these same parents create a more stimulating environment for their children.

The cMTS assumption also states that the potential probability of children obtaining a college degree increases with the mother's education but is conditional on the father's schooling level. In other words, the father's schooling level is restricted to

[3] The stable unit treatment value assumption (known as the SUTVA Assumption (Rubin, 1974)).

[4] Manski (2013) studies the identification of the treatment response with social interactions.

[5] de Haan (2011) considers the effect of father's education using mother's education as the monotone instrument.

have no impact on the direction of the selection bias due to mother's education. If we consider the subsample of women who have married less educated men, it seems less clear that the children of highly educated women have greater educational attainment. This could be due to unobserved factors that explain why these women self-select themselves into marriage with less educated men. The women who "married down" to low educated men might have done so to compensate for unobserved low ability. Also, Search theory (Becker, 1974; Oppenheimer, 1988) suggests that women tend to marry down if the cost of additional search for a partner is high, which is true especially for older women (Lichter, 1990) as the supply of potential partners decreases with time for women (Goldman et al., 1984). Therefore, we can expect these women to be on average older when having their children. Children of older mothers generally perform worse when it comes to cognitive skills (Zybert et al., 1978) and there is some empirical support for the direct causal effect of mother's age at birth on child's schooling (Kalmijn and Kraaykamp, 2005). Therefore, it is important to distinguish between the MTS and cMTS assumptions because their meanings are different and because of their consequences for the identification of the average treatment effect.

The cMTS assumption does not necessarily imply the MTS assumption and this situation is known in the literature as Simpson's paradox (Freedman et al., 2007). Subsequently, the bounds based on the MTS assumption need not be wider than the bounds based on the cMTS assumption.

If we strengthen the cMTS assumption so that regardless of the value of the instrument, the sample of individuals with higher treatment will have higher mean potential outcome, that is $\forall t, i_1, i_2, \forall t_2 \geq t_1 : E[y(t)|z = t_2, v = i_2] \geq E[y(t)|z = t_1, v = i_1]$, then the MTS assumption holds.[6] This means that conditioning on the treatment is the dominant determinant of the mean potential outcome. Under the assumption that the potential outcome $y(t)$ is independent of the instrument v, conditional on the treatment z for all values of z and t, the two assumptions are equivalent. We can think

[6] Proof:

$$E[y(t)|z = t_2] = \sum_{i \in V} E[y(t)|z = t_2, v = i]P(v = i|z = t_2) \geq \min_{i \in V} E[y(t)|z = t_2, v = i] \geq$$
$$\geq \max_{i \in V} E[y(t)|z = t_1, v = i] \geq \sum_{i \in V} E[y(t)|z = t_1, v = i]P(v = i|z = t_1) = E[y(t)|z = t_1].$$

of this as an exclusion restriction: the instrument does not affect the outcome directly, only via its explanation for the treatment.

To identify the lower and upper bounds on the average treatment effects, we conduct a search in the space of the joint probability distribution functions of $(y(0), y(1), y, z, v)$ that satisfy the MTR, MTS and MIV assumptions, and that are compatible with the probability distribution of the observed component (y, z, v). The joint distribution describes the complete probabilistic behavior of all the variables that we model; therefore, this approach guarantees the sharpness of the bounds by construction. The MTR, MTS and MIV identifying assumptions translate into *linear* restrictions on the joint distribution of $(y(0), y(1), y, z, v)$. Because the average treatment effect is also linear in the joint distribution, finding an upper bound for the treatment effect corresponds to a linear program.[7,8] The joint distribution that maximizes the average treatment effect under the MTR+MTS+MIV assumption is shown on Figure 4.1 and the corresponding linear program in Figure 4.2. Under this assumption, the upper bound on the average treatment effect is 36.5%, which is substantially greater than 21.4% – the upper bound under the MTR+cMTS+MIV assumption. Hence, the bounds reported in de Haan (2011) are too narrow if the MTS assumption (rather than the cMTS assumption) is assumed.

The results in the Table 4.1 show the bounds on the effect of an increase in the mother's college education on the probability of her child having a college degree under different menus of assumptions. We observe that the identifying power of the cMTS assumption is greater than the MTS assumption in the present application. We can also see that neither the MTS nor the cMTS assumption determines the sign of the average treatment effect and that the upper bounds differ. The MTR assumption rules out the negative effect on the average treatment effect. In addition, the MIV assumption plays no role once we assume either the MTS assumption or the cMTS assumption. However, if we did not make a distinction between the MTS and cMTS assumptions, we would falsely conclude that the MIV assumption contracted

Bounds on the Effect of an Increase in the Mother's College Education on the Probability the Child has a College Degree		
Setup	Assumptions	[Lower Bound, Upper Bound]
LinProg (this paper)	No Assumptions	[-0.358, 0.641]
LinProg (this paper)	MTS	[-0.358, 0.365]
LinProg (this paper)	cMTS	[-0.358, 0.214]
LinProg (this paper)	MTR	[0, 0.641]
LinProg (this paper)	MTR + MTS	[0, 0.365]
LinProg (this paper)	MTR + cMTS	[0, 0.214]
LinProg (this paper)	MTR + MTS + MIV	[0, 0.365]
LinProg (this paper)	MTR + cMTS + MIV	[0, 0.214]
de Haan (2011)	(MTR + MTS + MIV ?)	[0, 0.214]
Note: Estimates not bias corrected, $n = 16,912$		

Table 4.1: Presented are the bounds obtained from a search in the space of the joint distribution functions of $(y(1), y(2), y, z, v)$, that satisfy the identifying assumptions and are compatible with the observed distribution of (y, z, v). The results in the final row are from de Haan (2011) (p. 881), in which the MIV bounds (4.3) are applied to sharpen the MTR+MTS bounds (4.4) on $E[y(t)|v = i]$ for $i \in \{1, 2, 3, 4\}$.

the MTR+MTS upper bound from 36.5% to 21.4% and therefore provided an important source of identification.

4.4 Conclusion

Whenever an MIV is used to sharpen the bounds based on the MTS assumption, the latter should be stated conditional on all the values of the instrument. This applies to all past studies that employ the MTR+MTS+MIV assumption. To avoid any misunderstanding, we recommend that future studies explicitly state the conditioning on the monotone instrument when invoking the MTS assumption.

[7] A general identification scheme with examples is presented in Laffers (2013b), which builds upon the work in Galichon and Henry (2009b). The Matlab source code is available upon request.

[8] Convex combinations of the joint distributions corresponding to the lower and the upper bound yield values inside this interval.

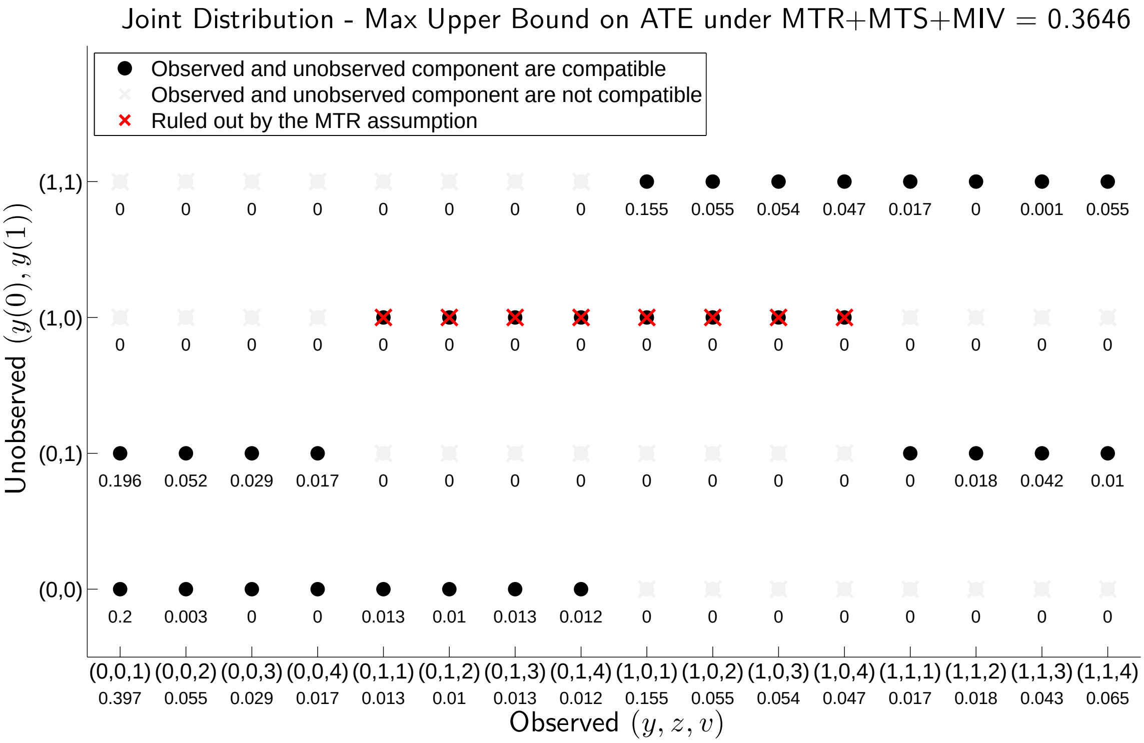


Figure 4.1: This figure depicts the joint probability distribution of the observed and unobserved components that satisfy the MTR, MTS and MIV assumptions and that are compatible with the probability distribution of the observed (y, z, v) (values under the labels on the horizontal axis are the observed probabilities). The joint distribution implies the average treatment effect of 0.365, which is higher that the upper bound of 0.214 reported in de Haan (2011). Points for which the unobserved component is not compatible with the observed component ($\forall i, t : z_i = t \nrightarrow y_i = y_i(t)$ - without dots) must be assigned zero probability, which reduces the space of the distribution functions from R_+^{64} to R_+^{32} . Also, points in the second column that correspond to the unobserved component $(y(0) = 1, y(1) = 0)$ are ruled out by the MTR assumption and this further shrinks the space of the distribution functions to R_+^{24}. The joint distribution is the optimal solution of the linear program shown in Figure 4.2.

$$\max_{\pi} \overbrace{[0\,1\,0\,1\,0\,1\,0\,1\,0\,0\,0\,0\,0\,0\,0\,0\,1\,0\,1\,0\,1\,0\,1\,0] \times \pi}^{\text{Average Treatment Effect}}$$

subject to

$$\text{DATA}\left\{\begin{bmatrix}
1\,1\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\\
0\,0\,1\,1\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\\
0\,0\,0\,0\,1\,1\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\\
0\,0\,0\,0\,0\,0\,1\,1\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\\
0\,0\,0\,0\,0\,0\,0\,0\,1\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\\
0\,0\,0\,0\,0\,0\,0\,0\,0\,1\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\\
0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,1\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\\
0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,1\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\\
0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,1\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\\
0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,1\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\\
0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,1\,0\,0\,0\,0\,0\,0\,0\,0\,0\\
0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,1\,0\,0\,0\,0\,0\,0\,0\,0\\
0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,1\,1\,0\,0\,0\,0\,0\,0\\
0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,1\,1\,0\,0\,0\,0\\
0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,1\,1\,0\,0\\
0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,1\,1
\end{bmatrix} \times \pi = \begin{bmatrix}0.397\\0.055\\0.029\\0.017\\0.013\\0.01\\0.013\\0.012\\0.155\\0.055\\0.054\\0.047\\0.017\\0.018\\0.043\\0.065\end{bmatrix}\right\}\begin{array}{c}\text{Observed}\\\text{probabilities}\end{array}$$

$$\begin{array}{l}
\text{MTS}\left\{\begin{bmatrix}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & .19 & .19 & .19 & .19 & 0 & -.80 & 0 & -.80 & 0 & -.80 & 0 & -.80\\
0 & .19 & 0 & .19 & 0 & .19 & 0 & .19 & 0 & 0 & 0 & 0 & .19 & .19 & .19 & .19 & -.80 & -.80 & -.80 & -.80 & -.80 & -.80 & -.80 & -.80
\end{bmatrix}\right. \\
\text{MIV}\left\{\begin{bmatrix}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & .13 & -.58 & 0 & 0 & 0 & .13 & 0 & -.58 & 0 & 0 & 0 & 0\\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & .13 & -.13 & 0 & 0 & 0 & 0 & .13 & 0 & -.13 & 0 & 0\\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & .14 & -.13 & 0 & 0 & 0 & 0 & 0 & .14 & 0 & -.13\\
0 & .13 & 0 & -.58 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & .13 & -.58 & 0 & 0 & .13 & .13 & -.58 & -.58 & 0 & 0 & 0 & 0\\
0 & 0 & 0 & .13 & 0 & -.13 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & .13 & -.13 & 0 & 0 & 0 & .13 & .13 & -.13 & -.13 & 0 & 0\\
0 & 0 & 0 & 0 & 0 & .14 & 0 & -.13 & 0 & 0 & 0 & 0 & 0 & 0 & .14 & -.13 & 0 & 0 & 0 & 0 & .14 & .14 & -.13 & -.13
\end{bmatrix}\right.
\end{array} \times \pi \le \begin{bmatrix}0\\0\\0\\0\\0\\0\\0\\0\end{bmatrix}$$

$$\pi \ge \begin{bmatrix}0\\\vdots\\0\end{bmatrix},$$

$$\pi^* = \begin{array}{l}[0.2\ \ 0.2\ \ 0.003\ 0.052\quad 0\quad 0.029\quad 0\quad 0.017\ 0.013\ \ 0.01\ \ 0.013\ 0.012\quad \ldots\\ \ldots\ 0.16\ 0.055\ 0.054\ 0.047\quad 0\quad 0.017\ 0.018\quad 0\quad 0.042\ 0.001\ \ 0.01\ \ 0.055]'.\end{array}$$

Figure 4.2: This linear program searches in the space of the joint probability distributions assigned to all combinations of the observed component (y, z, v) and the unobserved component $(y(0), y(1))$ that are compatible $(\forall i, t : z_i = t \rightarrow y_i = y_i(t))$ and satisfy the MTR assumption (as depicted in Figure 4.1). The space of the joint distributions is further restricted to satisfy the MTS assumption, the MIV assumption and to be compatible with the observed probabilities. The optimal solution π^* maximizes the average treatment effect.

Bibliography

ANDREWS, D. W. K. (2000): "Inconsistency of the Bootstrap when a Parameter is on the Boundary of the Parameter Space." *Econometrica*, 68, 399–405.

ANDREWS, D. W. K. AND X. SHI (2013): "Inference Based on Conditional Moment Inequalities," *Econometrica*, 81, 609–666.

ANTONOVICS, K. L. AND A. S. GOLDBERGER (2005): "Does Increasing Women's Schooling Raise the Schooling of the Next Generation? Comment," *American Economic Review*, 95, 1738–1744.

ARTSTEIN, Z. (1983): "Distributions of Random Sets and Random Selections," *Israel Journal of Mathematics*, 46, 313–324.

BABBAR, M. M. (1955): "Distributions of Solutions of a Set of Linear Equations (with an Application to Linear Programming)," *Journal of American Statistical Association*, 50, 854–869.

BALKE, A. AND J. PEARL (1994): "Counterfactual Probabilities: Computational Methods, Bounds, and Applications," in *Uncertainty in Artificial Intelligence 10*, ed. by L. R. de Mantaras and D. Poole, Morgan Kaufmann, 46–54.

——— (1997): "Bounds on treatment effects from studies with imperfect compliance," *Journal of the American Statistical Association*, 439, 1172–1176.

BECKER, G. S. (1974): "A Theory of Marriage," in *Economics of the Family: Marriage, Children, and Human Capital*, National Bureau of Economic Research, Inc, NBER Chapters, 299–351.

BEHRMAN, J. R. AND M. R. ROSENZWEIG (2002): "Does Increasing Women's Schooling Raise the Schooling of the Next Generation?" *American Economic Review*, 92, 323–334.

——— (2005): "Does Increasing Women's Schooling Raise the Schooling of the Next Generation? Reply," *American Economic Review*, 95, 1745–1751.

BERESTEANU, A., I. MOLCHANOV, AND F. MOLINARI (2011): "Sharp Identification Regions in Models With Convex Moment Predictions," *Econometrica*, 79, 1785–1821.

——— (2012): "Partial identification using random set theory," *Journal of Econometrics*, 166, 17 – 32.

BERESTEANU, A. AND F. MOLINARI (2008): "Asymptotic Properties for a Class of Partially Identified Models," *Econometrica*, 76, 763–814.

BERTAIL, P., D. POLITIS, AND J. ROMANO (1999): "On subsampling estimators with unknown rate of convergence," *Journal of American Statistical Association*, 94, 569–579.

BJÖRKLUND, A., M. LINDAHL, AND E. PLUG (2006): "The origins of intergenerational associations: Lessons from Swedish adoption data," *The Quarterly Journal of Economics*, 121, 999–1028.

BLACK, S., P. DEVEREUX, AND K. SALVANES (2005): "Why the Apple Doesn't Fall Far: Understanding Intergenerational Transmission of Human Capital," *American Economic Review*, 95, 437–42.

BROCK, W. A. AND S. N. DURLAUF (2001): "Discrete Choice with Social Interactions," *Review of Economic Studies*, 68, 235–260.

BUGNI, F. A. (2010): "Bootstrap Inference in Partially Identified Models Defined by Moment Inequalities: Coverage of the Identified Set," *Econometrica*, 78, 735–753.

——— (2011): "A Comparison of Inferential Methods in Partially Identified Models in Terms of Error in the Coverage Probability," Working paper, Department of Economics, Duke University.

Canay, I. A. (2010): “EL Inference for Partially Identified Models: Large Deviations Optimality and Bootstrap Validity,” *Journal of Econometrics*, 156, 408–425.

Carneiro, P., C. Meghir, and M. Parey (2013): “Maternal education, home environments, and the development of children and adolescents,” *Journal of the European Economic Association*, 11, 123–160.

Chernozhukov, V., C. Hansen, and M. Jansson (2009): “Finite sample inference for quantile regression models,” *Journal of Econometrics*, 152, 93–103.

Chernozhukov, V., H. Hong, and E. Tamer (2007): “Estimation and Confidence Regions for Parameter Sets in Econometric Models1,” *Econometrica*, 75, 1243–1284.

Chernozhukov, V., S. Lee, and A. M. Rosen (2013): “Intersection Bounds: Estimation and Inference,” *Econometrica*, 81, 667–737.

Chesher, A. (2009): “Single equation endogenous binary reponse models,” CeMMAP working papers CWP23/09, Centre for Microdata Methods and Practice, Institute for Fiscal Studies.

——— (2010): “Instrumental Variable Models for Discrete Outcomes,” *Econometrica*, 78, 575–601.

Chesher, A., A. M. Rosen, and K. Smolinski (2013): “An instrumental variable model of multiple discrete choice,” *Quantitative Economics*, 4, 157–196.

Chevalier, A. (2004): “Parental education and child’s education: A natural experiment,” IZA discussion paper.

Chiburis, R. C. (2010): “Bounds on Treatment Effects Using Many Types of Monotonicity,” Working paper, Department of Economics, University of Texas at Austin.

Conley, T. G., C. B. Hansen, and P. E. Rossi (2012): “Plausibly Exogenous,” *Review of Economics and Statistics*, 94, 260–272.

Currie, J. and E. Moretti (2003): “Mother’s education and the intergenerational transmission of human capital: Evidence from college openings,” *The Quarterly Journal of Economics*, 118, 1495–1532.

de Haan, M. (2011): "The Effect of Parents' Schooling on Child's Schooling: A Nonparametric Bounds Analysis," *Journal of Labor Economics*, 29, 859–892.

DiCiccio, T. J. and B. Efron (1996): "Bootstrap Confidence Intervals," *Statistical Science*, 11, 189–212.

Efron, B. (1979): "Bootstrap Methods: Another Look at the Jackknife," *Annals of Statistics*, 7, 1–26.

——— (1981): "Nonparametric standard errors and confidence intervals," *The Canadian Journal of Statistics*, 9, 139–172.

——— (1987): "Better Bootstrap Confidence Intervals," *Journal of the American Statistical Association*, 82, 171–185.

Efron, B. and R. J. Tibshirani (1993): *An Introduction to the Bootstrap*, New York: Chapman & Hall.

Ekeland, I., A. Galichon, and M. Henry (2010): "Optimal transportation and the falsifiability of incompletely specified economic models," *Economic Theory*, 42, 355–374.

Freedman, D., R. Pisani, and R. Purves (2007): *Statistics.*, W W Norton & Company Inc.

Freyberger, J. and J. Horowitz (2012): "Identification and shape restrictions in nonparametric instrumental variables estimation," CeMMAP working papers CWP15/12, Centre for Microdata Methods and Practice, Institute for Fiscal Studies.

Galichon, A. and M. Henry (2009a): "A test of non-identifying restrictions and confidence regions for partially identified parameters," *Journal of Econometrics*, 152, 186–196.

——— (2009b): "A test of non-identifying restrictions and confidence regions for partially identified parameters," *Journal of Econometrics*, 152, 186–196.

——— (2011): "Set Identification in Models with Multiple Equilibria," *Review of Economic Studies, Vol. 78, No. 4, pp. 1264-1298, 2011*.

GOLDMAN, N., C. F. WESTOFF, AND C. HAMMERSLOUGH (1984): "Demography of the marriage market in the United States," *Population Index*, 50, 5–26.

GONZÁLEZ, L. (2005): "Nonparametric bounds on the returns to language skills," *Journal of Applied Econometrics*, 20, 771–795.

GUNDERSEN, C. AND B. KREIDER (2009): "Bounding the effects of food insecurity on children's health outcomes," *Journal of Health Economics*, 28, 971–983.

GUNDERSEN, C., B. KREIDER, AND J. PEPPER (2012): "The impact of the National School Lunch Program on child health: A nonparametric bounds analysis," *Journal of Econometrics*, 166, 79–91.

HAHN, J. AND J. HAUSMAN (2005): "Estimation with Valid and Invalid Instruments," *Annals of Economics and Statistics / Annales d'Économie et de Statistique*, pp. 25–57.

HAUSER, R. M. (2005): "Survey response in the long run: The Wisconsin Longitudinal Study," *Field Methods*, 17, 3–29.

HENRY, M., R. MEANGO, AND M. QUEYRANNE (2011): "Combinatorial Bootstrap Inference in Partially Identified Incomplete Structural Models," *SSRN eLibrary*.

HENRY, M. AND I. MOURIFIÉ (2013): "Set inference in latent variables models," *The Econometrics Journal*, 16, S93–S105.

HENRY, M. AND A. ONATSKI (2012): "Set Coverage and Robust Policy," *Economics Letters*, 115, 256–257.

HOLMLUND, H., M. LINDAHL, AND E. PLUG (2011): "The Causal Effect of Parents' Schooling on Children's Schooling: A Comparison of Estimation Methods," *Journal of Economic Literature*, 49, 615–51.

HONORÉ, B. E. AND E. TAMER (2006): "Bounds on Parameters in Panel Dynamic Discrete Choice Models," *Econometrica*, 74, 611–629.

HOROWITZ, J. L. (2001): "The Bootstrap," in *Handbook of Econometrics*, ed. by J. Heckman and E. Leamer, Elsevier, vol. 5 of *Handbook of Econometrics*, chap. 52, 3159–3228.

IMBENS, G. W. AND J. D. ANGRIST (1994): "Identification and estimation of local average treatment effects," *Econometrica*, 62, 467–475.

IMBENS, G. W. AND C. F. MANSKI (2004): "Confidence Intervals for Partially Identified Parameters," *Econometrica*, 72, 1845–1857.

KALMIJN, M. AND G. KRAAYKAMP (2005): "Late or later? A sibling analysis of the effect of maternal age on children's schooling," *Social Science Research*, 34, 634–650.

KITAGAWA, T. (2012): "Estimation and Inference for Set-identified Parameters Using Posterior Lower Probability," Cemmap working papers, Centre for Microdata Methods and Practice, Institute for Fiscal Studies.

KOMAROVA, T. (2013): "Binary choice models with discrete regressors: Identification and misspecification," *Journal of Econometrics*, 177, 14 – 33.

KREIDER, B. AND J. V. PEPPER (2007): "Disability and Employment: Reevaluating the Evidence in Light of Reporting Errors," *Journal of the American Statistical Association*, 102, 432–441.

KREIDER, B., J. V. PEPPER, C. GUNDERSEN, AND D. JOLLIFFE (2012): "Identifying the Effects of SNAP (Food Stamps) on Child Health Outcomes When Participation Is Endogenous and Misreported," *Journal of the American Statistical Association*, 107, 958–975.

LAFFERS, L. (2013a): "Bounding Average Treatment Effects using Linear Programming," Working paper.

——— (2013b): "Identification in Models with Discrete Variables," NHH Dept. of Economics Discussion Paper No. 01/2013, available at http://ssrn.com/abstract=2205827.

——— (2013c): "Inference in Partially Identified Models with Discrete Variables," Working paper.

——— (2013d): "A Note on Bounding Average Treatment Effects," *Economics Letters*, 120, 424–428.

LICHTER, D. (1990): "Delayed Marriage, Marital Homogamy, and the Mate Selection Process Among White Women," *Social Science Quarterly*, 71, 802–811.

MANSKI, C. F. (1990): "Nonparametric Bounds on Treatment Effects," *American Economic Review*, 80, 319–23.

——— (1995): *Identification Problems in the Social Sciences*, Cambridge, Harvard University Press.

——— (1997): "Monotone Treatment Response," *Econometrica*, 65, 1311–1334.

——— (2003): *Partial Identification of Probability Distributions*, New York: Springer-Verlag.

——— (2007): "Partial Indentification of Counterfactual Choice Probabilities," *International Economic Review*, 48, 1393–1410.

——— (2008): "Partial Identification in Econometrics," in *The New Palgrave Dictionary of Economics*, ed. by S. N. Durlauf and L. E. Blume, Basingstoke: Palgrave Macmillan.

——— (2013): "Identification of treatment response with social interactions," *The Econometrics Journal*, 16, S1–S23.

MANSKI, C. F. AND J. V. PEPPER (2000): "Monotone Instrumental Variables, with an Application to the Returns to Schooling," *Econometrica*, 68, 997–1012.

——— (2009): "More on monotone instrumental variables," *Econometrics Journal*, 12, S200–S216.

MANSKI, C. F. AND T. S. THOMPSON (1986): "Operational characteristics of maximum score estimation," *Journal of Econometrics*, 32, 85–108.

MARTIN, D. (1975): "On the continuity of the maximum in parametric linear programming," *Journal of Optimization Theory and Applications*, 17, 205–210.

MAURIN, E. AND S. MCNALLY (2008): "Vive la Révolution! Long-Term Educational Returns of 1968 to the Angry Students," *Journal of Labor Economics*, 26, 1–33.

MOLCHANOV, I. (2005): *Theory of Random Sets*, Springer-Verlag, London.

MOON, H. R. AND F. SCHORFHEIDE (2012): "Bayesian and Frequentist Inference in Partially Identified Models," *Econometrica*, 80, 755–782.

MUNKRES, J. R. (2000): *Topology*, Prentice Hall, second ed.

NEVO, A. AND A. M. ROSEN (2012): "Identification with Imperfect Instruments," *Review of Economics and Statistics*, 93, 127–137.

OPPENHEIMER, V. K. (1988): "A Theory of Marriage Timing," *American Journal of Sociology*, 94, pp. 563–591.

OREOPOULOS, P., M. E. PAGE, AND A. H. STEVENS (2006): "The intergenerational effects of compulsory schooling," *Journal of Labor Economics*, 24, 729–760.

POLITIS, D., J. ROMANO, AND M. WOLF (1999): *Subsampling*, Springer Series in Statistics, Springer Verlag.

PREKOPA, A. (1966): "On the Probability Distribution of the Optimum of a Random Linear Program," *Journal SIAM Control*, 4, 211–222.

ROMANO, J. AND A. SHAIKH (2012): "On the uniform asymptotic validity of subsampling and the bootstrap," *Annals of Statistics*, 40, 2798–2822.

ROMANO, J. P. AND A. M. SHAIKH (2010): "Inference for the Identified Set in Partially Identified Econometric Models," *Econometrica*, 78, 169–211.

ROSEN, A. M. (2008): "Confidence sets for partially identified parameters that satisfy a finite number of moment inequalities," *Journal of Econometrics*, 146, 107–117.

RUBIN, D. B. (1974): "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology*, 66, 688–701.

SHAIKH, A. M. AND E. J. VYTLACIL (2011): "Partial Identification in Triangular Systems of Equations With Binary Dependent Variables," *Econometrica*, 79, 949–955.

STOYE, J. (2009): "More on Confidence Intervals for Partially Identified Parameters," *Econometrica*, 77, 1299–1315.

TAMER, E. T. (2010): "Partial Identification in Econometrics," *Annual Review of Economics, Vol. 2, pp. 167-195, 2010.*

Tintner, G. (1960): “A Note on Stochastic Linear Programming,” *Econometrica*, 28, 490–495.

Wagner, H. M. (1958): “On the Distribution of Solutions in Linear Programming Problem,” *Journal of the American Statistical Association*, 53, 161–163.

Woutersen, T. and J. C. Ham (2013): “Calculating Confidence Intervals for Continuous and Discontinuous Functions of Parameters,” Working paper, Department of Economics, University of Arizona.

Zybert, P., Z. Stein, and B. Lillian (1978): “Demography of the marriage market in the United States,” *Perceptual and Motor Skills*, 47, 815–818.