A General Theory of Identification

Guillaume Basse
Iavor Bojinov

Working Paper 20-086

A General Theory of Identification

Guillaume Basse
Stanford University

Iavor Bojinov
Harvard Business School

Working Paper 20-086

Funding for this research was provided in part by Harvard Business School.

A general theory of identification

Guillaume Basse
Department of MS&E and Department of Statistics
Stanford

Iavor Bojinov
Harvard Business School
Harvard University

February 14, 2020

Abstract

What does it mean to say that a quantity is identifiable from the data? Statisticians seem to agree on a definition in the context of parametric statistical models — roughly, a parameter θ in a model $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ is identifiable if the mapping $\theta \mapsto P_\theta$ is injective. This definition raises important questions: Are parameters the only quantities that can be identified? Is the concept of identification meaningful outside of parametric statistics? Does it even require the notion of a statistical model? Partial and idiosyncratic answers to these questions have been discussed in econometrics, biological modeling, and in some subfields of statistics like causal inference. This paper proposes a unifying theory of identification that incorporates existing definitions for parametric and nonparametric models and formalizes the process of identification analysis. The applicability of this framework is illustrated through a series of examples and two extended case studies.

1 Introduction

Statistical inference teaches us "how" to learn from data, whereas identification analysis explains "what" we can learn from it. Although "what" logically precedes "how," the concept of identification has received relatively less attention in the statistics community. In contrast, economists have been aware of the identification problem since at least the 30's (Frisch, 1934, Chapter 9) and have pioneered most of the research on the topic. Koopmans (1949) coined the term "identifiability" and emphasized a "clear separation between problems of statistical inference arising from the variability of finite samples, and problems of identification in which [the statistician] explore[s] the limits to which inference even from an infinite number of observations is subject."

Hurwicz (1950) and Koopmans and Reiersol (1950) formalized the intutive idea of identification and developed a general theory for statistical models. The literature then fractures, with specialized definitions of identifiability arising in different fields, including biological systems modelling (Jacquez and Perry, 1990); parametric models (Rothenberg, 1971; Hsiao, 1983; Paulino and de Bragança Pereira, 1994); ecological regression (Goodman, 1959; Cross & Manski, 2002); nonparametric models (Matzkin, 2007); causal models (Pearl, 2009; Shpitser, 2008); and nonparametric finite population models (Manski, 1989, 2009). This divergence and lack of coherent unifying theory has obfuscated some central ideas and slowed down the development of the field.

This paper proposes a general framework for studying identifiability that encompasses existing definitions as special cases. We make three main contributions. First, we study the common structure of the specialized definitions and extract a single general — and mathematically rigorous — definition of identifiability. Abstracting away the specifics of each domain allows us to recognize the commonalities and make the concepts more transparent as well as easier to extend to new settings. Second, we use our definition to develop a set of results and a systematic approach for determining whether a quantity is identifiable and, if not, what is its identification region (*i.e.,* the set of values of the quantity that are coherent with the data and assumptions). This process of *identification analysis*, formalizes ideas introduced in the literature on partial identification (Manski, 2003, 2009; Tamer, 2010). Third, we provide concrete examples of how to apply our definition in different settings and include two in-depth case studies of identification analysis.

The paper proceeeds as follows. Section 2 introduces our general theory, starting with some backgound on binary relations (Section 2.1), which are the key mathematical objects underpinning our definition of identification (Section 2.2). We illustrate the flexibility and broad applicability of our definition in Section 3 and discuss identification analysis in Section 4. Finally, we provide two case studies in Section 5.

2 General theory of identification

2.1 Background on binary relations

Let Θ and Λ be two sets. A binary relation R from Θ to Λ is a subset of the cartesian product $\Theta \times \Lambda$. For $\vartheta \in \Theta$ and $\ell \in \Lambda$, we say that ϑ is R-related to ℓ if $(\vartheta, \ell) \in R$. Following convention (Halmos, 2017), we use the notation $\vartheta R \ell$ as an abbreviation for $(\vartheta, \ell) \in R$. Below, we define four important properties that a binary relation may have (Freedman, 2015); see Lehman et al. (2010) for an in-depth discussion.

Definition 1. *A binary relation R from Θ to Λ is said to be:*

- injective *if*

$$\forall \vartheta, \vartheta' \in \Theta, \forall \ell \in \Lambda, \qquad \vartheta R \ell \textit{ and } \vartheta' R \ell \quad \Rightarrow \quad \vartheta = \vartheta'$$

- surjective *if*

$$\forall \ell \in \Lambda, \exists \vartheta \in \Theta : \quad \vartheta R \ell$$

- functional *if*

$$\forall \vartheta \in \Theta,\ \forall \ell, \ell' \in \Lambda, \qquad \vartheta R \ell \textit{ and } \vartheta R \ell' \quad \Rightarrow \quad \ell = \ell'$$

- left-total *if*

$$\forall \vartheta \in \Theta,\ \exists \ell \in \Lambda : \quad \vartheta R \ell$$

A binary relation that is both functional and left-total is called a function.

Example 1. *Let Θ be the set of prime numbers, Λ be the set of integers, and R the "divides" relation such that $\vartheta R \ell$ if ϑ divides ℓ (e.g., $3R3$, $3R6$, but 3 is not in relation with 2). In this case, R is surjective and left-total, but not injective nor functional.*

Example 2. *Let $\Theta = \mathbb{R}$, $\Lambda = \mathbb{R}$, and R be the "square" relation defined by $\vartheta R \ell$ if $\vartheta^2 = \ell$. In this case, R is left-total and functional, but it is not surjective (e.g., there is no $\vartheta \in \Theta$ such that $\vartheta R(-4)$) nor injective (e.g., $2R4$ and $-2R4$). If we instead consider $\Lambda = \mathbb{R}_{\geq 0}$, the set of all positive real numbers and 0, then R is both surjective and injective.*

Example 2, shows that the properties described in Definition 1 depend on both the binary relation and the sets Λ and Θ. Throughout this paper, whenever we refer to properties of binary relations, the dependence on Λ and Θ will always be implied.

2.2 Identification in sets and functions

We start by defining identifiability for a binary relation. The definition forms the basis of our unifying framework as all the other definitions of identifiability are obtainable by specifying appropriate Λ, Θ, and R.

Definition 2 (Identifiability)**.** *Let Θ and Λ be two sets, and R a surjective and left-total binary relation from Θ to Λ. Then,*

- *Θ is* R-identifiable *at $\ell_0 \in \Lambda$ if there exists a unique $\vartheta_0 \in \Theta$ such that $\vartheta_0 R \ell_0$;*

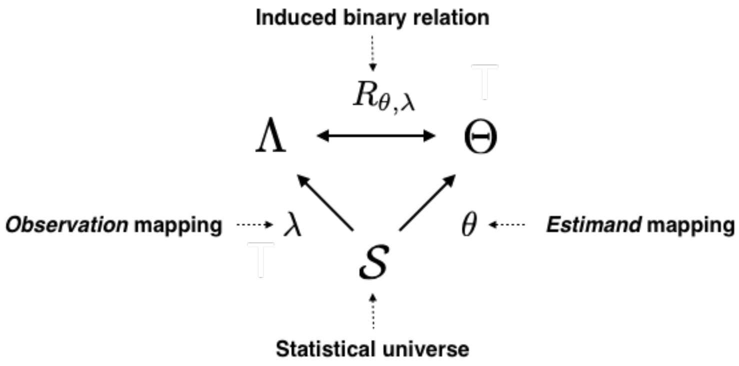

Figure 1: Diagram of the main objects.

- Θ *is* everywhere R-identifiable *in* Λ *if it is* R-identifiable *at* ℓ_0 *for all* $\ell_0 \in \Lambda$*. In this case, we usually say that* Θ *is R-identifiable.*

The distinction between R-identifiable at ℓ_0 and everywhere has important practical implications. For example, when handling missing data, the missing at random assumption aims to obtain identification at the observed missing data pattern (*i.e.*, at ℓ_0); whereas, the stronger missing always at random aims to be everywhere identifiable (Mealli and Rubin, 2015; Bojinov et al., 2020).

Formally, identifiability everywhere is equivalent to the binary relation being injective.

Proposition 1. *Let* Θ *and* Λ *two sets, and* R *a surjective and left-total binary relation from* Θ *to* Λ*.* Θ *is R-identifiable if and only if* R *is injective.*

Proof. This is a restatement of the definition. □

In most practical applications, we can derive a natural specification of the problem by working with an *induced binary relation.* Intuitively, an induced binary relation connects "what we know" to "what we are trying to learn" through a "statistical universe" in which we operate.

Definition 3 (Induced binary relation)**.** *Let* $\mathcal{S}$ *be a set and* $G(\mathcal{S})$ *be the set of all functions with domain* $\mathcal{S}$*. Let* $\lambda, \theta \in G(\mathcal{S})$*, and* $\Theta = \text{Img}(\theta)$ *and* $\Lambda = \text{Img}(\lambda)$ *their respective images. The binary relation from* Θ *to* Λ *defined as* $R_{\theta,\lambda} = \{(\theta(S), \lambda(S)), S \in \mathcal{S}\}$ *is called the* induced binary relation *associated with* (θ, λ)*.*

The examples in Section 3 show how $\mathcal{S}$, λ and θ map to real problems. In broad terms, the *statistical universe* $\mathcal{S}$ contains all the objects relevant to a given problem; the *observation mapping* λ maps $\mathcal{S}$ to "what we know"; and the *estimand mapping* θ maps $\mathcal{S}$ to "what we are trying to learn". Figure 1, illustrates how these concepts are connected.

The following proposition follows immediately from the definition of an induced binary relation.

Proposition 2. *Let $\mathcal{S}$ be a set, and $\theta, \lambda \in G(\mathcal{S})$. The induced binary relation $R_{\theta,\lambda}$ is surjective and left-total.*

Applying Definition 3 to the induced binary relation allows us to extend the notion of identification from sets to functions.

Definition 4 (Identifiability of a function)**.** *Consider $\mathcal{S}$ and $\theta, \lambda \in G(\mathcal{S})$, and let $\Theta = \mathrm{Img}(\theta)$ and $\Lambda = \mathrm{Img}(\lambda)$.*

- *The function θ is said to be identifiable at $\ell_0 \in \Lambda$ if Θ is $R_{\theta,\lambda}$-identifiable at ℓ_0. That is, for $\ell_0 \in \Lambda$ let $\mathcal{S}_0 = \{S \in \mathcal{S} : \lambda(S) = \ell_0\}$ then $R_{\theta,\lambda}$ is identifiable at ℓ_0 iff there exists $\vartheta_0 \in \Theta$, such that, for all $S \in \mathcal{S}_0$, we have that $\theta(S) = \vartheta_0$.*
- *The function θ is said to be identifiable everywhere from λ if Θ is $R_{\theta,\lambda}$-identifiable everywhere in Λ. We will usually simply say that θ is identifiable.*

Definition 4 is the workhorse allowing us to unify the notions of identifiability used in the literature for parametric and nonparametric models, as well as for finite populations.

Remark 1. *Both Definitions 2 and 4 use the adjective "identifiable" to qualify a set Θ or a mapping θ. The terminology arises naturally from the interpretation of θ and λ; indeed, we write that the estimand mapping θ is identifiable from the observation mapping λ. Proposition 1, however, makes it clear that identifiability is fundamentally a property of the binary relation R, whether we apply the adjective identifiable to θ, λ, or the whole model is mostly a matter of semantics.*

3 Identification in statistical models and finite populations

There are two significant benefits of using our framework to tackle identification in statistical models. First, the flexibility of our general formulation allows us to work directly with both parametric and nonparametric models, without having to introduce separate definitions. Second, relying on binary relations instead of functions enriches the class of questions that can be addressed through the lens of identification.

In this section, we make extensive use of examples to illustrate the broad applicability of our framework. All examples follow a common structure: first we explain the context; second, we ask an informal identification question; third, we show how to formalize the question in our framework by specifying $\mathcal{S}$, λ, and θ appropriately. The process of answering these questions, which we call *identification analysis* is described in Section 4 and illustrated in Section 5.

3.1 Parametric models

Consider a parametric model $\Lambda = \{P_\vartheta, \vartheta \in \Theta\}$, where Θ is a finite dimensional parameter space and P_ϑ is a distribution indexed by $\vartheta \in \Theta$. The standard definition of parametric identification centers around the injectivity of the parametrization (*e.g.*, Definition 11.2.2 of Casella and Berger (2002) and Definition 5.2 of Lehmann and Casella (2006)).

Definition 5 (Parametric identification). *The parameter ϑ of statistical model $\Lambda = \{P_\vartheta, \vartheta \in \Theta\}$ is said to be identifiable if the function $\vartheta \to P_\vartheta$ is injective.*

For parametric statistical models, Definition 5 is equivalent to Definition 4 with appropriately chosen statistical universe $\mathcal{S}$, observation mapping λ, and estimand mapping θ.

Theorem 1. *For a parameter set Θ, define the statistical universe to be $\mathcal{S} = \{(P_\vartheta, \vartheta), \vartheta \in \Theta\}$. Let the inference and estimand mappings $\lambda, \theta \in G(\mathcal{S})$ be $\lambda(S) = P_\vartheta$ and $\theta(S) = \vartheta$, respectively. In this setting, Definition 4 is equivalent to Definition 5.*

Proof. By construction, the induced binary relation $R_{\theta,\lambda}$ is functional and left-total; therefore, $R_{\theta,\lambda}$ is a function mapping ϑ to P_ϑ. The conclusion follows from Proposition 1 □

One of the classic textbook examples is the identification of the parameters in a linear regression.

Example 3 (Linear regression). *Consider a p-dimensional random vector $X \sim P(X)$ for some distribution P_X such that $E[X^t X]$ has rank $r < p$, where E denotes the expectation with respect to the law of X. Let $P(Y \mid X; \beta, \sigma^2) = \mathcal{N}(X^t\beta, \sigma^2)$, where $\mathcal{N}(\mu, \sigma^2)$ is the normal distribution with mean μ and variance σ^2, and let $P_{\beta,\sigma^2}(X, Y) = P(Y \mid X; \beta, \sigma^2)P(X)$.*
Question: *Are the regression parameters β and σ^2 identifiable?*
Our framework: *We can establish the identifiability of the parameter $\vartheta = (\beta, \sigma^2)$ from the joint distribution $P_\vartheta(X, Y)$ by letting $\mathcal{S} = \{(P_\vartheta, \vartheta), \vartheta \in \Theta\}$, where $\Theta = \mathbb{R} \times \mathbb{R}^+$, $\lambda(S) = P_\vartheta$, and $\theta(S) = \vartheta$.*

Even in the simple parametric setting, the benefits of the added flexibility of our general formulation become apparent when we ask more subtle questions about identifiability. For instance, using the set up of Example 3, suppose we are only interested in identifying $\beta = \phi(\vartheta)$, rather than the pair $\vartheta = (\beta, \sigma^2)$. The standard Definition 5 does not apply here, since $\beta \to P_\vartheta$ is not a function; that is, each value of β is associated with an infinite number of distributions P_ϑ, with different values of the parameter σ^2. Indeed, the key limitation with existing definitions is that they focuse on the injectivity of a function. In contrast, our framework studies the injectivity of binary relations, which need not be functional. Therefore, Definition 4 is directly applicable to studying the identifiability of β by replacing $\theta(S) = (\beta, \sigma^2)$ by $\theta(S) = \beta$; generally, it allows us to consider any parameter of the model or combinations of parameters without having to introduce new definitions.

Example 4 (Mixtures)**.** *Let* $Y_1 \sim \mathcal{N}(\mu_1, 1)$*, let* $Y_2 \sim \mathcal{N}(\mu_2, 1)$*, and* $B \sim$ *Bernoulli*(π)*.*
Question: *Which of the distributional parameters of* $Y = BY_1 + (1 - B)Y_2$ *are identifiable?*
Our framework: *For* $\vartheta = (\mu_1, \mu_2, \pi)$*, let* P_ϑ *be the normal distribution with mean* $\pi\mu_1 + (1 - \pi)\mu_2$ *and variance* $\pi^2 + (1 - \pi)^2$*. Let* $\mathcal{S} = \{(P_\vartheta, \vartheta), \vartheta \in \Theta\}$ *where* $\Theta = \mathbb{R} \times \mathbb{R} \times [0, 1]$*. The observation mapping is defined as* $\lambda(S) = P_\vartheta$ *and the estimand mappings are* $\theta_\pi(S) = \pi$*,* $\theta_{\mu_1}(S) = \mu_1$*,* $\theta_{\mu_2}(S) = \mu_2$*.The question of the identifiability of the parameters can be resolved by studying the injectivity of the binary relations* $R_{\theta_{\mu_1}, \lambda}$*,* $R_{\theta_{\mu_2}, \lambda}$*, and* $R_{\theta_\pi, \lambda}$ *as in Definition 4.*

Clearly, the three induced binary relation $R_{\theta_{\mu_1}, \lambda}$, $R_{\theta_{\mu_2}, \lambda}$, and $R_{\theta_\pi, \lambda}$ are not functional, making Definition 5 nonapplicable. Traditionally, authors have tackled this problem by proposing a separate definition for identifying a function of ϑ (*e.g.*, Paulino and de Bragança Pereira (1994)[Definition 2.4]); Basu (2006) refers to this as *partial identifiability.* Our definition of identification for functions of the parameter agrees with both Paulino and de Bragança Pereira (1994) and Basu (2006), with the added benefit of working directly for both parameters and functions of parameters without requiring additional formulation.

3.2 Nonparametric models

Many authors have recognized the limitations of traditional definitions for parametric identifiability (Definition 5) when working with nonparametric models, and have proposed specialiazed frameworks (Hurwicz, 1950; Matzkin, 2007, 2013; Pearl, 2009). Consider, for instance, the framework described by Matzkin (2007) to define identifiability. Let $\mathcal{S}$ be the set of all functions and distributions that satisfy the restriction imposed by some model $\mathcal{M}$, and assume that any $S \in \mathcal{S}$ defines a distribution of the observable variables $P(.; S)$. Similar to our general set up, Matzkin (2007)[Section 3.1] considers a function $\theta : \mathcal{S} \to \Theta$ which defines a feature of S we would like to learn about. Matzkin (2007) then proposes the following definition.

Definition 6. *For* $\vartheta_0 \in \theta(S)$*, let*

$$\Gamma(\vartheta_0, \mathcal{S}) = \{P(.; S) | S \in \mathcal{S} \text{ and } \theta(S) = \vartheta_0)\},$$

be the set of all probability distributions that satisfy the constraints of model $\mathcal{M}$*, and are consistent with* ϑ_0

and $\mathcal{S}$. Then $\vartheta_1 \in \Theta$ is identifiable if for any $\vartheta_0 \in \Theta$ such that $\vartheta_0 \neq \vartheta_1$

$$\Gamma(\vartheta_0, \mathcal{S}) \cap \Gamma(\vartheta_1, \mathcal{S}) = \emptyset$$

This definition of nonparametric identifiability can be obtained as a special case of our general definition:

Theorem 2. *Let $\mathcal{S}$ be the set of all functions and distributions that satisfy the restriction imposed by some model $\mathcal{M}$. Define $\lambda(S) = P(.;S)$, then Definition 4 is equivalent to Definition 6.*

Proof. In our notation, we can write Definition 6 as:

$$\forall \vartheta_0, \vartheta_1 \in \Theta, \quad \vartheta_0 \neq \vartheta_1 \quad \Rightarrow \quad \Gamma(\vartheta_0, \mathcal{S}) \cap \Gamma(\vartheta_1, \mathcal{S}) = \emptyset, \tag{1}$$

which is equivalent to

$$\begin{aligned}
(1) \qquad & \Longleftrightarrow \qquad \neg\Big(\Gamma(\vartheta_0, \mathcal{S}) \cap \Gamma(\vartheta_1, \mathcal{S}) = \emptyset\Big) \quad \Rightarrow \quad \neg\Big(\vartheta_0 \neq \vartheta_1\Big) \\
& \Longleftrightarrow \qquad \Gamma(\vartheta_0, \mathcal{S}) \cap \Gamma(\vartheta_1, \mathcal{S}) \neq \emptyset \quad \Rightarrow \quad \vartheta_0 = \vartheta_1 \\
& \Longleftrightarrow \qquad \exists \ell \in \Lambda : \ \vartheta_0 R_{\theta,\lambda} \ell \text{ and } \vartheta_1 R_{\theta,\lambda} \ell \quad \Rightarrow \quad \vartheta_0 = \vartheta_1
\end{aligned}$$

which is the definition of injectivity (see Definition 1). The conclusion follows from Proposition 1. □

Theorem 2 shows that Matzkin's nonparametric identification definition is a special case of our more general framework, with a specific choice of statistical universe $\mathcal{S}$ and observation mapping λ. We now provide three examples that cannot be addressed with Definition 6 and require the additional flexibity afforded by Definition 4.

Example 5 (Fixed margins problem)**.** *Consider two distributions $P_X(X)$ and $P_Y(Y)$, and denote by $P_{XY}(X,Y)$ their joint distribution. The fixed margin problem (Fréchet, 1951) asks what information the marginal distributions P_X and P_Y contain about the joint distribution P_{XY}.*

Question: *Is P_{XY} identifiable from P_X and P_Y?*

Our framework: *Let $\mathcal{S}$ be a family of joint distributions for X and Y. Let $\lambda(S) = (P_X, P_Y)$ and let $\theta(S) = P_{XY}$. The question of the identifiability of P_{XY} from P_X and P_Y can be answered by studying the injectivity of the induced mapping $R_{\theta,\lambda}$ as in Definition 4 (see Section 5.1 for a detailed treatment).*

In the first example, Definition 6 falls short by not allowing observation mappings of the form $\lambda(S) =$

(P_X, P_Y) – a problem transparently addressed by our definition. The following example describes another setting in which the same issue arises.

Example 6 (Missing data)**.** *If Y is a random variable representing a response of interest, let Z be a missing data indicator that is equal to 1 if the response Y is observed, and 0 otherwise. The observed outcome of interest is then drawn from $P(Y \mid Z = 1)$.*
Question: *Is the distribution of the missing outcomes $P(Y \mid Z = 0)$ identifiable from that of the observed outcomes $P(Y \mid Z = 1)$?*
Our framework: *Let $\mathcal{S}$ be a family of joint distributions for Z and Y, and define $\lambda(S) = (P(Y \mid Z = 1), P(Z))$, and $\theta(S) = P(Y \mid Z = 0)$. The question can be answered by studying the injectivity of the induced mapping $R_{\theta,\lambda}$ as in Definition 4.*

Example 6 shows that θ need not be the identity function: here for instance, we are interested in the conditional distribution $\theta(S) = P(Y \mid Z = 0)$. In fact, $\theta(S)$ does not even need to be a distribution: in the following example, it is a conditional expectation.

Example 7 (Ecological regression)**.** *Ecological inference is concerned with extracting individual-level information from aggregate data (King, 2013). An instance of the ecological inference problem is the ecological regression problem (Cross & Manski, 2002) which can be summarized as follows: suppose we know the distributions $P(Y \mid X)$ and $P(Z \mid X)$. What information does this give us about the expectation $E[Y \mid X, Z]$?*
Question: *Is $E[Y \mid X, Z]$ identifiable from $P(Y \mid X)$ and $P(Z \mid X)$.*
Our framework: *Let $\mathcal{S}$ be a family of joint distributions for Y, X, and Z. Define $\lambda(S) = (P(Y \mid X), P(Z \mid X))$ and $\theta(S) = E[Y \mid X, Z]$. The question can be answered by studying the injectivity of the induced mapping $R_{\theta,\lambda}$ as in Definition 4.*

3.3 Identification in finite populations

The examples presented so far asked questions about identifiability in the context of statistical models: the statistical universe, estimand mappings and observation mappings involved entire distributions (or summaries of distributions). Implicitly, this corresponds to the "infinite observations" perspective of Koopmans (1949) quoted in introduction. The missing data problem of Example 6, for instance, asks about the identifiability of $P(Y \mid Z = 0)$ from $P(Y \mid Z = 1)$. Consider instead a finite population of N units and denote by Y_i an outcome of interest for unit $i = 1, \ldots, N$. Suppose we only observe Y_i if $Z_i = 1$ and that the outcome is missing when $Z_i = 0$; formally, we observe $\{Y_i^*, Z_i\}_{i=1}^N$, where

$$Y_i^* = \begin{cases} Y_i & \text{if } Z_i = 1 \\ * & \text{otherwise.} \end{cases}$$

This is the finite population analog to Example 6. A natural identification question would be: can we identify $\tau = \overline{Y} = \sum_{i=1}^N Y_i$ from $\{Y_i^*, Z_i\}$? Neither Definition 5 nor Definition 6 apply in this setting, since there are no statistical models (parametric or nonparametric) involved — and yet the identifiability question makes intuitive sense.

Manski (2009) addresses the problem by introducing a sampling model and applying concepts of identification from nonparametric statistical models. Specifically, let $I \sim \text{Unif}(\{1, \ldots, N\})$ be a random variable selecting a unit uniformly at random in the population. The population average τ can be rewritten as $E_I[Y_I]$, where E_I is the expectation with respect to the distribution of Y_I induced by $P(I)$. Manski's approach allows us to work with $P_I(Y_I^*, Z_I)$ instead of $\{Y_i^*, Z_i\}_{i \in I}$, and to rephrase the problem as whether $E_I[Y_I]$ is identifiable from $P_I(Y_I^*, Z_I)$ — the results of Section 3.2 are now directly applicable. A downside of this approach, however, is that it changes the objective somewhat. Indeed, while τ and $E_I[Y_I]$ refer to the same quantity, $P_I(Y_I^*, Z_I)$ contains strictly less information than $\{Y_I^*, Z_I\}$, making it impossible to formulate some seemingly simple identification questions; such as, whether Y_1 can be identified from $\{Y_i^*, Z_i\}$.

By contrast, our general definition accomodates this setting by simply specifying appropriate $\mathcal{S}$, θ, and λ. Let $\mathcal{S}_Y$ be a set of possible outcome vectors, $\mathcal{S}_Z$ be a set of possible assignment vectors, and define $\mathcal{S} = \mathcal{S}_Y \times \mathcal{S}_Z$. An element $S \in \mathcal{S}$ is then a pair $(\{Y_i\}_{i=1}^N, \{Z_i\}_{i=1}^N)$. Finally, define the observation mapping $\lambda(S) = \{Y_i^*, Z_i\}_{i=1}^N$ and the estimand mapping as, for instance, $\theta(S) = \overline{Y}$ or $\theta(S) = Y_1$. The following example illustrates our framework in a slightly more involved finite-population setting.

Example 8 (Population identification of causal effects). *With N units, let each unit be assigned to one of two treatment interventions, $Z_i = 1$ for treatment and $Z_i = 0$ for control. Under the stable unit treatment value assumption (Rubin, 1980) each unit i has two potential outcomes $Y_i(1)$ and $Y_i(0)$, corresponding to the outcome of unit i under treatment and control, respectively. For each unit i, the observed outcome is $Y_i^* = Y_i(Z) = Y_i(1)Z_i + Y_i(0)(1 - Z_i)$. Let $Y(1) = \{Y_1(1), \ldots, Y_N(1)\}$ and $Y(0) = \{Y_1(0), \ldots, Y_N(0)\}$ be the vectors of potential outcomes and $Y = (Y(1), Y(0))$.*

Question: *Is $\tau(Y) = \overline{Y(1)} - \overline{Y(0)}$ identifiable from the observed data (Y^*, Z).*

Our framework: *Let $\mathcal{S}_Y = \mathbb{R}^N \times \mathbb{R}^N$ be the set of all possible values for Y, $\mathcal{S}_Z = \{0, 1\}^N$ and $\mathcal{S} = \mathcal{S}_Y \times \mathcal{S}_Z$. Take $\theta(S) = \tau(Y)$ and $\lambda(S) = (Y^*, Z)$ as the estimand and observation mapping, respectively. The question* *is then answerable by studying the injectivity of the induced binary relation $R_{\theta,\lambda}$ as in Definition 4.*

4 Identification analysis

So far, we have shown how a variety of identification questions can be formulated in our framework, but we have said nothing about how they can answered. Identification analysis is a three-steps process for answering such questions. The first step is to establish whether θ is identifiable or not (Section 4.1). For not idenfiable θ, the second step is to determine its identification region (Section 4.2). The third step is to incorporate different assumptions and assessing their impact on the structure of the identification region (Section 4.3).

4.1 Determining if θ is identifiable

The most direct—but usually challenging—approach to determine if θ is $R_{\theta,\lambda}$-identifiable is to use Definition 4. A simpler alternative is to instead show that $\theta(S)$ is a function of $\lambda(S)$ for all $S \in \mathcal{S}$; ensuring that each $\lambda(S)$ is in relation with a single $\theta(S)$.

Proposition 3. *If there exists a funtion $f : \Lambda \to \Theta$ such that $\theta(S) = f(\lambda(S))$ for all $S \in \mathcal{S}$, then θ is $R_{\theta,\lambda}$-identifiable.*

Proof. Fix $\ell_0 \in \Lambda$, let $\vartheta_0 = f(\ell_0)$ and consider $\mathcal{S}_0 = \{S \in \mathcal{S} : \lambda(S) = \ell_0\}$. For any $S \in \mathcal{S}_0$ we have that $\theta(S) = f(\lambda(S)) = f(\ell_0) = \vartheta_0$; therefore, θ is identifiable at ℓ_0. Since this holds for any $\ell_0 \in \Lambda$, θ is $R_{\theta,\lambda}$-identifiable. □

To illustrate the concepts in this section, we draw on an extended treatment of Example 6, which discusses identification in missing data. It is, for instance, easy to show that the marginal probability $\theta_1(S) = P(Z = 1)$ is identifiable by noticing that it can be written as a function of $\lambda(S) = (P(Y|Z = 1), P(Z = 1))$ and applying Proposition 3.

In many applications Proposition 3 is still difficult to apply directly either because $\theta(S)$ is a complicated function of $\lambda(S)$, or because it is not even a function of $\lambda(S)$. Either way, it is often better to first break up the estimand mapping $\theta(S)$ into simpler pieces, establish identifiability for each of them, and then leverage the fact that a function of identifiable quantities is itself identifiable.

Proposition 4. *Let $\theta, \theta_1, \theta_2 \in G(\mathcal{S})$, and f a function such that:*

$$\forall S \in \mathcal{S}, \qquad \theta(S) = f(\theta_1(S), \theta_2(S)).$$

If θ_1 is $R_{\theta_1,\lambda}$-identifiable and θ_2 is $R_{\theta_2,\lambda}$-identifiable, then θ is $R_{\theta,\lambda}$-identifiable. This trivially generalizes to $\theta_1, \dots, \theta_T \in G(\mathcal{S})$.

Proof. Fix $\ell_0 \in \Lambda$ and let $\mathcal{S}_0 = \{S \in \mathcal{S} : \lambda(S) = \ell_0\}$. Since θ_1 is $R_{\theta_1,\lambda}$-identifiable and θ_2 is $R_{\theta_2,\lambda}$-identifiable, then by Definition 4, there exists $\vartheta_1 \in \mathrm{Img}(\theta_1)$ and $\vartheta_2 \in \mathrm{Img}(\theta_2)$ such that:

$$\forall S \in \mathcal{S}_0, \quad \theta_1(S) = \vartheta_1 \ \text{ and } \ \theta_2(S) = \vartheta_2$$

and so:

$$\forall S \in \mathcal{S}_0, \quad \theta(S) = f(\theta_1(S), \theta_2(S)) = f(\vartheta_1, \vartheta_2) \equiv \vartheta_0 \in \mathrm{Img}(\theta).$$

□

In our missing data example, the quantity of interest is the average response $\theta(S) = E[Y]$. Applying the strategy described above, we can write it as a function of simpler quantities:

$$\underbrace{E[Y]}_{\theta(S)} = \underbrace{E[Y \mid Z = 1]P(Z = 1)}_{\theta_a(S)} + \underbrace{E[Y \mid Z = 0]P(Z = 0)}_{\theta_b(S)}.$$

Starting with the first term θ_a,

$$\theta_a(S) = \underbrace{E[Y \mid Z = 1]}_{\theta_2(S)}\underbrace{P(Z = 1)}_{\theta_1(S)},$$

we have already shown that θ_1 is identifiable. Another application of Proposition 3 establishes the identifiability of $\theta_2(S) = E[Y \mid Z = 1]$ and Proposition 4 stitches these results together to establish the identifiability of θ_a. The second term θ_b, which also decomposes into two parts

$$\theta_b(S) = \underbrace{E[Y \mid Z = 0]}_{\theta_3(S)}\underbrace{P(Z = 0)}_{1-\theta_1(S)},$$

is not identifiable because: although $1 - \theta_1(S)$ is identifiable (by Proposition 3), θ_3 generally is not. To see this, consider the following simple counter-example. Suppose Y is binary, the elements of $\mathcal{S}$ are then of the form $S = \{P(Y = 1 \mid Z = 1) = \alpha, P(Y = 1 \mid Z = 0) = \beta, P(Z = 1) = \gamma\}$. Recall the observation mapping $\lambda(S) = (P(Y = 1 \mid Z = 1) = \alpha, P(Z = 1) = \gamma)$. Fix $\ell_0 = (\alpha_0, \gamma_0) \in \Lambda$ and define $S_1 = (\alpha_0, \beta_1 = 0, \gamma_0)$ and $S_2 = (\alpha_0, \beta_2 = 1, \gamma_0)$. By construction, $\theta_3(S_1) = 0 \neq 1 = \theta_3(S_2)$ while $\lambda(S_1) = \lambda(S_2)$; applying Definition 4 shows that θ_3 is not identifiable. The counter-example illustrates that θ_b is generally not identifiable which implies that θ is not identifiable (except when there is no missing data $\theta_1(S) = 1$).

4.2 Finding θ's identification region

By Definition 4, we know that if an estimand mapping θ is not identifiable at ℓ_0, then there exists at least two values $\vartheta_1, \vartheta_2 \in \Theta$ such that $\vartheta_1 R_{\theta,\lambda} \ell_0$ and $\vartheta_2 R_{\theta,\lambda} \ell_0$. The second step in identification analysis is to determine the set of all $\vartheta \in \Theta$ such that $\vartheta R_{\theta,\lambda} \ell_0$. This set is generally called the identification region of θ at ℓ_0 (Manski, 1990; Imbens and Manski, 2004; Romano and Shaikh, 2008).

Definition 7. *Consider $\mathcal{S}$ and $\theta, \lambda \in G(\mathcal{S})$. We define the identification region of θ at $\ell_0 \in \Lambda$ as:*

$$H\{\theta; \ell_0\} \equiv R^{-1}_{\theta,\lambda}(\ell_0) \subseteq \Theta$$

where:

$$R^{-1}_{\theta,\lambda}(\ell_0) \equiv \{\vartheta \in \Theta : \vartheta R_{\theta,\lambda} \ell_0\} = \{\theta(S) : S \in \mathcal{S}, \lambda(S) = \ell_0\}$$

is the pre-image of ℓ_0 in Θ.

Informally, the identification region is the set of all values of the estimand that are equally compatible with the observation ℓ_0. If an estimand mapping θ is identifiable at ℓ_0, then a single estimand is compatible with ℓ_0 and the identification region reduces to a singleton.

Proposition 5. *For $\theta, \lambda \in G(\mathcal{S})$, we have that*

- *θ is $R_{\theta,\lambda}$-identifiable at $\ell_0 \in \Lambda$ if and only if $H\{\theta; \ell_0\} = \{\vartheta_0\}$ for some $\vartheta_0 \in \Theta$,*
- *θ is $R_{\theta,\lambda}$-identifiable everywhere if and only if $H\{\theta; \ell_0\}$ is a singleton for every $\ell_0 \in \Lambda$.*

Proof. We only need to prove the first element, since the second follows by definition. Fix $\ell_0 \in \Lambda$ and define $\mathcal{S}_0 = \{S \in \mathcal{S} : \lambda(S) = \ell_0\}$. Suppose that θ is $R_{\theta,\lambda}$-identifiable at ℓ_0. By Definition 4, there exits ϑ_0 such that $\theta(S) = \vartheta_0$ for all $S \in \mathcal{S}_0$. The identification region is then

$$\begin{aligned} H\{\theta; \ell_0\} &= \{\theta(S) : S \in \mathcal{S}, \lambda(S) = \ell_0\} \\ &= \{\theta(S) : S \in \mathcal{S}\} \\ &= \{\vartheta_0\}. \end{aligned}$$

This completes the first part of the proof. Now suppose that θ is such that $H\{\theta;\ell_0\} = \{\vartheta_0\}$ for some $\vartheta_0 \in \Theta$. Then, by Definition 7, $\theta(S) = \vartheta$ for all $S \in \mathcal{S}$ such that $\lambda(S) = \ell_0$ which is the definition of identifiability (Definition 4). □

To illustrate the concept of an identification region, consider Example 6, with the additional information that $Y \in \mathbb{R}$ (making $\Theta = \mathbb{R}$), and $P(Z = 1) \neq 1$. Fixing $\ell_0 = (P^{(0)}(Y \mid Z = 1), P^{(0)}(Z = 1)) \in \Lambda$ we need to check for every $\vartheta \in \Theta$ whether it belongs to the identification region of ℓ_0. The previous section showed that both $\theta_1 = E[Y|Z = 1]$ and $\theta_2 = P(Z = 1)$ are identifiable so let ϑ_1 and ϑ_2 be the unique elements of Θ such that $\vartheta_1 R_{\theta_1,\lambda} \ell_0$ and $\vartheta_2 R_{\theta_2,\lambda} \ell_0$. Now take $P(Y \mid Z = 0)$ to be any distribution with expectation $E[Y \mid Z = 0] = (\vartheta - \vartheta_2\vartheta_1)/(1 - \vartheta_1)$. By construction, $S = (P^{(0)}(Y \mid Z = 1), P(Y \mid Z = 0), P^{(0)}(Z = 1))$ is such that $\lambda(S) = \ell_0$ and $\theta(S) = \vartheta$, therefore $\vartheta \in H\{\theta;\ell_0\}$. Since this is true for all $\vartheta \in \mathbb{R}$, we conclude that $H\{\theta;\ell_0\} = \mathbb{R}$.

Generally, we interpret Θ as the values of the estimand *a priori* possible, and $H\{\theta;\ell_0\}$ as the values of the estimands compatible with the observation ℓ_0. If $H\{\theta;\ell_0\} = \Theta$ then, not only is θ not identifiable, but ℓ_0 carries no information about the range of plausible values for θ. We call such a quantity *strongly non-identifiable*.

Definition 8 (Strong non-identifiability)**.** *A mapping $\theta \in G(\mathcal{S})$ is said to be strongly non-identifiable if*

$$\forall \ell \in \Lambda, \quad H\{\theta;\ell\} = \Theta$$

We will usually write $H\{\theta;\ell\} = H\{\theta\}$.

When discussing how to establish identifiability, we argued that it was often helpful to break the quantity of interest into simpler pieces and work separately on each piece before re-combining them — the same strategy applies to the identification region. The following proposition gives simple rules for combining identification regions.

Proposition 6. *Let $\lambda, \theta_1, \theta_2 \in G(\mathcal{S})$. Let $\theta \in G(\mathcal{S})$, and consider a function f such that:*

$$\forall S \in \mathcal{S}, \theta(S) = f(\theta_1(S), \theta_2(S))$$

Then if θ_1 is $R_{\theta_1,\lambda}$-identifiable, the identification region of θ at $\ell_0 \in \Lambda$ is

$$H\{\theta;\ell_0\} = \{f(\vartheta_0, \vartheta) : \vartheta \in H\{\theta_2;\ell_0\}\},$$

where $H\{\theta_2; \ell_0\}$ is the identification region of θ_2 with respect to $R_{\theta_2,\lambda}$, at ℓ_0, and $\{\vartheta_0\} = R^{-1}_{\theta_1,\lambda}(\ell_0)$.

Proof. By Definition 7,

$$H\{\theta; \ell_0\} = \{f(\theta_1(S), \theta_2(S)) : S \in \mathcal{S}, \lambda(S) = \ell_0\}$$

Since θ_1 is $R_{\theta_1,\lambda}$-identifiable, by Definition 4, there exists $\vartheta_0 \in \Theta$ such that $\theta_1(S) = \vartheta_0$ for all $S \in \mathcal{S}$ such that $\lambda(S) = \ell_0$. The identification region is then,

$$H\{\theta; \ell_0\} = \{f(\vartheta_0, \theta_2(S)) : S \in \mathcal{S}, \lambda(S) = \ell_0\}.$$

Now notice that $H\{\theta_2(S), \ell_0\} = \{\theta_2(S) : S \in \mathcal{S}, \lambda(S) = \ell_0\}$ by definition, and so:

$$\begin{aligned} H\{\theta; \ell_0\} &= \{f(\vartheta_0, \vartheta) : \vartheta \in \{\theta_2(S) : S \in \mathcal{S}, \lambda(S) = \ell_0\}\} \\ &= \{f(\vartheta_0, \vartheta) : \vartheta \in H\{\theta_2; \ell_0\}\} \end{aligned}$$

□

When we proved that $H\{\theta; \ell_0\} = \mathbb{R}$ earlier in this section, we relied implicitly on this result. In fact, the decomposition

$$\underbrace{E[Y]}_{\theta(S)} = \underbrace{E[Y \mid Z = 1]}_{\theta_2(S)}\underbrace{P(Z = 1)}_{\theta_1(S)} - \underbrace{E[Y \mid Z = 0]}_{\theta_3(S)}\underbrace{P(Z = 0)}_{\theta_4(S)} \tag{2}$$

has an additional property of interest: the terms θ_1, θ_2 and θ_4 are identifiable while θ_3 is strongly non-identifiable. Intuitively, this factorization isolates the non-identifiable part of θ in a single term. We call this factorization a *reduced form*.

Proposition 7. *Fix $\mathcal{S}$, λ, and θ such that we can factorize as $\theta(S) = f(\{\theta_k(S)\}_{k=1}^K, \theta_*(S))$ for some f and $\{\theta_k\}_k^K \in G(\mathcal{S})$. If*

1. *θ_k is $R_{\theta_k,\lambda}$-identifiable at ℓ_0 for $k = 1, \ldots, K$, and*
2. *θ_* is strongly non-$R_{\theta_*,\lambda}$-identifiable at ℓ_0.*

then the reduced form *of $H\{\theta; \ell_0\}$ is*

$$H\{\theta; \ell_0\} = \left\{ f(\{\vartheta_k\}_{k=1}^K, \vartheta),\ \vartheta \in H\{\theta_*\} \right\},$$

where ϑ_k is the unique element of Θ such that $\vartheta_k R_{\theta_k,\lambda} \ell_0$, for $k = 1, \ldots, K$.

Proof. The proof follows the same lines as that of Proposition 6, the difference being that since θ_* is strongly-non identifiable, we have $H\{\theta_*, \ell_0\} = H\{\theta_*\}$. □

In our running example, both $\theta_1(S) = P(Z = 1)$ and $\theta_2(S) = E[Y|Z = 1]$ are identifiable, whereas $\theta_3(S) = E[Y|Z = 0]$ is strongly non-identifiable. The reduced form of the identification region for $\theta = E[Y]$ at $\ell_0 \in \Lambda$ is then

$$H\{\theta; \ell_0\} = \{(\vartheta_2\vartheta_1 + \vartheta(1 - \vartheta_1)), \vartheta \in H\{\theta_3\}\},$$

where ϑ_1 and ϑ_2 are the unique elements of Θ such that $\vartheta_1 R_{\theta_1,\lambda}\ell_0$ and $\vartheta_2 R_{\theta_2,\lambda}\ell_0$, respectively. Since θ_3 is strongly non-identifiable $H\{\theta_3; \ell_0\} = \mathbb{R}$ and, therefore, $H\{\theta; \ell_0\} = \mathbb{R}$, when $P(Z = 1) \neq 1$.

4.3 Incorporating assumptions

In the derivation of the identification region of $H\{\theta; \ell_0\} = \mathbb{R}$, we made no assumption about the outcomes Y (only requiring them to be real numbers). Suppose that we assume $Y \in [0, 1]$, how does this affect the identification region of θ at ℓ_0? That is the type of question that the third step of identification analysis seeks to answer. In our framework, we formalize assumptions as functions inducing restrictions on the statsistical universe $\mathcal{S}$.

Definition 9 (Assumption)**.** *An assumption is a function $A : \mathcal{S} \to \mathbb{R}$. The set $\mathcal{A} = \{S \in \mathcal{S} : A(S) = 0\}$ is called the subset of $\mathcal{S}$ satisfying assumption A.*

To incorporate assumptions in our framework, we augment our notation with a superscript A; defining, for instance, $R^A_{\theta,\lambda}$ to be the restriction of $R_{\theta,\lambda}$ to the set $\mathcal{A}$. In general, the purpose of an assumption is to make $R^A_{\theta,\lambda}$ "closer" to injective (Figure 4.3, provides an intuitive visualization). The restricted identification region is then a subset of the full identification region,

$$H^A\{\theta; \ell_0\} = \{\theta(S) : S \in \mathcal{A}, \lambda(S) = \ell_0\} \subseteq H\{\theta; \ell_0\}.$$

In particular, an assumption makes θ identifiable when $H^A\{\theta; \ell_0\}$ is a singleton. For instance, continuing with our missing data example, let $A(S) = \delta(P(Y, Z), P(Y)P(Z))$ where δ is the total variation distance. This specification is equivalent to assuming that Y and Z are independent[1]. Under this assumption, it is easy to verify that $\theta(S) = E[Y]$ is identifiable. Following Manski (2009) we distinguish two types of assumptions.

[1] Mealli and Rubin (2015) call this assumption missing always completely at random.

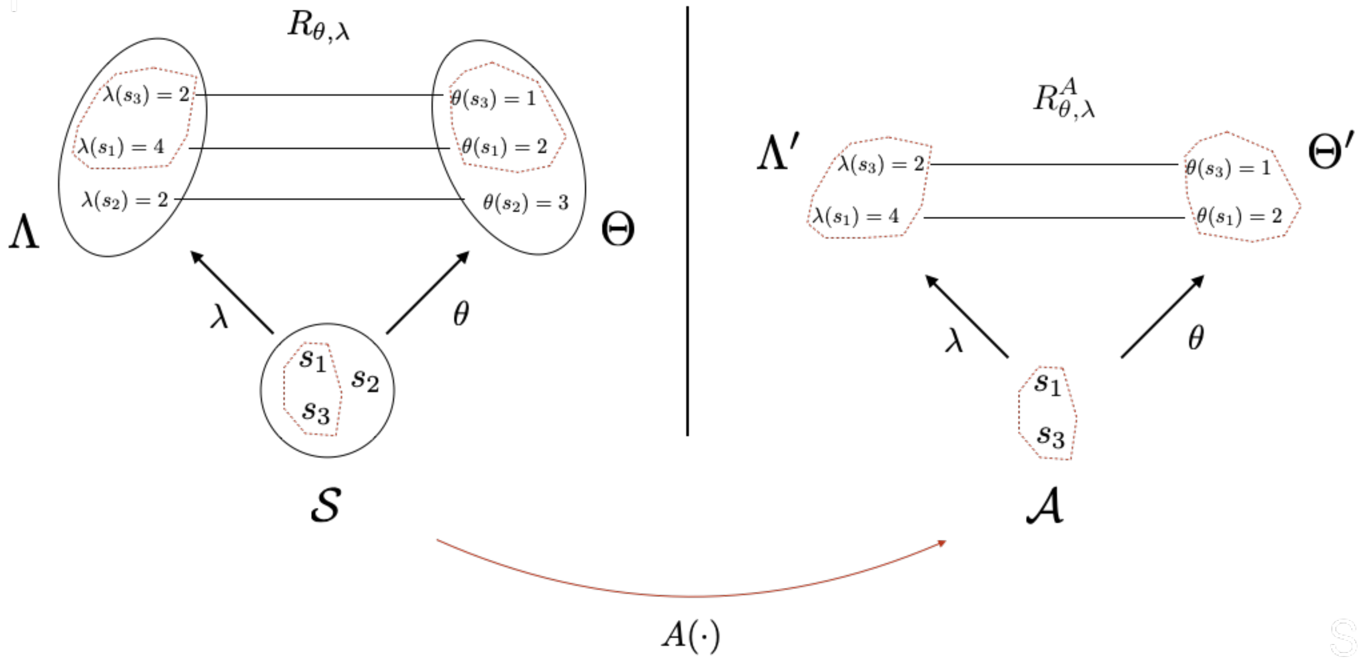

Figure 2: Illustration of an assumption. On the left, $R_{\theta,\lambda}$ is not injective. On the right, $R^A_{\theta,\lambda}$ is injective making θ identifiable.

Definition 10. *Define* $Img_A(\lambda) = \{\lambda(S) : S \in \mathcal{A}\}$. *If* $Img_A(\lambda) = Img(\lambda)$, *then assumption* A *is said to be a priori irrefutable. Otherwise, it is called a priori refutable.*

We use the term *a priori* to stress that we can determine if an assumption is refutable before observing any data. Once we observe $\lambda(S) = \ell_0$ an *a priori* refutable assumption is either refuted, or it is not, depending on whether $\ell_0 \in \text{Img}_A(\lambda)$; in other words, an assumption may be *a priori* refutable and yet not be refuted by a given ℓ_0.

In the context of our running example, the assumption that $Y \in [0, 1]$ is *a priori* refutable — observing data points outside of this interval would refute the assumption. Under this assumption, it is easy to verify that $H^A\{E[Y|Z = 0]; \ell_0\} = [0, 1]$ and, therefore, the reduced form of the identification region of $\theta(S) = E[Y]$ at ℓ_0 is:

$$H^A\{\theta; \ell_0\} = \{(\vartheta_2\vartheta_1 + \vartheta(1 - \vartheta_1)), \vartheta \in [0, 1]\},$$

which provides a natural bound for $E[Y]$.

5 Case studies

5.1 Fixed margin problem

Continuing with the setup of Example 5 with the added simplifying assumption that all the probabilities are continuous, let F_X, F_Y and F_{XY} be the cumulative density functions (CDFs) of P_X, P_Y, and P_{XY} respectively. Recall that $S = P_{XY}$ and $\mathcal{S}$ is the set of all continuous joint probability distributions. The observation mapping is $\lambda(S) = (F_X, F_Y)$ and the estimand mappings are $\theta_X(S) = F_X$, $\theta_Y(S) = F_Y$ and $\theta(S) = F_{XY}$.

We begin by establishing that θ is not everywhere $R_{\theta,\lambda}$-identifiable. Let $\mathcal{C}$ be the set of all copulas and let $\ell_0 = (F_X^{(0)}, F_Y^{(0)})$. By Sklaar's theorem, for all $C \in \mathcal{C}$, the function

$$F_{XY}^{(0,C)} : (x, y) \mapsto C(F_X^{(0)}(x), F_Y^{(0)}(y))$$

is a valid joint CDF with margins $F_X^{(0)}$ and $F_Y^{(0)}$. In particular, let $C_1 \neq C_2 \in \mathcal{C}$ and $S_1 = F_{XY}^{(0,C_1)} \in \mathcal{S}$, $S_2 = F_{XY}^{(0,C_2)} \in \mathcal{S}$. Then $\lambda(S_1) = \lambda(S_2) = (F_X^{(0)}, F_Y^{(0)}) = \ell_0$ by construction, but $\theta(S_1) \neq \theta(S_2)$ since $C_1 \neq C_2$. That is, θ is not $R_{\theta,\lambda}$-identifiable at ℓ_0.

Next, we will derive the identification region of θ at ℓ_0—the set of all CDFs of continuous joint distributions with margins $F_X^{(0)}$ and $F_Y^{(0)}$—in its reduced form. By Sklaar's theorem, for any joint CDF F_{XY} with margins F_X and F_Y, there exists a unique (since we focus on continuous probabilities) copula $C \in \mathcal{C}$ such that:

$$F_{XY}(x, y) = C(F_X(x), F_Y(y)), \qquad \forall x, y.$$

Denote by $\theta^* \in G(\mathcal{S})$ the function that maps each $S = P_{XY}$ to its unique associated copula, and let f be the function that maps $(\{F_X, F_Y\}, C)$ to the valid joint CDF $F_{XY}(x, y) = C(F_X(x), F_Y(y))$. In our notation, we have:

$$\theta(S) = f(\{\theta_X(S), \theta_Y(S)\}, \theta^*(S)).$$

Since θ_X and θ_Y are identifiable, we can write:

$$H\{\theta; \ell_0\} = \left\{ f(\{\vartheta_X^{(0)}, \vartheta_Y^{(0)}\}, \vartheta), \quad \vartheta \in H\{\theta^*; \ell_0\} \right\} \tag{3}$$

But it is easy to verify that $H\{\theta^*; \ell_0\} = H\{\theta^*\} = \mathcal{C}$. Indeed, reasoning by contradiction, suppose that $\exists C_0 \in \mathcal{C} \backslash H\{\theta^*; \ell_0\}$. Let $F_{XY}^{(0)} = C_0(F_X^{(0)}, F_Y^{(0)})$. By Sklaar's theorem, $F_{XY}^{(0)} \in \mathcal{S}_0$. But then by definition

$C_0 = \theta^*(F_{XY}^{(0)}) \in H\{\theta^*; \ell_0\}$ which is a contradiction. Therefore $\mathcal{C} = H\{\theta^*; \ell_0\}$ That is, θ^* is strongly nonidentifiable. So Equation 3 is the reduced form representation of the identification region of θ. From Theorem 2.2.3 of Nelsen (1999), we have that

$$W(x,y) \leq C(x,y) \leq M(x,y)$$

with $W(x,y) = \max(x+y-1,0)$ and $M(x,y) = \min(x,y)$. Since W and M are both copulas, we have:

$$W(F_X(x), F_Y(y)) = \min_{\vartheta \in H\{\theta;\ell_0\}} \vartheta(x,y) \quad \leq \quad \vartheta(x,y) \quad \leq \quad \max_{\vartheta \in H\{\theta;\ell_0\}} \vartheta(x,y) = M(F_X(x), F_Y(y))$$

for all $\vartheta \in H\{\theta; \ell_0\}$. This corresponds exactly to the Hoeffding-Fréchet bounds.

We now assume that the joint distribution is a non-degenerate bivariate normal. Formally, we define the function $A(S) = 0$ iff $S = P_{XY}$ is bivariate normal and $A(S) = 1$ otherwise. In this case, copulas are of the form $C_\rho(u,v) = \Phi_\rho(\Phi^{-1}(u), \Phi^{-1}(v))$ where Φ is the standard normal CDF and Φ_ρ is the CDF of a bivariate normal with mean zero, unit variance, and correlation ρ. Now let $\ell_0 = (F_X^{(0)}, F_Y^{(0)})$ a pair of univariate normal distributions with parameters $(\mu_X^{(0)}, \sigma_X^{(0)})$ and $(\mu_Y^{(0)}, \sigma_Y^{(0)})$ respectively, and define $\tau = (\mu_X^{(0)}, \sigma_X^{(0)}, \mu_Y^{(0)}, \sigma_Y^{(0)})$. In this setting, we can show by the same reasoning as above that the parameter τ is identifiable while the parameter ρ is strongly non-identifiable. The reduced form of the identification region of θ is

$$H^A\{\theta; \ell_0\} = \{\Phi_{\tau,\rho}, \rho \in [-1,1]\}.$$

where $\Phi_{\tau,\rho}$ is the CDF of the bivariate normal distribution with parameters τ and correlation ρ.

If we further assume that the joint distribution is a non-degenerate bivariate normal with correlation ρ_0 (corresponding to assumption A_0) then the identification region reduces to $H^{A_0}\{\theta; \ell_0\} = \{\Phi_{\tau,\rho_0}\}$; that is, θ is identifiable.

Remark 2. *The assumption A is an example of a priori refutable assumption. Indeed, assumption A implies that the marginal distributions F_X and F_Y are themselves normal, so observing non-gaussian marginals would refute the assumption.*

5.2 Causal inference

For an infinite population of experimental units, we are interested in studying the relative effectiveness of a treatment, denoted by $Z = 1$, relative to a control, denoted by $Z = 0$, on an outcome of interest. Under the potential outcomes framework, each unit has two outcomes[2] corresponding to what would happen if the unit receives treatment, $Y(1) \in \mathbb{R}$, or control, $Y(0) \in \mathbb{R}$. Since each unit only receives one treatment and, in turn, only has one observed outcome, causal inference is essentially an identification problem.

Under the nonparametric setup, the statistical universe $\mathcal{S}$ is the space of all possible joint distribution on $(Z, Y(1), Y(0))$, the observation mapping is $\lambda(S) = (P(Y^*|Z), P(Z))$, where $Y^* = ZY(1) + (1-Z)Y(0)$ is the observed outcome, and the estimand mapping $\theta(S) = E[Y(1) - Y(0)]$ is the average causal effect.

The estimand mapping naturally splits into two parts,

$$\theta(S) = \underbrace{E[Y(1)|Z=1]P(Z=1) - E[Y(0)|Z=0]P(Z=0)}_{\theta_a} + \underbrace{E[Y(1)|Z=0]P(Z=0) - E[Y(0)|Z=1]P(Z=1)}_{\theta_b}.$$

To determine the identifiability of the first term, write

$$\theta_a(S) = \underbrace{E[Y(1)|Z=1]}_{\theta_2(S)}\underbrace{P(Z=1)}_{\theta_1(S)} - \underbrace{E[Y(0)|Z=0]}_{\theta_3(S)}\underbrace{P(Z=0)}_{1-\theta_1(S)}.$$

By Proposition 3, both $\theta_1(S)$ and $1-\theta_1(S)$ are identifiable as they are a simple function of $\lambda(S)$. To see that $\theta_2(S)$ is identifiable, notice that

$$\theta_2(S) = E[Y(1)|Z=1] = E[ZY(1) + (1-Z)Y(0)|Z=1] = E[Y^*|Z=1]$$

is a function of $\lambda(S)$, Proposition 3 establishes the result. Similarly, $\theta_3(S)$ is identifiable. Applying Proposition 4 stitches the results showing that θ_a is identifiable.

The second term,

$$\theta_b(S) = \underbrace{E[Y(1)|Z=0]}_{\theta_4(S)}\underbrace{P(Z=0)}_{1-\theta_1(S)} - \underbrace{E[Y(0)|Z=1]}_{\theta_5(S)}\underbrace{P(Z=1)}_{\theta_1(S)},$$

is not identifiable because both $\theta_4(S)$ and $\theta_5(S)$ are generally not. To show $\theta_4(S)$ is not identifiable, assume the outcomes are binary and consider

$$S_1 = (P(Y(1), Y(0)) = (1,0)|Z=1) = 1, P((Y(1), Y(0)) = (0,0)|Z=0) = 1, P(Z=1) = \frac{1}{2}) \text{ and}$$
$$S_2 = (P(Y(1), Y(0)) = (1,0)|Z=1) = 1, P((Y(1), Y(0)) = (1,0)|Z=0) = 1, P(Z=1) = \frac{1}{2}).$$

[2] We implicitly assume the stable unit treatment value assumption Rubin (1980).

Clearly, $\lambda(S_1) = \lambda(S_2)$ while $\theta_4(S_1) = 0 \neq 1 = \theta_4(S_2)$, applying Definition 4 shows that θ_4 is not identifiable. A similarly approach shows that $\theta_5(S)$ is not identifiable.

To derive the reduced form for the identification region it remains to show that θ_b is strongly non-identifiable. Fix $\ell_0 = (P^{(0)}(Y^*|Z)P^{(0)}(Z))$, and let ϑ_1, ϑ_2, and ϑ_3 be the unique values in Θ corresponding to $\vartheta_1 R_{\theta_1,\lambda}\ell_0$, $\vartheta_2 R_{\theta_2,\lambda}\ell_0$, and $\vartheta_3 R_{\theta_3,\lambda}\ell_0$, respectively. Let $Y^m = Y(1)(1-Z)+Y(0)Z$ by the unobserved potential outcome. Notice that we can generally write elements of $\mathcal{S}$ as $S = (P(Y^m|Y^*, Z)P(Y^*|Z)P(Z))$. For $\alpha \in \mathbb{R}$, let $S_\alpha = (P^{(\alpha)}(Y^m|Y^*, Z)P^{(0)}(Y^*|Z)P^{(0)}(Z))$, where $P^{(\alpha)}(Y^m|Y^*, Z)$ is a distribution independent of Y^* with mean α. For all values of α, $\lambda(S_\alpha) = \ell_0$ and $\theta_b(S_\alpha) = \alpha$; giving us that $H\{\theta_b, \ell_0\} = \mathbb{R}$, which shows that θ_b is strongly non-identifiable. The reduced form is then,

$$H\{\theta, \ell_0\} = \{(\vartheta_2\vartheta_1 - \vartheta_3(1 - \vartheta_1) - \vartheta),\ \vartheta \in H\{\theta_b\}\} = \mathbb{R}.$$

We now consider how the reduced form is impacted by different assumptions. Assume the data will be collected from a Bernoulli randomized experiment, that is Z is independent of $(Y(1), Y(0))$ or, using the notation from the previouse section, $A_1(S) = \delta(P(Y(1), Y(0), Z), P(Y(1), Y(0))P(Z))$. This assumption is *a priori* irrefutable as $Img_{A_1}(\lambda) = Img(\lambda)$.

Assumption A_1 makes $\theta_4(S)$ point identifiable as $\theta_4(S) = E[Y(1)|Z = 0] = E[Y(1)|Z = 1] = \theta_2(S)$ — which we already showed is identifiable. Similarly $\theta_5(S) = \theta_3(S)$, and is also identifiable. Some basic algebra shows that the reduced form for $\ell_0 \in \Lambda$ is the singleton,

$$H\{\theta, \ell_0\} = \{\vartheta_2 - \vartheta_3\}.$$

Now consider the alternative assumption asserting that $Y(1), Y(0) \in [0, 1]$. This assumption is *a priori* refutable as $Img_{A_2}(\lambda) \neq Img(\lambda)$, that is, we have changed the image of the observation mapping. In this setting, it is useful to rewrite the estimand mapping as,

$$\begin{aligned}\theta(S) =& E[Y(1)|Z = 1]P(Z = 1) + E[Y(1)|Z = 0]P(Z = 0) - E[Y(0)|Z = 1]P(Z = 1) - E[Y(0)|Z = 0]P(Z = 0)\\ =& E[Y^*|Z = 1]P(Z = 1) - E[Y^*|Z = 0](1 - P(Z = 1))\\ &+ \left(E[Y(1)|Z = 0] - E[Y(0)|Z = 1]\frac{P(Z = 1)}{1 - P(Z = 1)}\right)(1 - P(Z = 1))\end{aligned}$$

The identification region for θ is,

$$H^{A_2}\{\theta, \ell_0\} = \left\{ \left(\vartheta_2\vartheta_1 - \vartheta_3(1-\vartheta_1) + \left[\vartheta - \frac{\vartheta'\vartheta_1}{1-\vartheta_1} \right](1-\vartheta_1) \right),\ \vartheta \in H^{A_2}\{\theta_4\}, \vartheta' \in H^{A_2}\{\theta_5\} \right\}$$

Since both θ_4 and θ_5 are strongly non-identifiable, the $Img(\theta_4) = Img(\theta_5) = [0,1]$. We can then rewrite the identification region as

$$\begin{aligned} H^{A_2}\{\theta, \ell_0\} &= \{[\vartheta_2\vartheta_1 - \vartheta_3(1-\vartheta_1) + (1-\vartheta_1)] - \vartheta,\ \vartheta \in [0,1]\} \\ &= \{[\vartheta_2\vartheta_1 - \vartheta_3(1-\vartheta_1)] - \vartheta,\ \vartheta \in [-\vartheta_1, 1-\vartheta_1]\}. \end{aligned}$$

If we further restricted the treatment assignment probabilities to be between 0.4 and 0.6, the reduced form would become

$$H^{A_2}\{\theta, \ell_0\} = \{[\vartheta_2\vartheta_1 - \vartheta_3(1-\vartheta_1)] - \vartheta,\ \vartheta \in [-0.6, 0.6]\}.$$

The reduced form naturally provides nonparametric bounds for the causal effects. Adding further assumptions can tighten the bound until we are left with a single point, as was the case when we assumed the data were collected from a Bernoulli randomized experiment.

6 Discussion

In this paper, we propose a unifying perspective on identification. Our theory centers around the idea that identifiability can be defined in terms of the injectivity of a certain binary relation. Examining the literature through this lens, we show that existing ad-hoc definitions are special cases of our general framework. One benefit of our flexible formulation is that it brings a new level of transparency and transferability to the concept of identification, allowing us to apply it in settings in which traditional definitions can not be used (Examples 5, 6, 7). In addition to providing a flexible — and completely general — definition of identifiability, we formalize a three-step process, called identification analysis, for studying identification problems.Identification logically precedes estimation: this paper has focused exclusively on the former. A challenge, when thinking about these concepts, is that identification deals with the idealized "infinite number of observations" setting, while estimation happens in finite samples. A quantity can, therefore, be identifiable, but difficult to estimate precisely in practice. Nevertheless, thinking about identification first is a critical step, and we hope that our framework will help in that regard.

References

Basu, A. P. (2006). *Identifiability.* American Cancer Society.

Bojinov, I. I., N. S. Pillai, and D. B. Rubin (2020, 03). Diagnosing missing always at random in multivariate data. *Biometrika 107*(1), 246–253.

Casella, G. and R. L. Berger (2002). *Statistical inference*, Volume 2. Duxbury Pacific Grove, CA.

Cross & Manski, C. (2002). Regression, short and long. *Econometrica 70.*

Fréchet, M. (1951). Sur les tableaux de corrélation dont les marges sont données. *Ann. Univ. Lyon, 3ˆ e serie, Sciences, Sect. A 14*, 53–77.

Freedman, R. S. (2015). Some new results on binary relations. *arXiv preprint arXiv:1501.01914*.

Frisch, R. (1934). *Statistical confluence analysis by means of complete regression systems*, Volume 5. Universitetets Økonomiske Instituut.

Goodman, L. A. (1959). Some alternatives to ecological correlation. *American Journal of Sociology 64*(6), 610–625.

Halmos, P. R. (2017). *Naive set theory.* Courier Dover Publications.

Hsiao, C. (1983). Identification. *Handbook of econometrics 1*, 223–283.

Hurwicz, L. (1950). Generalization of the concept of identification. In T. Koopmans (Ed.), *Statistical Inference in Dynamic Economic Models*, Number 10 in Cowles Commission for Research in Economics, pp. 245–257. New York: John Wiley & Sons, Inc.

Imbens, G. W. and C. F. Manski (2004). Confidence intervals for partially identified parameters. *Econometrica 72*(6), 1845–1857.

Jacquez, J. A. and T. Perry (1990). Parameter estimation: local identifiability of parameters. *American Journal of Physiology-Endocrinology and Metabolism 258*(4), E727–E736.

King, G. (2013). *A solution to the ecological inference problem: Reconstructing individual behavior from aggregate data.* Princeton University Press.

Koopmans, T. C. (1949). Identification problems in economic model construction. *Econometrica, Journal of the Econometric Society*, 125–144.

Koopmans, T. C. and O. Reiersol (1950). The identification of structural characteristics. *The Annals of Mathematical Statistics 21*(2), 165–181.

Lehman, E., F. T. Leighton, and A. R. Meyer (2010). Mathematics for computer science. Technical report, Technical report, 2006. Lecture notes.

Lehmann, E. L. and G. Casella (2006). *Theory of point estimation.* Springer Science & Business Media.

Manski, C. F. (1989). Anatomy of the selection problem. *Journal of Human resources*, 343–360.

Manski, C. F. (1990). Nonparametric bounds on treatment effects. *The American Economic Review 80*(2), 319–323.

Manski, C. F. (2003). *Partial identification of probability distributions.* Springer Science & Business Media.

Manski, C. F. (2009). *Identification for prediction and decision.* Harvard University Press.

Matzkin, R. L. (2007). Nonparametric identification. *Handbook of Econometrics 6*, 5307–5368.

Matzkin, R. L. (2013). Nonparametric identification in structural economic models. *Annu. Rev. Econ. 5*(1), 457–486.

Mealli, F. and D. B. Rubin (2015). Clarifying missing at random and related definitions, and implications when coupled clarifying missing at random and related definitions, and implications when coupled with exchangeability. *Biometrika 102*(4), 995–1000.

Nelsen, R. B. (1999). An introduction to copulas, vol. 139 of lecture notes in statistics.

Paulino, C. D. M. and C. A. de Bragança Pereira (1994). On identifiability of parametric statistical models. *Journal of the Italian Statistical Society 3*(1), 125–151.

Pearl, J. (2009). Causal inference in statistics: An overview. *Statistics surveys 3*, 96–146.

Romano, J. P. and A. M. Shaikh (2008). Inference for identifiable parameters in partially identified econometric models. *Journal of Statistical Planning and Inference 138*(9), 2786–2807.

Rothenberg, T. J. (1971). Identification in parametric models. *Econometrica: Journal of the Econometric Society*, 577–591.

Rubin, D. B. (1980). Randomization analysis of experimental data: The fisher randomization test comment. *Journal of the American statistical association 75*(371), 591–593.

Shpitser, I. (2008). *Complete identification methods for causal inference.* Ph. D. thesis, UCLA.

Tamer, E. (2010). Partial identification in econometrics. *Annu. Rev. Econ. 2*(1), 167–195.