A User's Guide for Inference in Models Defined by Moment Inequalities[*]

Ivan A. Canay[†] Gastón Illanes[‡] Amilcar Velez[§]

July 6, 2023

Abstract

Models defined by moment inequalities have become a standard modeling framework for empirical economists, spreading over a wide range of fields within economics. From the point of view of an empirical researcher, the literature on inference in moment inequality models is large and complex, including multiple survey papers that document the non-standard features these models possess, the main novel concepts behind inference in these models, and the most recent developments that bring advances in accuracy and computational tractability. In this paper we present a guide to empirical practice intended to help applied researchers navigate all the decisions required to frame a model as a moment inequality model and then to construct confidence intervals for the parameters of interest. We divide our template into four main steps: (a) a behavioral decision model, (b) moving from the decision model to a moment inequality model, (c) choosing a test statistic and critical value, and (d) accounting for computational challenges. We split each of these steps into a discussion of the "how" and the "why", and then illustrate how to take these steps to practice in an empirical application that studies identification of expected sunk costs of offering a product in a market.

KEYWORDS: Partially Identified Model, Confidence Regions, Uniform Asymptotic Validity, Moment Inequalities

JEL classification codes: C12, C14

[*] We would like to thank Matthew Thomas for excellent research assistance in the production of the companion Github repository (https://github.com/iacanay/guide-inequalities) with codes in `Matlab`, `Python`, and `R`. We also thank the editor, Elie Tamer, and two anonymous referees for comments and suggestions. Researcher(s)' own analyses calculated (or derived) based in part on data from Nielsen Consumer LLC and marketing databases provided through the NielsenIQ Datasets at the Kilts Center for Marketing Data Center at The University of Chicago Booth School of Business. The conclusions drawn from the NielsenIQ data are those of the researcher(s) and do not reflect the views of NielsenIQ. NielsenIQ is not responsible for, had no role in, and was not involved in analyzing and preparing the results reported herein.

[†] Department of Economics, Northwestern University. iacanay@northwestern.edu

[‡] Department of Economics, Northwestern University & NBER. gaston.illanes@northwestern.edu

[§] Department of Economics, Northwestern University. amilcare@u.northwestern.edu

1 Introduction

Partially identified models have become a standard modeling framework for empirical economists in the last two decades, spreading over a long list of topics in several areas. Examples include measurement error (Klepper and Leamer, 1984; Horowitz and Manski, 1995), missing data (Manski, 1989, 1994; Horowitz and Manski, 1998; Manski and Tamer, 2002), industrial organization (Tamer, 2003; Haile and Tamer, 2003; Ho, 2009; Holmes, 2011; Crawford and Yurukoglu, 2012; Eizenberg, 2014; Ho and Pakes, 2014; Pakes et al., 2015; Wollmann, 2018; Houde et al., 2023; Maini and Pammolli, 2023), finance (Hansen and Jagannathan, 1991; Hansen et al., 1995), labor economics (Blundell et al., 2007; Kline et al., 2013; Kline and Tartari, 2015), trade (Dickstein and Morales, 2018; Morales et al., 2019), program evaluation (Manski, 1990, 1997; Manski and Pepper, 2000; Heckman and Vytlacil, 2001; Bhattacharya et al., 2008, 2012; Shaikh and Vytlacil, 2011), and limited consideration sets (Cattaneo et al., 2020; Barseghyan et al., 2021), to name a few. Within the family of partially identified models, the set of models that can be represented by a set of moment inequalities has received particular attention for two reasons. First, a large class of behavioral decision models in industrial organization, labor, and other subfields, admit a representation in terms of moment inequalities; making moment inequality models a versatile way of representing behavior of agents across economic settings. Second, the structure imposed by moment inequalities facilitates the study and development of econometric tools to conduct inference in these models that help compensate for the challenging non-standard features that partially identified models introduce relative to traditional, point identified, models.

While there are challenges and questions that have not yet been resolved, the literature on inference in moment inequality models is mature enough to have produced multiple survey papers to this date. For example, Tamer (2010) provides a comprehensive summary of the history and early developments of the literature. Canay and Shaikh (2017) review the fundamentals behind the construction of confidence regions for parameters in identified sets characterized by a finite number of moment inequalities, with an emphasis on the importance of requiring confidence regions to be uniformly consistent in level over relevant classes of distributions. Ho and Rosen (2017) survey a class of econometric models used in a variety of empirical applications and discuss the roles that assumptions and data play in partial identification analysis. Molinari (2020) reviews the microeconometrics literature on partial identification with a focus on random set theory as a mathematical framework to unify a variety of models and methodologies. Finally, Kline et al. (2020) and Kline and Tamer (2022) focus on the most recently developed tools for inference in partially identified models and their applications to industrial organization. The rise in the popularity of these models is in part due to the advocacy of Charles Manski, who, in a series of books and articles, argued forcefully that partially identified models enable researchers to make more credible inferences (see the book-length treatments in Manski, 1995, 2003, 2007, 2013).

In this paper we present a guide to empirical practice intended to help applied researchers navigate all the decisions required to frame a model as a moment inequality model and then to construct confidence intervals for the parameters of interest. While we survey the literature and the most recent theoretical advances along the way, our main goal is not to provide a comprehensive chronological road-map of all the methods that are available up to this date but rather to provide a template that hopefully lowers the entry cost to the literature, both to newcomers and researchers with some exposure to the basic tools. The template we provide is divided into four main steps: (a) developing a behavioral decision model, (b) stating the assumptions required to go from the decision model to one defined in terms of moment inequalities, (c) choosing an appropriate test statistic and critical value, and (d) accounting for the computational considerations behind the construction of confidence intervals. We split each of these steps into a discussion of the "how", which succinctly describes the steps we recommend and follow in our empirical application, and a discussion of the "why", which discusses the considerations that led to our recommendations as well as the alternatives that are currently available to the researcher at each particular step. A reader can then choose to focus on the "how", and learn an established approach to inference in moment inequality models without digging into the overwhelming number of alternatives available at each stage, or have a more in-depth exposure, by not only learning the approach we adopt, but also becoming aware of the pros and cons of alternative tools.

We illustrate how to use the template we develop in an empirical application that studies sunk costs of offering products in the context of the acquisition of Energy Brands by The Coca Cola Company in 2007. We provide computer codes in `Matlab`, `Python`, and `R` that not only replicate our empirical application (using simulated data), but are also flexible enough for researchers to use them as a starting point in the development of their own code for similar empirical settings. All in all, we expect the combination of the guiding template with the computer codes to provide an easy to digest introduction to inference in moment inequality models that fosters the adoption of such models in empirical research.

The rest of the paper is structured as follows. Section 2 introduces basic definitions and notation. Section 3 describes the behavioral choice model that we use in our empirical application. Section 4 introduces the main assumptions that are needed to write the behavioral model as a model defined by a finite number of unconditional moment inequalities. Section 5 describes the test we use to test the hypothesis that the moment inequality model holds at a given value of the parameters of interest, which involves a choice of test statistic and critical value, while Section 6 in turn explains how we invert such a test in order to obtain confidence regions for those parameters. The first subsection in each of these sections, that is, Sections 3.1, 4.1, 5.1, and 6.1, provide a deep dive into the choices we made and discusses other alternatives available in the literature that the analyst may consider. With this structure, a reader mostly interested in a quick, hands-on, introduction to the topic may choose to skip the first subsection in each section. Finally, we adopt the template we developed in an empirical application we present in Sections 7-9. These sections not only illustrates in detail how we use the tools discussed previously, but also describe required preliminary steps, like demand and marginal cost estimation, making the section self-contained from beginning to end.

2 Setup and Notation

Suppose that a researcher observes an i.i.d. sample $\{W_i : 1 \leq i \leq N\}$ from a distribution $P \in \mathbf{P}$ on $\mathbf{R}^{d_W}$. The set $\mathbf{P}$ constitutes the model for the distribution of the observed data, and may include assumptions like finite second moments, certain matrices having full rank, among others. We discuss some of these conditions as we introduce the specifics of our model. A model defined by moment inequalities states that for a finite dimensional parameter vector $\theta_0 \in \Theta \subseteq \mathbf{R}^{d_\theta}$,

$$E_P[m(W_i, \theta_0)] \leq 0 \ , \tag{1}$$

where $m \equiv (m_1, \cdots, m_k)'$ is a known measurable function of the observed data and the parameter θ_0, and where the inequality is interpreted component-wise. The identified set for θ_0 is the set of values satisfying the moment inequalities in (1), i.e.,

$$\Theta_0(P) = \{\theta \in \Theta : E_P[m(W_i, \theta)] \leq 0\} \ . \tag{2}$$

We focus on the construction of confidence regions C_n for points θ in the identified set $\Theta_0(P)$ that are uniformly consistent in level, as defined in Canay and Shaikh (2017). Concretely, the random set C_n satisfies

$$\liminf_{n\to\infty} \inf_{P\in\mathbf{P}} \inf_{\theta\in\Theta_0(P)} P\{\theta \in C_n\} \geq 1 - \alpha \ , \tag{3}$$

which is usually accomplished by exploiting the duality between confidence regions and inverting tests of each of the individual null hypotheses

$$H_\theta : E_P[m(W_i, \theta)] \leq 0 \ . \tag{4}$$

More precisely, suppose that for each θ a test of H_θ, $\phi_n(\theta)$, is available that satisfies

$$\limsup_{n\to\infty} \sup_{P\in\mathbf{P}} \sup_{\theta\in\Theta_0(P)} E_P[\phi_n(\theta)] \leq \alpha \ . \tag{5}$$

It then follows that the confidence region that collects all values of $\theta \in \Theta$ that are not rejected by $\phi_n(\theta)$, i.e.,

$$C_n \equiv \{\theta \in \Theta : \phi_n(\theta) = 0\} \ , \tag{6}$$

satisfies (3). Here we do not discuss the assumptions on $\mathbf{P}$ that are required for each alternative test to lead to (5). We instead focus on practical and computational considerations, and refer the reader to the original papers, or to Canay and Shaikh (2017) for a unified theoretical framework. Finally, we focus on tests $\phi_n(\theta)$ for the null hypotheses H_θ that take the form,

$$\phi_n(\theta) \equiv I\{T_n(\theta) > c_n(1-\alpha, \theta)\} \ , \tag{7}$$

where $I\{\cdot\}$ denotes the indicator function, $T_n(\theta)$ denotes a test statistic, and $c_n(1-\alpha,\theta)$ denotes a critical value. In what follows, we remove the dependence on P from the expectations to simplify exposition, and so we use $E[\cdot]$ instead of $E_P[\cdot]$.

It is important to point out that the type of coverage for C_n in (3) is only one of the two types that have been proposed in the literature. The second type requires that the random set C_n covers the entire identified set with some pre-specified probability $1-\alpha$, i.e.,

$$\liminf_{n\to\infty} \inf_{P\in\mathbf{P}} P\{\Theta_0(P) \subseteq C_n\} \geq 1-\alpha \ . \tag{8}$$

When θ_0 is point-identified (i.e., $\Theta_0(P)$ is a singleton), then both (3) and (8) reduce to

$$\liminf_{n\to\infty} \inf_{P\in\mathbf{P}} P\{\theta_0 \in C_n\} \geq 1-\alpha \ .$$

In this sense, both (3) and (8) generalize the standard requirement for confidence regions for parameters that are point-identified.

As emphasized by Imbens and Manski (2004), confidence region satisfying (8) of course satisfy (3) as well, so confidence regions for points in the identified set are typically smaller than confidence regions for the identified set. Imbens and Manski (2004) argue further that confidence regions for points in the identified set are generally of greater interest than confidence regions for the identified set itself, as there is still only one "true" value for θ in the identified set. Other authors, however, have argued that in some instances confidence regions for the identified set are more desirable. See, for example, Henry and Onatski (2012), who describe a decision problem in which constructing a confidence region satisfying (3) is undesirable. In this user's guide, we focus on confidence regions satisfying (3), which are the type that have received the most attention in the literature on inference in partially identified models. Notable exceptions include Chernozhukov et al. (2007), Bugni (2010), Romano and Shaikh (2010), and Chen et al. (2018)

Remark 2.1. The literature on inference in models defined by moment inequalities has also distinguished between two notions of confidence regions: those that are pointwise consistent in levels and those that are *uniformly* consistent in levels. From a purely technical standpoint, the distinction boils down to confidence regions that are pointwise consistent in levels having the order in which the limit and the infimum appear in (3) reversed. More fundamentally, the importance of relying on uniformly valid confidence regions relates to the superior finite sample performance of such confidence regions relative to those that only provide pointwise guarantees. This has been extensively discussed in the literature and so we take it as well understood throughout our discussion. We refer the reader to Canay and Shaikh (2017), and reference therein, for an introduction to these alternative notions of coverage. ■

Remark 2.2. The model in (1) does not explicitly include moment equalities. The methods we review here could all incorporate equality constraints by either writing each equality as two inequalities or by choosing a test statistic in Section 5 that accounts for both, equalities and inequalities. ■

Remark 2.3. In many applications the identified set for θ is determined by conditional moment inequalities, in which case

$$E[g(O_i,\theta)|Z_i] \leq 0 \; a.s.$$

for a known function g, a set of instruments Z, and other observed data O. In this paper we focus on the case where the analyst replaces the conditional inequality with an unconditional inequality $E[g(O_i,\theta)h(Z_i)] \leq 0$ for a non-negative function $h(\cdot)$, so that $m(W_i,\theta) = g(O_i,\theta)h(Z_i)$ in (1) with $W_i = (O_i, Z_i)$. We acknowledge that this keeps the implementation simple at the cost of possibly losing information, as explained in the literature on inference in conditional moment inequality models; see Andrews and Shi (2013); Chernozhukov et al. (2013); Chetverikov (2013); Armstrong (2014b,a, 2015), and Armstrong and Chan (2014), among others. ■

The goal of this paper is to provide a step by step guide to construct the confidence region C_n in (6). We do so by describing the steps we take to compute C_n in the context of the empirical application in Section 8, while also discussing the thought process behind the decisions we made along the way. In this sense, we do not view this paper as a survey of the literature but rather as a blueprint, or starting point, for researchers interested in using models defined by moment inequalities in empirical settings. Specifically, in the next sections we describe in detail what we view as the four essential steps that are required to compute confidence regions in models defined by the inequalities in (1). These four steps are:

Step 1: The behavioral decision model.

Step 2: The moment functions $m(W_i, \theta)$.

Step 3: The test statistic $T_n(\theta)$ and the critical value $c_n(1-\alpha, \theta)$.

Step 4: The algorithm leading to C_n.

Relative to papers introducing methods for constructing confidence regions C_n, which usually proceed by taking the model in (1) as given and then provide a test statistic and a critical value that leads to a test with desirable properties for the hypotheses in (4), here we start with the fundamental behavioral decision model (Step 1) and explain how this may lead to the moment inequality functions $m(W_i, \theta)$ in Step 2. Different empirical applications will necessarily involve a different decision model, but many of the modeling choices we discuss apply broadly and go beyond the empirical setting we consider in Section 8. In Step 2, we follow the "profit inequalities" approach in Pakes (2010) to derive the moment inequalities and illustrate how different sets of moment inequalities can be obtained under different assumptions on the unobservable random variables entering the behavioral decision model. For a more comprehensive discussion on alternative ways to derive moment inequalities in behavioral choice models, we refer the reader to Pakes (2010) and Kline et al. (2020). For Step 3 we rely specifically on the recent work by Chernozhukov et al. (2019) and on the framework presented by Canay and Shaikh (2017) to discuss alternatives to the ones we take in this paper. Step 4 is usually the bottleneck in empirical papers doing inference in models defined by moment inequalities, and has been traditionally neglected in the theoretical literature. Here, we summarize what we view as good practices and share some of the shortcuts that we use to guide practitioners on how to deal with the potentially heavy computational burden associated with these models. Despite our efforts, we view the computational barrier as the main force preventing models defined by moment inequalities to be widely adopted.

3 The behavioral decision model

In order to illustrate how a behavioral choice model may lead to a model defined by moment inequalities as in (1), we focus on the model that leads to the empirical application in Section 8. This model is one where firms decide, for each of the products they produce, whether to offer the product in each of several available markets. We start by introducing some basic notation.

We index firms by $s \in \mathcal{S}$, products by $j \in \mathcal{J}$, and markets by $i \in \mathcal{N}$, each with cardinality $S \equiv |\mathcal{S}|$, $J \equiv |\mathcal{J}|$, and $n \equiv |\mathcal{N}|$, respectively. Each product is produced by only one firm, which means that $\mathcal{J}$ can be partitioned into S disjoint subsets, i.e., $\mathcal{J} = \cup_{s \in \mathcal{S}} \mathcal{J}_s$. Let $J_s \equiv |\mathcal{J}_s|$ for each $s \in \mathcal{S}$ and assume $J_s > 0$ for all $s \in \mathcal{S}$.

On a given market i and for each product j, firm s makes a product offering decision denoted by $D_{i,j} \in \{0, 1\}$, where $D_{i,j} = 1$ when product j is offered in market i and

$D_{i,j} = 0$ otherwise. The entire product portfolio in market i is then given by the following J-dimensional vector,

$$D_i \equiv (D_{i,1}, \ldots, D_{i,J})' \in \{0,1\}^J \ . \tag{9}$$

$D_{i,j}$ is not indexed by s since in our setting each product is produced by only one firm.

We denote the variable profits that firm s gets in market i given a product portfolio D_i by $r_{s,i}(D_i)$. We also denote the sunk cost that firm s pays to introduce product j to market i by $e_{i,j}(\theta)$, which in turn depends on the parameter of interest θ. The total stochastic profits that firm s obtains in market i when the product portfolio is D_i are then given by,

$$\pi_{s,i}(D_i, \theta) \equiv r_{s,i}(D_i) - \sum_{j \in \mathcal{J}_s} D_{i,j} e_{i,j}(\theta) \ . \tag{10}$$

We assume that firm s chooses $\{D_{i,j} : j \in \mathcal{J}_s\}$ by maximizing its expected total profits conditional on the information at its disposal at the time of making the decision. If we denote the information set of firm s by I_s, then firms maximize

$$E[\pi_{s,i}(D_i, \theta)|I_s] = E[r_{s,i}(D_i)|I_s] - \sum_{j \in \mathcal{J}_s} D_{i,j} E[e_{i,j}(\theta)|I_s] \ , \tag{11}$$

where $E[\cdot]$ is an expectation with respect to the distribution across markets. This means

$$\max_{d_i \in \Lambda_s(D_i)} E[\pi_{s,i}(d_i, \theta)|I_s] \leq E[\pi_{s,i}(D_i, \theta)|I_s] \ , \tag{12}$$

where the set $\Lambda_s(D_i)$ captures that firm s can only choose the elements in d_i corresponding to products in $\mathcal{J}_s$, i.e.,

$$\Lambda_s(D_i) \equiv \{d_i : d_{i,j} \in \{0,1\} \text{ if } j \in \mathcal{J}_s \text{ and } d_{i,j} = D_{i,j} \text{ if } j \notin \mathcal{J}_s\} \ . \tag{13}$$

Finally, we introduce notation to capture counter-factual deviations from the observed product portfolio D_i. For simplicity, we focus on one-product deviations and introduce two operators to denote such deviations: ∂_j, which is an operator on D_i defined as

$$\partial_j D_i = (D_{i,1}, \ldots, D_{i,j-1}, 1 - D_{i,j}, D_{i,j+1}, \ldots, D_{i,J}) \ , \tag{14}$$

and Δ_j, which is an operator on $\pi_{s,i}(\cdot)$ or $r_{s,i}(\cdot)$ defined as

$$\Delta_j \pi_{s,i}(D_i, \theta) = \pi_{s,i}(\partial_j D_i, \theta) - \pi_{s,i}(D_i, \theta) \ . \tag{15}$$

That is, ∂_j counter-factually changes the offering decision of product j by making it available in market i ($1 - D_{i,j} = 1$) if it was not previously offered ($D_{i,j} = 0$) or by removing the product from market i ($1 - D_{i,j} = 0$) if it was previously offered ($D_{i,j} = 1$).

In turn, Δ_j computes the difference in profits (variable or total) that would arise if the product portfolio were to change from D_i to $\partial_j D_i$. In Section 3.1 we discuss multiple-product deviations.

In equilibrium and given the profit maximizing behavior in (12), we obtain

$$E[\Delta_j \pi_{s,i}(D_i,\theta)|I_s] \leq 0 \qquad \text{for all } j \in \mathcal{J}_s \text{ and } s \in \mathcal{S} \ . \tag{16}$$

We refer to the inequality in (16) as the inequality induced by profit maximizing behavior. This inequality will lead to the model defined by moment inequalities in (1), but cannot be used directly as both expected variable profits and expected sunk costs are unobserved to the analyst. In what follows we then impose additional structure on costs and variable profits in order to go from the inequality in (16) to the one in (1).

Remark 3.1. The model we focus on here is determined by our empirical application in Section 8, where the main indices are markets and products. In other applications there could be additional dimensions that may introduce other indices, most notably, a time index. Aside from dynamic models, which we do not discuss in this paper, additional dimensions do not fundamentally change our analysis. ■

Expected variable profits and sunk costs

The analyst does not observe expected variable profits, $E[r_{s,i}(d_i)|I_s]$, for any potential product portfolio but rather realized variable profits, $E[r_{s,i}(D_i)|I_s]$, at the observed portfolio. We then require a model capable of predicting expected variable profits both for the observed product assortment and for deviations. In the empirical setting we consider in Section 8, this is a structural model of demand and supply (we present the details of this model in Section 7.2). In order to estimate this model, we assume the analyst observes data on product offering decisions D_i across markets, a vector of other product and market characteristics that we denote by $X_{i,j}$, and a vector of instrumental variables $L_{i,j}$ that are used to estimate demand and supply. The characteristics in $X_{i,j}$ could include variables that are market specific (e.g., market size), product specific (e.g., flavor or input costs), or market and product specific (e.g., price). The observed data to estimate the structural model of demand and supply is then

$$\{O_i \equiv (D_i, X_{i,1}, \ldots, X_{i,J}, L_{i,1}, \ldots, L_{i,J})' : i \in \mathcal{N}\} \ , \tag{17}$$

and is assumed to satisfy the conditions required to invoke appropriate versions of the law of large numbers (LLN) and central limit theorem (CLT) that are needed later on. We also denote by $Z_{i,j}$ a set of additional instruments will be used to instrument for the moment inequalities in Section 4. These instruments could be variables not included in O_i or functions of variables in O_i. The entire data set available to the researcher can be denoted by

$$\{W_i \equiv (O_i, Z_{i,1}, \ldots, Z_{i,J})' : i \in \mathcal{N}\} \ , \tag{18}$$

which is again assumed to satisfy the conditions required to invoke appropriate versions of the LLN and CLT. In the empirical setting we study in Section 8 we have information for hundreds of markets and about 30 products, which explains why we choose markets, $i \in \mathcal{N}$, to be the index determining the data generating process. This choice may vary depending on the setting, as long as there is at least one dimension that gives sufficient independence to reliably invoke LLNs and CLTs.

We make the following assumption on the estimated variable profit differentials, denoted by $\Delta_j \hat{r}_{s,i}(O_i)$, for each product $j \in \mathcal{J}$.

Assumption 3.1. *The estimated variable profit differential $\Delta_j \hat{r}_{s,i}(O_i)$ satisfies, for all $j \in \mathcal{J}$,*

$$\Delta_j \hat{r}_{s,i}(O_i) = E[\Delta_j r_{s,i}(D_i)|I_s] + U_{i,j} \quad \textit{where} \quad E[U_{i,j} \mid D_{i,j}, I_s] = 0 \ .$$

Assumption 3.1 requires that variable profit differentials are sufficiently well estimated. The error $U_{i,j}$ is sometimes referred to as a specification error (or expectational error in some simpler settings) and the conditional mean independence of $U_{i,j}$ with both $D_{i,j}$ and I_s have separate implications. First, conditional mean independence with $D_{i,j}$ requires $E[\Delta_j r_{s,i}(D_i)|I_s]$ to be equally well estimated regardless of the firm's decision about product j in market i, i.e., $D_{i,j}$. Second, conditional mean independence with I_s means that $E[\Delta_j r_{s,i}(D_i)|I_s]$ is the true conditional mean of $\Delta_j \hat{r}_{s,i}(O_i)$. We discuss alternatives to this assumption in Section 3.1.

The analyst does not observe expected sunk costs, $E[e_{i,j}(\theta)|I_s]$, either. Importantly, expected costs capture the parameter of interest θ. We start with the simplest possible specification for sunk costs, as described in the assumption below. In Section 3.1 we discuss alternative specifications and their implications on the construction of confidence sets.

Assumption 3.2. *Sunk costs $e_{i,j}(\theta)$ are known by firms at the time of making product offering decisions, and they admit the representation*

$$e_{i,j}(\theta) = \theta_s + V_{i,j} \ \ \textit{with} \ \ \theta_s \in \mathbf{R} \ \ \textit{and} \ \ E[V_{i,j}] = 0 \ \ \textit{for all} \ \ j \in \mathcal{J} \ .$$

Assumption 3.2 has two parts. First, sunk costs being known means that $e_{i,j}(\theta)$ is part of the information set and so $E[e_{i,j}(\theta)|I_s] = e_{i,j}(\theta)$. Second, the simple parametrization means that the stochastic sunk cost $e_{i,j}(\theta)$ has a firm specific mean θ_s and a product and market specific error term $V_{i,j}$ that is mean zero across markets. As before, this error term is part of the information set, I_s, of the firms and so $E[V_{i,j}|I_s] = V_{i,j}$. The parameter we intend to recover by means of a model defined by moment inequalities is θ_s for all $s \in \mathcal{S}$, i.e.,

$$\theta \equiv (\theta_s : s \in \mathcal{S})' \in \mathbf{R}^S \ . \tag{19}$$

Remark 3.2. In contrast to $U_{i,j}$ in Assumption 3.1, the error term $V_{i,j}$ is not uncorrelated with $D_{i,j}$ by virtue of being known at the time of making the product offering decision. In fact, one would expect $E[V_{i,j}D_{i,j}] = E[V_{i,j}|D_{i,j} = 1]P\{D_{i,j} = 1\}$ to be negative if, conditional on offering a product, costs tend to be lower. Errors such as $V_{i,j}$ are sometimes referred to as “structural”. ■

Combining the profit maximizing behavior in (12) with the structure on expected variable profits and sunk costs imposed by Assumptions 3.1 and 3.2, we can re-write (16) as,

$$\Delta_j \hat{r}_{s,i}(O_i) - U_{i,j} - (1 - 2D_{i,j})(\theta_s + V_{i,j}) \leq 0 \ , \tag{20}$$

or by grouping according to whether products were or were not offered, as

$$(\Delta_j \hat{r}_{s,i}(O_i) - U_{i,j} - (\theta_s + V_{i,j}))(1 - D_{i,j}) \leq 0 \text{ if not offered} \tag{21}$$

$$(\Delta_j \hat{r}_{s,i}(O_i) - U_{i,j} + (\theta_s + V_{i,j}))D_{i,j} \leq 0 \text{ if offered.} \tag{22}$$

The group of inequalities (21) and (22) may appear similar to the one in (1), but there is one important difference: while (1) is a function of the parameter of interest and the observed data only, the inequalities in (21) and (22) depend on the unobserved variables $U_{i,j}$ and $V_{i,j}$. This would not be particularly problematic if these variables were mean independent across markets conditional on $D_{i,j} = 0$ and $D_{i,j} = 1$. This is indeed the case for $U_{i,j}$ but, as discussed in Remark 3.2, likely not the case for $V_{i,j}$. In Section 4 we impose assumptions on $V_{i,j}$ that allow us to derive moment inequality functions that do not depend on $U_{i,j}$ and $V_{i,j}$. In Section 4.1 we review how similar assumptions have been used in the literature, as well as alternative versions that lead to potentially different inequalities.

3.1 Why? Modeling choices

The behavioral decision model we introduce has three features that deserve discussion. The first feature is the profit maximizing behavior in (12) that led to the inequality in (16), which we re-state here for readability:

$$E[\Delta_j \pi_{s,i}(D_i, \theta)|I_s] \leq 0 \qquad \text{for all } j \in \mathcal{J}_s \text{ and } s \in \mathcal{S} \ .$$

This inequality depends on the operator Δ_j and so it depends on one-product deviations; that is, counter-factuals where a single product is added or removed from the market. One-product deviations have been considered in a number of papers, including

Nosko (2010), Eizenberg (2014), Wollmann (2018) and Fan and Yang (2022), among others. However, the profit maximizing behavior in (12) considers *any* type of deviation from the observed actions, and so in principle one could consider q-product deviations, with $q \geq 1$, and derive inequalities associated with the union of such deviations. This approach quickly increases the number of moment inequalities and thus necessarily requires confidence intervals that work well when the number of moment inequalities is large relative to the sample size. For example, in the application we consider in Section 8 the model that considers one-product deviations can contain $k = 2J = 62$ moment inequalities while the one that considers one and two-product deviations can contain $k = 1250$ moment inequalities.

The second feature is the error term $U_{i,j}$ introduced in Assumption 3.1. This error captures the difference between the estimated variable profit differential $\Delta_j \hat{r}_{s,i}(O_i)$ and the true expected profits $E[\Delta_j r_{s,i}(D_i)|I_s]$. This assumption could be relaxed by decomposing $U_{i,j}$ into a specification error and a so-called "structural error", where the structural error would be allowed to be correlated with $D_{i,j}$, see Pakes (2010). We do not consider this case here to keep our exposition simple and because our goal is not to provide the most realistic empirical model but rather to present a relatively simple one that still showcases the difficulties that commonly appear in practice. We note also that it would be possible to treat such an error term similarly to the way we treat the error term $V_{i,j}$ at the expense of additional notation.

The third feature relates to Assumption 3.2. We have decided to focus on a straightforward parameterization to keep exposition simple and to make the role of the structural error $V_{i,j}$ starker: it is likely that sunk costs differ within firms across products and markets, and that firms take this into account when making product offering decisions. In general, we expect most moment inequality models to feature a structural error of some kind, so discussing strategies to deal with this issue is important. A natural extension of Assumption 3.2 is to add covariates to the sunk cost, i.e.,

$$e_{i,j}(\theta) = X_{i,j}'\theta_s + V_{i,j} \ .$$

While this increases the dimensionality of the confidence set and increases computational complexity, it is straightforward from a theoretical perspective. We discuss and implement this extension in Section 8.2.

Remark 3.3. The decision model we use in this paper is one that leads to profit inequalities, as in (16). A popular alternative to this approach is the "generalized discrete choice" model introduced by Tamer (2003) and Ciliberto and Tamer (2009). This approach adds additional structure to the model, that the analyst has a model for $\Delta_j r_{s,i}(D_i)$ with no error (i.e., $U_{i,j} = 0$), that the analyst knows the distribution of $V_i \equiv (V_{i,1}, \ldots, V_{i,J})$ conditional on observable characteristics, and complete information across firms. Under these additional assumptions this alternative approach leads to a set of moment inequalities that we discuss further in Section 4.1. We refer the reader to Pakes (2010) for a deeper discussion of the differences between the profit inequalities and generalized discrete choice approaches, and Fan and Yang (2022) for an application of the generalized discrete choice approach to a model of product assortment. ■

Remark 3.4. Assumption 3.1 implies that the analyst is able to estimate demand and supply sufficiently well, in order for the differentials in variable profits to have a conditional mean given by $E[\Delta_j r_{s,i}(D_i)|I_s]$. It also assumes that firms make product variety decisions as a function of this conditional mean, rather than following another criterion. If either assumption is not true, $U_{i,j}$ becomes a structural error as well and one could then treat this error similarly to how we treat the structural unobservable $V_{i,j}$ - say, by bounding its conditional expectation as in Assumption 4.2. Here, we choose to maintain Assumption 3.1 for simplicity. ■

4 The moment functions

In order to derive the moment inequalities in (1) from those in (21) and (22), we need assumptions that allow us to deal with the error terms $U_{i,j}$ and $V_{i,j}$, or at least with their expectations conditional on the product offering decision $D_{i,j}$. We do so by discussing the role of the instrumental variables, $Z_{i,j}$, that we introduced in (18). We start with the following assumption.

Assumption 4.1. *The instrumental variables $Z_{i,j}$ satisfy*

$$E[U_{i,j}|Z_{i,j}, D_{i,j}] = 0 \quad \textit{and} \quad E[V_{i,j}|Z_{i,j}] = 0 \ . \tag{23}$$

The requirement in (23) requires $U_{i,j}$ to be mean independent of $Z_{i,j}$ conditional on $D_{i,j} = 1$ (and $D_{i,j} = 0$) and $V_{i,j}$ to be mean independent of $Z_{i,j}$. The additional conditioning on $D_{i,j}$ may be viewed as a reasonable assumption for the specification error $U_{i,j}$ given Assumption 3.1, though in Section 8.2.1 we argue that intuitive arguments that lead to powerful instruments may put this assumption at risk. This may be less reasonable for $V_{i,j}$, as the fact that $V_{i,j}$ belongs to the information set of the firms suggests that

$$E[V_{i,j}|Z_{i,j}, D_{i,j}] \neq 0 \ .$$

That is, one of the difficulties in our setting is that the exogeneity of the instrument needs to hold conditional on products being offered, $D_{i,j} = 1$, and conditional on products not being offered, $D_{i,j} = 0$, and this is unlikely to hold for a structural error term $V_{i,j}$ since, as discussed in Remark 3.2, $E[V_{i,j}|D_{i,j}]$ is expected to be negative. It follows that the instrumental variables $Z_{i,j}$ are not expected to play a significant role in dealing with the problems introduced by the error $V_{i,j}$, unless the analyst is willing impose assumptions on $E[V_{i,j}|Z_{i,j}, D_{i,j}]$. Below we consider one such assumption.

Assumption 4.2. *The conditional expectation $E[V_{i,j}|Z_{i,j}, D_{i,j}]$ satisfies*

$$\Big|E[V_{i,j}|Z_{i,j}, D_{i,j}]\Big| \leq \bar{V}$$

for some known value $\bar{V} \geq 0$.

Assumption 4.2 does not pin down a unique value for $E[V_{i,j}|Z_{i,j}, D_{i,j}]$ but it limits the range of values this expectation can take. A simple sufficient condition for this assumption would be that $V_{i,j}$ is supported on $[-\bar{V}, \bar{V}]$. While this may be a strong condition in certain settings, it is arguably weaker than assuming that sunk costs do not have a structural error component or assuming $\bar{V} = 0$. In particular, it allows for the conditional expectations $E[V_{i,j}|Z_{i,j}, D_{i,j} = 1]$ and $E[V_{i,j}|Z_{i,j}, D_{i,j} = 0]$ to be different from each other, consistent with the discussion in Remark 3.2.

Let $h(Z_{i,j})$ be some known, positive valued function of the instrument. In Section 4.1 we discuss some of the functions $h(\cdot)$ commonly used in applications, including the ones we use in our empirical application. Our goal is to show that Assumptions 4.1 and 4.2 allow us to use the group of inequalities in (21) and (22) to construct a group of moment inequalities, like those in (1), with moment functions $m(W_i, \theta)$ that do not depend on unobserved random variables like $U_{i,j}$ and $V_{i,j}$. We show the details for (21) as (22) follows from similar arguments. Recall that (21) states that

$$(\Delta_j \hat{r}_{s,i}(O_i) - U_{i,j} - (\theta_s + V_{i,j}))(1 - D_{i,j})h(Z_{i,j}) \leq 0 \ ,$$

where we have multiplied each side of the inequality by $h(Z_{i,j})$. Taking expectations,

$$E[(\Delta_j \hat{r}_{s,i}(O_i) - \theta_s)(1 - D_{i,j})h(Z_{i,j})] + E[V_{i,j}(1 - D_{i,j})h(Z_{i,j})] \leq 0 \ , \tag{24}$$

where we used $E[U_{i,j}(1 - D_{i,j})h(Z_{i,j})] = 0$ by Assumption 4.1. The above expression still depends on the unobservable $V_{i,j}$. To deal with this error note that

$$E[V_{i,j}(1 - D_{i,j})h(Z_{i,j})] = E[V_{i,j}h(Z_{i,j})] - E[V_{i,j}D_{i,j}h(Z_{i,j})] \geq -E[\bar{V}D_{i,j}h(Z_{i,j})] \ , \tag{25}$$

where we used $E[V_{i,j}h(Z_{i,j})] = 0$ and $E[V_{i,j}D_{i,j}h(Z_{i,j})] \leq E[\bar{V}D_{i,j}h(Z_{i,j})]$ by invoking by Assumption 4.1, the law of iterated expectations, and Assumption 4.2. It follows from these derivations that (21) leads to the following moment inequality,

$$E[\big((\Delta_j \hat{r}_{s,i}(O_i) - \theta_s)(1 - D_{i,j}) - \bar{V}D_{i,j}\big)\, h(Z_{i,j})] \leq 0 \ , \tag{26}$$

where the term inside the expectation is now a known function of the data and the parameter of interest. We conclude that under Assumptions 4.1 and 4.2, the following pair of moment functions have non-positive expectations,

$$m_j^l(W_i,\theta) \equiv \big(\big(\Delta_j \hat{r}_{s,i}(O_i) - \theta_s\big)(1 - D_{i,j}) - \bar{V} D_{i,j}\big)\, h(Z_{i,j}) \tag{27}$$

$$m_j^u(W_i,\theta) \equiv \big(\big(\Delta_j \hat{r}_{s,i}(O_i) + \theta_s\big) D_{i,j} - \bar{V}(1 - D_{i,j})\big)\, h(Z_{i,j}) \ . \tag{28}$$

These moment functions can be interpreted as re-centered versions of the moment functions that could be derived under a conditional independence assumption of the form $V_{i,j} \perp D_{i,j}|Z_{i,j}$; see Section 4.1 for details. Importantly, these are *known* functions of the data, $\{W_i : i \in \mathcal{N}\}$ in (18), and the parameter of interest, θ in (19).

We conclude that the $k = 2J$ dimensional vector of moment functions is

$$m(W_i,\theta) = (m_1^l, \ldots, m_J^l, m_1^u, \ldots, m_J^u)' \ , \tag{29}$$

where $m(W_i,\theta)$ satisfies $E[m(W_i,\theta)] \leq 0$, as in (1), by the arguments leading to (26).

Remark 4.1. The moment functions in (27) and (28) do not include the error terms $U_{i,j}$ and $V_{i,j}$ by construction to emphasize the fact that $m(W_i,\theta)$ is, by definition, a known function of observed variables, W_i, and the parameters of interest, θ. This makes the notation consistent with the vast majority of theoretical papers that develop inference tools for inference in moment inequality models. ■

4.1 Why? Bias and alternative moment functions

In this paper we have adopted the profit inequalities approach proposed by Pakes (2010) as the main mechanic to obtain moment inequalities. This approach fits our empirical application well and is computationally tractable but is by no means the only alternative we could have followed. To clarify this point, we divide this section into two levels of discussion, where we first discuss alternative steps we could have taken *within* the profit inequalities approach, and then mention how one could have derived inequalities using the leading alternative to the profit inequalities approach, generalized discrete choice models as in Ciliberto and Tamer (2009).

Dealing with bias in the profit inequalities approach

The derivations leading to (26) show that $E[m^l_j(W_i,\theta)] \leq 0$. To see that $E[m^u_j(W_i,\theta)] \leq 0$ as well, consider the following derivation,

$$\begin{aligned} E\left[m^u_j(W_i,\theta)\right] &= E\left[\left(\left(\Delta_j\hat{r}_{s,i}(O_i)+\theta_s\right)D_{i,j} - \bar{V}(1-D_{i,j})\right)h(Z_{i,j})\right] \\ &= E\left[\left(E[\Delta_j r_{s,i}(D_i)|I_s]+\theta_s+V_{i,j}\right)h(Z_{i,j})D_{i,j}\right] \\ &\quad + E\left[U_{i,j}D_{i,j}h(Z_{i,j})\right] - E\left[V_{i,j}D_{i,j}h(Z_{i,j})\right] \\ &\quad - \bar{V}E\left[(1-D_{i,j})h(Z_{i,j})\right] \\ &\leq -E\left[V_{i,j}D_{i,j}h(Z_{i,j})\right] - \bar{V}E\left[(1-D_{i,j})h(Z_{i,j})\right] \\ &\leq 0 \ , \end{aligned}$$

where the first inequality follows from the profit maximizing behavior in (12) and Assumption 4.1. The second inequality follows from

$$\begin{aligned} -E\left[V_{i,j}D_{i,j}h(Z_{i,j})\right] &= -E\left[V_{i,j}D_{i,j}h(Z_{i,j})\right] + E\left[V_{i,j}h(Z_{i,j})\right] \\ &= E\left[V_{i,j}(1-D_{i,j})h(Z_{i,j})\right] \\ &\leq \bar{V}E\left[(1-D_{i,j})h(Z_{i,j})\right] \ , \end{aligned}$$

where the first equality now follows from Assumption 4.1 implying $E\left[V_{i,j}h(Z_{i,j})\right] = 0$ and the inequality follows from Assumption 4.2.

It is important to note that we derive moment inequalities from (21) and (22), as opposed to (20), where both inequalities are combined into one. For convenience, we re-state (20) below imposing $U_{i,j} = V_{i,j} = 0$ to provide a clean intuition for this choice,

$$\Delta_j\hat{r}_{s,i}(O_i) - (1-2D_{i,j})\theta_s \leq 0 \ .$$

Since $(1-2D_{i,j}) = 1$ when $D_{i,j} = 0$ and $(1-2D_{i,j}) = -1$ when $D_{i,j} = 1$, writing the two sets of inequalities separately leads to

$$\frac{1}{n_{0,j}}\sum_{i\in\mathcal{N}_{0,j}}\Delta_j\hat{r}_{s,i}(O_i) \leq \theta_s \quad \text{and} \quad \theta_s \leq \frac{1}{n_{1,j}}\sum_{i\in\mathcal{N}_{1,j}}\Delta_j\hat{r}_{s,i}(O_i) \ , \tag{30}$$

providing a lower and upper bound for θ_s, respectively. Here, we use the notation $\mathcal{N}_{a,j} \equiv \{i \in \mathcal{N} : D_{i,j} = a\}$ and $n_{a,j} \equiv |\mathcal{N}_{a,j}|$. In contrast, the sample average of the single equation above would simply lead to a weighted average of the bounds,

$$\frac{1}{n}\sum_{i\in\mathcal{N}}\Delta_j\hat{r}_{s,i}(O_i) + \left(\frac{n_{1,j}}{n} - \frac{n_{0,j}}{n}\right)\theta_s \leq 0 \ , \tag{31}$$

which is weakly informative about θ_s in the sense that it only provides a lower (or an upper) bound depending on the sign of $n_{1,j} - n_{0,j}$.

To construct the moment functions in (27) and (28), Assumptions 3.1 and 3.2 were insufficient and thus we additionally introduced Assumptions 4.1 and 4.2. Assumption 4.1 essentially requires a variable in the information set I_s of the firms to be mean independent of $V_{i,j}$, since the requirement on $U_{i,j}$ holds for any variable in I_s by Assumption 3.1. Assumption 4.2, on the other hand, deserves further discussion. Perhaps the best way to understand why this particular assumption appears reasonable in the models we consider and why it could be interpreted as weaker than other common alternatives is to start with a conditional independence assumption,

$$V_{i,j} \perp D_{i,j} \mid Z_{i,j} \ . \tag{32}$$

This assumption states that any relationship between $V_{i,j}$ and the decision $D_{i,j}$ is fully captured by $Z_{i,j}$. It follows immediately that $E[V_{i,j}|Z_{i,j}, D_{i,j}] = E[V_{i,j}|Z_{i,j}] = 0$, and from here it follows that we could simply work with the following moment functions,

$$\tilde{m}^l_j(W_i, \theta) = \big(\Delta_j \hat{r}_{s,i}(O_i) - \theta_s\big) h(Z_{i,j})(1 - D_{i,j}) \tag{33}$$
$$\tilde{m}^u_j(W_i, \theta) = \big(\Delta_j \hat{r}_{s,i}(O_i) + \theta_s\big) h(Z_{i,j}) D_{i,j} \ . \tag{34}$$

Here $h(Z_{i,j})$ is again some known, positive valued, function of the instrument. The moment functions in (33) and (34) are functions of (W_i, θ) and do not include unobserved random variables. In addition, note that the moment $\tilde{m}^l_j(W_i, \theta)$ provides a lower bound for the parameter θ while the moment $\tilde{m}^u_j(W_i, \theta)$ provides an upper bound. To see that $\tilde{m}^l_j(W_i, \theta)$ and $\tilde{m}^u_j(W_i, \theta)$ have non-positive expectations under the condition in (32), let's consider $\tilde{m}^u_j(\cdot)$ as the arguments are symmetric for both set of moments. Note that

$$E\left[\tilde{m}^u_j(W_i, \theta)\right] \leq -E\left[V_{i,j} h(Z_{i,j}) D_{i,j}\right] = 0 \tag{35}$$

where the last equality follows from $E\left[V_{i,j} h(Z_{i,j}) D_{i,j}\right] = E[h(Z_{i,j}) D_{i,j} E\left[V_{i,j}|Z_{i,j}, D_{i,j}\right]] = 0$, due to (32). This derivation then shows that a conditional independence assumption actually allows us to derive a simpler set of moment functions, relative to the ones we derived in Section 4; see Ho (2009); Holmes (2011); Houde et al. (2023); Maini and Pammolli (2023) for examples of papers using this approach.[1] However, a concern with these moment functions is that their validity, interpreted here as having non-positive expectations, depends crucially on $E\left[V_{i,j} h(Z_{i,j}) D_{i,j}\right]$ being zero, as illustrated by (35). Whenever the conditional independence assumption fails, we would expect (35) to be strictly positive due to the reasons discussed in Remark 3.2, which then leads to the possibility that $E\left[\tilde{m}^u_j(W_i, \theta)\right] > 0$. We refer to the term $E\left[V_{i,j} h(Z_{i,j}) D_{i,j}\right]$ as the "bias" term induced by the error term $V_{i,j}$.

[1] We include into this category papers that assume $V_{i,j} = 0$

The approach we take in Assumption 4.2 essentially imposes enough structure to bound the magnitude of the bias term $E\left[V_{i,j}h(Z_{i,j})D_{i,j}\right]$. This approach was introduced by Eizenberg (2014). An alternative interpretation of this assumption would be to require that $V_{i,j}$ has known support on $[-\bar{V}, \bar{V}]$, in which case Assumption 4.2 immediately follows. The price we pay for a weaker assumption relative to the condition in (32), which essentially translates to $\bar{V} = 0$, is that $\bar{V}$ is now a tuning parameter that needs to be chosen by the researcher and that is difficult to pin down in a data dependent manner. Yet, the value of $\bar{V}$ matters in applications as it directly enters the moment functions in (27) and (28).

Remark 4.2. An alternative to imposing Assumption 4.2, which requires the choice of the tuning parameter $\bar{V}$, would be to work with the inequalities in (33) and (34) that assume $\bar{V} = 0$, and steer attention to the misspecification robust identified set instead of the original identified set. This approach, that accounts for model misspecification but re-interprets the object of interest, has been recently developed by Andrews and Kwon (2019). At a high level, this approach recenters the "biased" moments using a data-dependent adjustment to the moments that guarantees that (1) holds for at least one value of θ after the moment conditions are properly adjusted. We compare our results with this alternative approach in Section 8.2 and provide additional details on its implementation in Appendix C. ■

Remark 4.3. Assumption 4.2 is not the only strategy that has been used to deal with a structural error term. The most straightforward strategy is to enrich the parameterization of θ and to argue that there is no remaining structural error (Ho (2009); Holmes (2011); Houde et al. (2023); Maini and Pammolli (2023)). A closely related alternative is to use a control function for the structural error term, as discussed in Pakes (2010). Both of these alternatives hinge on the econometrician being able to perfectly parameterize the relevant conditional expectation - if there are any structural error components remaining, the model will be misspecified. Researchers who are considering following these approaches may still benefit from augmenting them with Assumption 4.2. ■

Alternative Approaches to derive moment functions

Assumption 4.2 is not the only strategy that has been used to deal with a structural error term, $V_{i,j}$, and to obtain moment functions with non-positive expectations. A common alternative is differencing (Assumption 4a in Pakes (2010), Crawford and Yurukoglu (2012); Ho and Pakes (2014); Morales et al. (2019)). Under this approach, the researcher assumes that $V_{i,j}$ does not vary across one dimension, products j or markets i, and adds the inequalities in equations 21 and 22 across observations where opposite decisions have been made. This addition cancels out the structural error term. Note that under our specification of sunk costs, this approach would lead to θ cancelling out, but in richer sunk cost models $X_{i,j}\theta_s$ that need not be the case. A similar approach is to work with unconditional inequalities, as in Assumption 4b in Pakes (2010) and Wollmann (2018). We do highlight the fact that this approach can only deal with a structural error component if the realizations of such an error are, in fact, identical across the pairs of firms that are used to form the differences.

Other alternatives include the approach in Dickstein and Morales (2018), that requires a parametric model for the structural unobservable, and the approach in Illanes (2016), that combines inequalities like the one in (20) with Assumption 4.1 to obtain bounds on the parameters of interest using a least favorable distribution for $V_{i,j}$ as in Schennach (2014). The first approach allows researchers to deal with a structural error with unbounded support, but it has only been applied to binary choice settings. The second one of these approaches increases the computational complexity significantly.

Perhaps the most notable alternative approach that we would like to mention is the generalized discrete choice model introduced by Ciliberto and Tamer (2009). As discussed in Remark 3.3, this approach requires additional assumptions on the profit equation, the information set of agents, and the knowledge of the analyst. Under such additional assumptions, this approach can lead to a characterization of the identified set for θ that is "sharp", in the sense that the moment inequalities that characterize $\Theta_0(P)$ contain all of the information assumed in the structure of the model. This feature would not be shared by the profit inequalities strategy we follow in this paper under the same assumptions (thus leading to the so-called "outer" identified sets which are larger than the "sharp" counterparts). We refer the reader to Beresteanu et al. (2011) for a method to construct sharp identified sets in a variety of partially identified models.

To be more concrete, assume that firms know the product portfolio D_i, that $U_{i,j} = 0$, and that the distribution of $V_i \equiv (V_{i,1}, \dots, V_{i,J})$ conditional on observable characteristics is known. Under these assumptions, the generalized discrete choice model leads to 2^{J+1} inequalities given by

$$E[m^l(d, Z, X, \theta)] = \underline{P}\{\theta|d, Z, X\} - \hat{P}\{d|X, Z\} \leq 0 \tag{36}$$

$$E[m^u(d, Z, X, \theta)] = \hat{P}\{d|X, Z\} - \bar{P}\{\theta|d, Z, X\} \leq 0 \ , \tag{37}$$

for each $d \in \{0, 1\}^J$, where $\hat{P}\{d|X, Z\}$ is the observed frequency of product portfolio d (conditional on covariates) across markets,

$$\bar{P}\{\theta|d, Z, X\} \equiv P\{d \text{ is st } \Delta_j \pi_{s,i}(d, \theta) \leq 0 \text{ for all } j \in \mathcal{J} \mid Z, X\}$$

is the probability that the model predicts for d being one (of the possibly many) equilibria of the game, and

$$\underline{P}\{\theta|d, Z, X\} \equiv P\{d \text{ is the only element st } \Delta_j \pi_{s,i}(d, \theta) \leq 0 \text{ for all } j \in \mathcal{J} \mid Z, X\}$$

is the probability the model predicts for d being the unique equilibrium of the game. The probabilities $\bar{P}\{\theta|d, Z, X\}$ and $\underline{P}\{\theta|d, Z, X\}$ are often computed via simulation, for each θ, by taking draws from the known conditional distribution of V_i. For additional details on the differences between the generalized discrete choice approach and the profit inequalities approach, we refer the reader to Pakes (2010).

5 The test

The derivations in the previous two sections led to the moment functions in (29) that in turn deliver the model in (1) under the stated assumptions. We now describe how to test the hypothesis in (4), which we re-state here for readability

$$H_\theta : E[m(W_i, \theta)] \leq 0 \ .$$

In order to do so, define for each $1 \leq \ell \leq k$,

$$\bar{m}_{n,\ell}(\theta) = \frac{1}{n} \sum_{i \in \mathcal{N}} m_\ell(W_i, \theta) \quad \text{and} \quad \hat{\sigma}_{n,\ell}(\theta) = \sqrt{\frac{1}{n} \sum_{i \in \mathcal{N}} (m_\ell(W_i, \theta) - \bar{m}_{n,\ell}(\theta))^2} \ , \tag{38}$$

and let

$$\bar{m}_n(\theta) = (\bar{m}_{n,1}(\theta), \ldots, \bar{m}_{n,k}(\theta))' \ .$$

The test statistic we use to construct our test is

$$T_n(\theta) = \max_{1 \leq \ell \leq k} \frac{\sqrt{n}\bar{m}_{n,\ell}(\theta)}{\hat{\sigma}_{n,\ell}(\theta)} \ . \tag{39}$$

This test statistic is known as a "max"-type test statistic and has been used by Chernozhukov et al. (2019) as it has favorable properties in settings where the number of moment inequalities, k, is large relative to the sample size, N. For this, and other reasons we discuss in Section 5.1, we chose this test statistic out of the many choices of test statistics available in the literature.

Much of the effort in developing a test of H_θ that satisfies (5) lies in the construction of the critical value $c_n(1-\alpha, \theta)$. The difficulty is due to the fact that, in models defined by moment inequalities, the limiting distribution of the usual test statistics, including all the ones we discuss in the next subsection, depends on which and how many of the moment conditions in (1) are in fact equal to zero, i.e., "binding". The literature has provided a variety of ways to circumvent this challenge that can be grouped into four groups that we briefly review in Section 5.1 (for a more comprehensive description of these and other critical values, see Canay and Shaikh, 2017).

Given our choice of test statistic in (39), we again use the approach proposed by

Chernozhukov et al. (2019) and use a critical value that requires two steps.

Step 1: Let $0 < \beta < \alpha/2$ be a tuning parameter and $\Phi(\cdot)$ be the distribution function of the standard normal distribution. Define

$$\hat{k}_n = \sum_{\ell=1}^{k} I\left\{ \frac{\sqrt{n}\bar{m}_{n,\ell}(\theta)}{\hat{\sigma}_{n,\ell}(\theta)} > -2\hat{c}^{\text{lf}}_{n,k}(1-\beta,\theta) \right\} \ , \tag{40}$$

where

$$\hat{c}^{\text{lf}}_{n,k}(1-\beta,\theta) = \frac{\Phi^{-1}(1-\beta/k)}{\sqrt{1-\Phi^{-1}(1-\beta/k)^2/n}} \ . \tag{41}$$

Step 2: Define the critical value of the test as

$$\hat{c}^{\text{ts}}_{n}(1-\alpha,\theta) = \left(\frac{\Phi^{-1}(1-(\alpha-2\beta)/\hat{k}_n)}{\sqrt{1-\Phi^{-1}(1-(\alpha-2\beta)/\hat{k}_n)^2/n}} \right) I\{\hat{k}_n \geq 1\} \ . \tag{42}$$

The resulting test is $\phi_n(\theta)$ in (7) for $T_n(\theta)$ in (39) and $c_n(1-\alpha,\theta)$ in (42). This critical value is fast to compute and does not involve a resampling technique (like, for example, the bootstrap). There are, however, variations that essentially maintain the two step nature of $\hat{c}^{\text{ts}}_{n}(1-\alpha,\theta)$ while using the bootstrap to approximate the quantiles. While there are no formal results that show that the bootstrap provides an asymptotic refinement in models that exhibit discontinuities of the form that are usually present in moment inequality models (and, in general, refinements are not expected in such settings), there is significant numerical evidence that shows that the bootstrap often leads to noticeable power gains (i.e., smaller confidence intervals) in practice. We describe the bootstrap version of the two-step critical value in Section 5.2 and use it in the empirical application of Section 8. In our application, the bootstrap once again leads to improved performance relative to the standard two-step critical value in (42).

5.1 Why? On the choice of test statistic and critical value

On the Test Statistic

The vast majority of tests that have been proposed in the literature on inference in models defined by moment inequalities reject H_θ for large values of a test statistic T that is weakly increasing in each component of the vector m in (1). In order to describe these test statistics succinctly, it is useful to introduce some additional notation. For

$\bar{m}_{n,\ell}(\theta)$ and $\hat{\sigma}_{n,\ell}(\theta)$ as in (38), let

$$\begin{aligned}
\hat{D}_n(\theta) &= \text{diag}(\hat{\sigma}_{n,\ell}(\theta) : 1 \leq \ell \leq k) \\
\hat{\Omega}_n(\theta) &= \hat{D}_n(\theta)^{-1} \frac{1}{n} \left(\sum_{i \in \mathcal{N}} (m(W_i, \theta) - \bar{m}_n(\theta))(m(W_i, \theta) - \bar{m}_n(\theta))' \right) \hat{D}_n(\theta)^{-1} \ .
\end{aligned}$$

The test statistics can be broadly represented as:

$$T_n(\theta) \equiv T\left(\hat{D}_n^{-1}(\theta) \sqrt{n} \bar{m}_n(\theta), \hat{\Omega}_n(\theta) \right) \ , \tag{43}$$

where T is a real-valued function that is weakly increasing in each component of its first argument, continuous in both arguments, and satisfies some additional mild conditions; see Andrews and Soares (2010) for details. The "max" test statistic we choose in (39) satisfies all these conditions. The choice of test statistic traditionally involves a trade-off in two dimensions: computational tractability and power properties. Statistics that do not directly depend on the sample correlation matrix $\hat{\Omega}_n(\theta)$; like the max statistic we use, are computationally attractive but could be less powerful than statistics that use $\hat{\Omega}_n(\theta)$; like the quasi-likelihood ratio proposed by Rosen (2008),

$$T_n^{\text{qlr}}(\theta) \equiv \inf_{t \in \mathbf{R}^k : t \leq 0} \left(\hat{D}_n^{-1}(\theta) \sqrt{n} \bar{m}_n(\theta) - t \right)' \hat{\Omega}_n^{-1}(\theta) \left(\hat{D}_n^{-1}(\theta) \sqrt{n} \bar{m}_n(\theta) - t \right) \ , \tag{44}$$

the adjusted quasi-likelihood ratio proposed by Andrews and Barwick (2012), where $\hat{\Omega}_n(\theta)$ is replaced by

$$\tilde{\Omega}_n(\theta) = \max\{\epsilon - \det(\hat{\Omega}_n(\theta)), 0\} I_k + \hat{\Omega}_n(\theta)$$

for some fixed $\epsilon > 0$, or the empirical likelihood ratio proposed by Canay (2010). The adjustment referred to in the adjusted quasi-likelihood ratio statistic stems from the desire to accommodate situations in which the correlation matrix is (nearly) singular.

A more recent consideration is the behavior of the test statistics in settings with a large number of moment inequalities relative to the sample size, as studied by Chernozhukov et al. (2019) and Bai et al. (2019), among others. The max statistic in (39) is particularly convenient in settings with large k because its quantile grows very slowly with k whereas the quantile of the modified method of moments statistic given by

$$T_n^{\text{mmm}}(\theta) \equiv \sum_{1 \leq \ell \leq k} \max \left\{ \frac{\sqrt{n} \bar{m}_{n,\ell}(\theta)}{\hat{\sigma}_{n,\ell}(\theta)}, 0 \right\}^2 \ ,$$

and used by Andrews and Soares (2010); Ciliberto and Tamer (2009); Bugni (2010) among others, would be expected to grow with k at a faster, polynomial rate. In this paper we take the stance that inference in models defined by moment inequalities is, in most cases, demanding from a computational standpoint and so we do not assign heavy weight on considerations that may improve power at the cost of introducing additional computational burdens or behaving poorly when k is large. For this reason, we use and recommend max-type test statistics as these are straightforward to compute, do not require the analyst to compute and invert $\hat{\Omega}_n(\theta)$, and behave particularly well in applications with a large number of moment inequalities, see Chernozhukov et al. (2019); Bai et al. (2019). Armstrong (2014a) discusses other desirable properties of tests based on this test statistic while Andrews et al. (2019) propose a conditional inference approach for linear conditional moment inequalities that is based on $T_n(\theta)$ in (39) as well.

On the Critical Value

As we mentioned earlier, the construction of a critical value that leads to a test satisfying (5) has been the center of attention in the literature on inference in models defined by moment inequalities. Broadly speaking, critical values can be divided into two groups: those that aim at being sufficiently big without using the data (least favorable), and those that use the data to determine which moments are likely binding and which ones are likely slack (moment selection). Least favorable critical values are simple and often exhibit computational advantages. Critical values involving moment selection are computationally more demanding, but they often lead to more powerful tests. We discuss the basic considerations that are relevant for the setting we consider in this paper below and refer the interest reader to Canay and Shaikh (2017) for a more comprehensive description of these and other critical values.

Least Favorable. Least favorable critical values are based on the observation that considering the distribution that arises when all k moments are binding represents the worst-case or least favorable case; a result that follows from $T_n(\theta)$ being increasing in each component of its first argument and H_θ stating that each component of $E_P[m(W_i, \theta)]$ does not exceed zero. For $T_n(\theta)$ in (39), Chernozhukov et al. (2019) show that the least favorable critical value is given by $\hat{c}^{\text{lf}}_{n,k}(1-\alpha, \theta)$ as defined in (41) but with α replacing β. For other test statistics, like the modified method of moments or the adjusted quasi-likelihood ratio, the least favorable critical value takes a different form, see Canay and Shaikh (2017); Rosen (2008); Andrews and Guggenberger (2009) for details in those cases. Least favorable critical values provide a convenient way to obtain an initial idea of the confidence set C_n as they are one-step, do not require information on which moments are binding or not, and, as a result, are typically fast to compute. Thus, even if the researcher decides to use a more powerful critical value that requires two steps and moment selection, least favorable critical values may still provide the analyst a quick idea of the shape, size, and location of C_n in settings that are computationally demanding. In addition, when the least favorable critical value is paired with the test statistic in (39), Armstrong (2014a) shows that the test is then close to optimal even without moment selection as we introduce next.

Moment Selection. Moment selection, also known as *generalized* moment selection, are critical values that use the data to decide which of the moments in (1) are binding and which ones are slack, in a way that guarantees that the selected number of binding moments is asymptotically conservative. They have been originally introduced in the literature by Andrews and Soares (2010), with closely related ideas appearing in Canay (2010) and Bugni (2014), and later on further refined by Romano et al. (2014) and Chernozhukov et al. (2019), among others. Critical values involving moment selection tend to be smaller than least favorable critical values and thus lead to more powerful tests and smaller confidence regions.

Two-step methods that involve moment selection can be divided in two groups: those that require a tuning parameter that drifts to infinity with the sample size (e.g., the ones in Andrews and Soares, 2010; Bugni, 2014; Canay, 2010, among several others); and those that require a fixed tuning parameter and that account for classification mistakes in the first stage (e.g., the ones in Romano et al., 2014; Chernozhukov et al., 2019; Bai et al., 2019). For example, the generalized moment selection approach proposed by Andrews and Soares (2010) treats the ℓ^{th} moment as binding if

$$\frac{\sqrt{n}\bar{m}_{n,\ell}(\theta)}{\hat{\sigma}_{n,\ell}(\theta)} > -\tau_n \ , \tag{45}$$

where τ_n is a sequence satisfying $0 < \tau_n \to \infty$ and $\tau_n/\sqrt{n} \to 0$. The tuning parameter τ_n is an example of a drifting tuning parameter and a common rule of thumb to set its value is $\tau_n = \sqrt{\log n}$. In contrast, the two-step approach in Chernozhukov et al. (2019) relies on a non-drifting sequence of tuning parameters; that we denoted by β in (42), and adjust the size in the second step to account for possible mistakes in the first stage; which explains why $\hat{c}^{\text{ts}}_{n,k}(1-\alpha,\theta)$ in (42) replaces α with $\alpha - 2\beta$ in the expression of the least favorable critical value. We should note that this is not the only inference approach that requires non-drifting tuning parameters and that allows for a large number of moment inequalities. In particular, the method proposed by Romano et al. (2014) pioneered the use of non-drifting tuning parameters like β for inference in moment inequality models, and has been recently shown to be valid in settings with a large number of moment inequalities by Bai et al. (2019). Overall, the fact that this second group of critical values accounts for the probability that some inequalities may be incorrectly labeled as binding tends to lead to better performance in finite samples. This is perhaps the main reason why we focus here on two-step critical values that rely on non-drifting sequences of tuning parameters.

Remark 5.1. Despite β being a non-drifting tuning parameter, its choice certainly affects the power properties of the test. More concretely, increasing β has two effects. First, increasing β leads to higher values of $\hat{c}^{\text{ts}}_{n}(1-\alpha,\theta)$ in (42) since $1-\alpha+2\beta$ is increasing in β. Second, increasing β lowers the value of $\hat{k}_n$ in (40), which in turn leads to smaller values of $\hat{c}_n^{\text{ts}}(1-\alpha,\theta)$. Since the test statistic T_n does not depend on β, the first effect decreases the power of the method and the second one increases it. Based on simulation evidence, Chernozhukov et al. (2019) recommend the rule of thumb $\beta = \alpha/50$ and we use this as our benchmark choice. ■

Remark 5.2. While the vast majority of recent papers proposing new tests for the problems we review here require tuning parameters for their implementation, there are some tuning parameter free alternatives. For example, Cox and Shi (2019) recently proposed a test that, while involving moment selection, does not involve tuning parameters. The critical value of this test is the quantile of a chi-square distribution with degrees of freedom equal to the rank of the active moment inequalities, where "active" is determined in sample without tuning parameters. Such tuning parameter free critical value necessarily requires the test statistic $T_n(\theta)$ to be the quasi-likelihood ratio test statistic in (44), which by construction requires to compute $\hat{\Omega}_n(\theta)$ and its inverse, and so this type of tuning parameter free critical value does not apply to the max test statistic we opted to choose in this paper. ■

Remark 5.3. In some settings the moment function $m(W_i,\theta)$ may depend on additional point-identified parameters that need to be estimated before the mechanics that are specific to inference in moment inequalities are implemented. This is indeed the case in our empirical application, where the profit differentials, $\Delta_j \hat{r}_{s,i}(O_i)$, require the analyst to estimate demand in the first place; see Section 7. Formally, one could let m depend on an additional parameter ϑ, so that $E[m(W_i,\theta,\vartheta)] \leq 0$, and then redefine $\bar{m}_{n,\ell}(\theta)$ in (38) as

$$\frac{1}{n}\sum_{i\in\mathcal{N}} m_\ell(W_i,\theta,\hat{\vartheta})$$

where $\hat{\vartheta}$ is a consistent and asymptotically normally distributed estimator of ϑ. The asymptotic variance of $\sqrt{n}\bar{m}_{n,\ell}(\theta)$ is different when ϑ is replaced by the estimator $\hat{\vartheta}$ and so $\hat{\sigma}_{n,\ell}(\ell)$ in (38) needs to be defined accordingly, but otherwise the rest of the mechanics remain unchanged; see Andrews and Soares (2010, Section 10.2 and footnote 15) for a discussion. This consideration still applies to cases where ϑ depends on θ, i.e., $\vartheta(\theta)$. Alternatively, one could modify the bootstrap approach in the next section to simultaneously account for demand estimation and product offering decisions (i.e., allowing Step 1(b) to re-estimate ϑ for each bootstrap sample while keeping the same expression for $\hat{\sigma}_{n,\ell}(\ell)$ in (38)); similar in spirit to the recent implementation by Ciliberto et al. (2021). Formal results on the properties of such a bootstrap modification have not been yet derived. ■

5.2 Bootstrap variant

The critical value defined in (42) can be easily computed in closed form and does not require any type of numerical approximation. However, several of the methods developed to construct confidence regions in models defined by moment inequalities end up using some form of bootstrap approximation in the first or second stage (and sometimes in both) in order to obtain more accurate confidence regions. In our empirical application in Section 8 we consider both the asymptotically normal version of the critical value as described in (42) and the bootstrap version, denoted by $\hat{c}_n^{\text{bs}}(1-\alpha,\theta)$, that we describe below following Chernozhukov et al. (2019, Section 4.2.2).

The bootstrap version of the two-step critical value for $T_n(\theta)$ in (39) involves the following three steps:

Step 1 (bootstrap draws): Generate a bootstrap sample $W_{b,1}^*,\ldots,W_{b,n}^*$ as i.i.d. draws from the empirical distribution of $\{W_i : i \in \mathcal{N}\}$.

Step 2 (moment selection): Let $0 < \beta < \alpha/2$ be the tuning parameter discussed in Remark 5.1. Label the moment inequalities as binding or slack via the following steps

(a) Compute $\bar{m}_{n,\ell}(\theta)$ as in (38) using the bootstrap sample and denote it by $\bar{m}_{b,\ell}^*(\theta)$.

(b) Construct the Bootstrap test statistic

$$T_b^*(\theta) \equiv \max_{1\le\ell\le k} \frac{\sqrt{n}(\bar{m}_{b,\ell}^*(\theta) - \bar{m}_{n,\ell}(\theta))}{\hat{\sigma}_{n,\ell}(\theta)} \ . \tag{46}$$

(c) Define the quantile of the bootstrap test statistic as

$$c_n^*(1-\beta,\theta) \equiv \{ \text{ conditional } 1-\beta \text{ quantile of } T_b^*(\theta) \text{ given } \{W_i : i\in\mathcal{N}\} \ \}.$$

(d) Collect the indices of the **binding moment inequalities**,

$$\mathcal{L}_n^* \equiv \left\{ 1 \le \ell \le k : \frac{\sqrt{n}\bar{m}_{n,\ell}(\theta)}{\hat{\sigma}_{n,\ell}(\theta)} > -2c_n^*(1-\beta,\theta) \right\} \ . \tag{47}$$

Step 3 (critical value): Define the critical value of the test as $\hat{c}_n^{\text{bs}}(1-\alpha,\theta)$ via the following steps:

(a) Compute $\bar{m}_{n,\ell}(\theta)$ as in (38) using the bootstrap sample and denote it by $\bar{m}_{b,\ell}^*(\theta)$.

(b) Construct the Bootstrap test statistic

$$T_{b,\mathcal{L}_n^*}^*(\theta) \equiv \left(\max_{\ell\in\mathcal{L}_n^*} \frac{\sqrt{n}(\bar{m}_{b,\ell}^*(\theta) - \bar{m}_{n,\ell}(\theta))}{\hat{\sigma}_{n,\ell}(\theta)} \right) I\{\mathcal{L}_n^* \neq \emptyset\} \ . \tag{48}$$

(c) Define the **critical value** as the quantile of the bootstrap test statistic,

$$\hat{c}_n^{\text{bs}}(1-\alpha,\theta) \equiv \{1-\alpha+2\beta \text{ quantile of } T^*_{b,\mathcal{L}^*_n}(\theta) \text{ given } \{W_i : i \in \mathcal{N}\}\}. \tag{49}$$

The resulting test is $\phi_n(\theta)$ in (7) for $T_n(\theta)$ in (39) and $c_n(1-\alpha,\theta)$ given by $\hat{c}_n^{\text{bs}}(1-\alpha,\theta)$ as described in Step 3 above, i.e.,

$$\phi_n(\theta) = I\left\{\max_{1\le\ell\le k}\frac{\sqrt{n}\bar{m}_{n,\ell}(\theta)}{\hat{\sigma}_{n,\ell}(\theta)} > \hat{c}_n^{\text{bs}}(1-\alpha,\theta)\right\} .$$

There exist a number of variations of the bootstrap implementation we just described that have been discussed in the literature, including the multiplier bootstrap and hybrid methods, among others. We do not discuss these variations here and refer the reader to Chernozhukov et al. (2019) and Bai et al. (2019) for additional details.

Remark 5.4. The bootstrap, in general, should not be expected to provide an asymptotic refinement in models defined by moment inequalities. However, extensive numerical evidence in a variety of papers do tend to show that the bootstrap approximation tends to be more accurate than approximations that are simply based on asymptotic normality, as it is the case in (42). In fact, Chernozhukov et al. (2019, Theorem 4.3) show that when all the moment inequalities are binding, the asymptotic size of the tests based on bootstrap methods coincides with the nominal size α; i.e., they are asymptotically non-conservative. ■

Remark 5.5. The literature on inference in moment inequalities has also considered other types of approximations to the asymptotic distribution of the test statistic, most notably subsampling; see Romano and Shaikh (2008) and Andrews and Guggenberger (2009). Subsampling has the advantages of automatically delivering moment selection and so it is usually implemented in one step: in fact, it is algorithmically the same as its implementation in other, perhaps more traditional, models. It does require the researcher to choose a subsample size, i.e., a number b that is smaller than n and satisfies $b \to \infty$ and $b/n \to 0$ and so, in this sense, it involves a drifting tuning parameter. In simulation studies, subsampling appears to work well for an appropriately chosen subsample size, but may behave poorly in finite samples for other choices of the subsample size. Since good data-dependent rules for choosing b are not currently well developed, we do not devote much attention to subsampling critical values in this paper. ■

6 Confidence intervals

Up until this point we discussed how to test the hypothesis $H_\theta : E[m(W_i,\theta)] \le 0$ for a given value of θ using the test $\phi_n(\theta)$. The duality between hypothesis testing and confidence regions then leads to an immediate characterization of a confidence region for θ, simply by collecting all values of $\theta \in \Theta$ that are not rejected by $\phi_n(\theta)$, i.e.,

$$C_n \equiv \{\theta \in \Theta : \phi_n(\theta) = 0\} \ .$$

Note that, as opposed to more traditional settings where the parameters of interest are point identified and asymptotically normally distributed, the confidence region C_n is not generally available in closed form and is difficult to report as soon as the dimension of θ is larger than 2. However, researchers are often mostly interested in marginal confidence intervals for individual coordinates of θ, either to follow the tradition of standard t-test-based inference or because only few individual coordinates of θ are of interest. The dominant practice consists indeed in a projection of C_n into some, or all, of its coordinates due to be the fact that such objects can be readily reported in standard output tables. Concretely, the marginal confidence interval for the sth coordinate of θ is given by

$$C_n^{\mathrm{s}} \equiv \left[\min_{\theta \in C_n} \mathrm{c}'\theta, \max_{\theta \in C_n} \mathrm{c}'\theta\right] \ ,$$

where c is a vector in $\mathbf{R}^{d_\theta}$ that selects the sth coordinate. Computing C_n^{s} takes C_n as an input, and so it still requires to “invert” the test $\phi_n(\theta)$ by evaluating $\phi_n(\theta)$ over *all* possible values θ takes to then collect the values that are not rejected. Importantly, even in cases where C_n satisfies (3) with equality, the projected confidence interval C_n^{s} is typically conservative in the sense that its asymptotic coverage for θ_s exceeds the normal level $1 - \alpha$.

Test inversion is conceptually simple, but when it comes to computational considerations it presents several challenges to practitioners that are often difficult to address. There are at least three alternative that have been widely discussed and used in the literature to compute C_n and C_n^{s}. The first one is *grid search*. When the dimension of θ is low, a natural way to tackle this problem is by a simple grid search over Θ. Mechanically, the analyst defines a finite set of points in Θ to evaluate $\phi_n(\theta)$, denotes such a grid by Θ_{grid}, and then collects all the points in Θ_{grid} that are not rejected, i.e.,

$$\{\theta \in \Theta_{\text{grid}} : \phi_n(\theta) = 0\} \ .$$

In the specification of Section 8.2 where θ is two-dimensional, we consider a grid with 1401 points for each of the two dimensions of θ, leading to a grid Θ_{grid} with $1401^2 \approx 2 \times 10^6$ evaluation points. While this is certainly tractable and reliable in low dimensional settings, a grid search does not scale up well with the dimension of Θ. This is relevant even for moderate dimensions of Θ, as the number of evaluations points grows exponentially with d_θ (i.e., in our case it leads to 1401^{d_θ}). We deal with this problem in one of the alternative specifications of Section 8.2, where $d_\theta = 6$ and where we instead opt to construct C_n^{s} by solving two non-linear optimization problems as we discuss next.

Finally, in addition to the scaling of the computational complexity, grid search may lead to confidence regions that are non-convex even if $\Theta_0(P)$ is a convex set.

The second alternative is to obtain C_n^{s} directly by *optimization*. To be concrete, let c be a vector in $\mathbf{R}^{d_\theta}$ that selects one of the coordinates of θ; e.g., $\mathrm{c} = (1, 0, \dots, 0)$ would be the vector that selects the first coordinate of θ. When c selects the sth coordinate we may construct C_n^{s} by solving the following two optimization problems,

$$\min_{\theta \in \Theta} \mathrm{c}'\theta \quad \text{subject to} \quad T_n(\theta) \le c_n(1-\alpha, \theta) \tag{50}$$

$$\max_{\theta \in \Theta} \mathrm{c}'\theta \quad \text{subject to} \quad T_n(\theta) \le c_n(1-\alpha, \theta) \ . \tag{51}$$

These problems are generally non-convex, which introduces non-trivial computational challenges. For example, there may not be guarantees that global optima is achievable independently of the starting values used by the optimization algorithm; a point we discuss further in Section 8.2. There are of course settings where these problems are indeed convex and, in fact, even linear. For example, when $T_n(\theta)$ is linear in θ and $c_n(1-\alpha, \theta)$ does not depend on θ, these two problems are not only tractable but can also be solved quite fast with modern computational resources; see Gafarov (2019); Cho and Russell (2018); Andrews et al. (2019) for recent examples along this line. Even when $T_n(\theta)$ is not linear in θ, the computational burden could be reduced by considering a critical value $c_n(1-\alpha, \theta)$ that does not depend on θ, as it is for example the case when $c_n(1-\alpha, \theta)$ is the least favorable critical value $\hat{c}_{n,k}^{\mathrm{lf}}(1-\alpha, \theta)$ defined in (41) with α replacing β. Approximating the critical value following Kaido et al. (2019), as we discuss in the next section, is another way to reduce the number of evaluation points of $c_n(1-\alpha, \theta)$. In the empirical application in Section 8, we found the optimization problems in (50) and (51) to be generally well behaved across languages (Matlab, Python, R) and consistent with a simple grid search. The notable exception was when the parameter θ was a non-linear function of other parameters, as in Section 8.2.3, where the results for $\theta_1(\mu)$ and $\theta_2(\mu)$ all led to different results in `Matlab`, `Python`, and `R`.[2] This illustrates the difficulties with blindly relying on (50) and (51) in non-linear settings.

Finally, a third alternative is to use an approach that directly computes a confidence interval for each coordinate θ_s without computing C_n in the first place; an approach known as *subvector inference* and that leads to confidence intervals that we denote by $\tilde{C}_n^{\mathrm{s}}$. A full description of the existing methods to construct $\tilde{C}_n^{\mathrm{s}}$ requires us to introduce significant additional notation, so we instead refer the reader to the original references in the next section. We note, however, that in some specific settings, including our empirical application, the structure of the model is such that the moment inequalities affecting each coordinate of θ are non-overlapping. That is, the vector of moment

[2] The codes in our Github repository `https://github.com/iacanay/guide-inequalities` replicate these results with simulated data.

functions in (29) admits the partition

$$m(W_i, \theta) = (m_1(W_i, \theta_1)', \ldots, m_{d_\theta}(W_i, \theta_{d_\theta})')' \ ,$$

where each set of moments in $m_s(W_i, \theta_s)$ only depend on θ_s, for $s \in \{1, \ldots, d_\theta\}$. Subvector inference in this special case can be easily done by following the same steps described in Section 5 *separately and sequentially* for each of the non-overlapping moments $m_s(W_i, \theta_s)$. We illustrate this approach in Section 8.1. Leaving aside computational considerations, using subvector inference to construct $\tilde{C}_n^{\text{s}}$ is generally expected to provide narrower confidence intervals for each coordinate relative to C_n^{s} or, conversely, projecting a confidence region C_n for the vector θ into each of its coordinates θ_s is known to lead to conservative confidence intervals.

Taking stock, constructing confidence regions becomes more difficult as the dimensionality of θ grows and this introduces a trade-off that researchers commonly face. On the one hand, using a richer model that controls for many observable characteristics and brings flexibility to variable profit specifications provides certain reassurances that key assumptions, like Assumption 3.1, are likely to hold. But since they increase the dimensionality of θ, the computational burden associated with such additional flexibility increases. On the other hand, a researcher that keeps the dimensionality low in order to keep the computational burden under control is more likely to question the credibility of key assumptions and be concerned about misspecification. Since the presence of flexible structural error terms like $V_{i,j}$ alleviates misspecification concerns, we advocate that researchers work with low-dimensional specifications of the parameter of interest and compensate such simplicity with explicit and flexible assumptions on the structural error term, along the lines of assumptions like Assumption 4.2.

6.1 Why? Computational Considerations

The computational challenges associated with inference in moment inequalities models are perhaps one of the main reasons that prevent a broader adoption of these methods in empirical work. As a result, whereas in the previous sections there were typically several alternatives to the specific choices we make in this paper, when it comes to computational tricks to compute C_n the number of existing alternatives are significantly reduced and restricted to essentially some very recent work in the area. We believe that some of these recent developments look quite promising, but we decided to stick to the "bread and butter" approach to computing confidence intervals in our empirical application given that: (a) this approach applies to a broad range of applications (taking the limitations on dimensionality as given), and (b) it is unclear to us, as of today, which one of the new methods we describe below will get traction in a wide range of applications. We instead discuss some of the most recent papers that we believe contribute to the computational challenges in concrete ways, without delving into the details that are required for the computational gains to kick in.

The literature on inference in conditional and unconditional moment inequalities has recently devoted significant attention to methods, models, and empirical settings that are aimed at reducing the computational challenges that are well known by practitioners in one way or another. Some of these papers present methods that are intended to reduce the computational cost for moment inequality models like those in (1) without imposing additional structure on the moment functions $m(W, \theta)$. Within this class of papers, Bugni et al. (2017), Kaido et al. (2019), and Belloni et al. (2018) proposed methods that are computationally attractive in settings where only a few components of the vector θ are of interest. This is known as a subvector inference problem, where all of the parameters that are not the main parameter of interest are "profiled-out", leading to a test statistic and a critical value that are a function of only the coordinate of interest. These approaches not only typically lead to some degree of computational gains, but they also lead to confidence intervals that are less conservative than projections of the joint confidence regions for the entire vector θ into each of its coordinates. In particular, Kaido et al. (2019) propose confidence intervals for moment inequalities models based on calibrated projections and the so-called E-A-M algorithm, which is related to the family of expected improvement algorithms described in Jones (2001). The authors show that calibrated projections could reduce the computational burden of constructing confidence intervals for θ to a significant extent, as well as providing numerical evidence that adapting the E-A-M algorithm to existing methods could also reduce computational time; see Kaido et al. (2019, Appendix C). Another proposal to improve the computational burden has been to combine frequentists and Bayesian tools, though in most cases these connections have been developed for confidence sets for $\Theta_0(P)$ satisfying (8), as discussed in Section 2. Along these lines, Chen et al. (2018) propose to compute critical values using quantiles of the sample test statistic $T_n(\theta)$ in (39) that are simulated from a quasi-posterior distribution, which requires a prior over the parameter space Θ. The goal of this approach is to benefit from the many existing algorithms, like MCMC or SMC, that are well-developed in the literature on Bayesian computation. Finally, other recent contributions where much of the emphasis relies on improving computational tractability include Cox and Shi (2019) and Syrgkanis et al. (2017).

More recently, several papers have drifted attention to models that impose additional structure in order to obtain a simpler moment function $m(W, \theta)$ that, in turn, can be exploited to improve computational times. This second class of papers has mostly focused on settings where the moment function $m(W, \theta)$ admits a linear representation as a function of θ. For example, Gafarov (2019); Cho and Russell (2018); Andrews et al. (2019) all propose novel methods that exhibit a higher degree of computational tractability when the model involves linear moment inequalities. Despite such linear requirements, these methods are relevant in several empirical settings, including some of the parametrizations we consider in Section 8. Finally, we note that in certain simpler models that were prominent in the early developments of the literature it is even possible to obtain the confidence region C_n in closed form (thereby avoiding the computational challenges previously mentioned altogether). A prominent example of such models are linear models with missing values in the outcome variable, as those studied by Imbens and Manski (2004) and Stoye (2009), among others.

Remark 6.1. Applied researchers are often interested in moving beyond inference on θ in order to study some type of counter-factual analysis, such as simulating equilibrium impacts in alternative settings. While moving to counter-factual analysis introduces certain new challenges, the specifics of their impact may be tightly connected to the details of the particular empirical application under consideration. We therefore go over the current practice and offer some guidance on how to conduct counter-factuals in Section 9, within the context of our empirical application. ■

7 Empirical Application: Preliminaries

Our empirical application focuses on The Coca Cola Company's acquisition of Energy Brands in May 2007 for US$ 4.1 billion. Prior to the acquisition, Energy Brands (also known as Glaceau) sold products under the Glaceau Smartwater and Vitaminwater brands. In turn, The Coca Cola Company produced Dasani water, Powerade sports drinks, and Minute Maid juices, as well as carbonated soft drinks. One potential benefit of this acquisition is that Energy Brands would be able to harness Coca Cola's distribution network, allowing them to enter new markets or to introduce new products to currently served markets. We aim to compare Energy Brands' and Coca Cola's pre-acquisition sunk cost of selling a product in a market. The setting directly fits into the behavioral decision model in Section 3 and the moment inequalities in Section 4.

It is important to note that the point of this paper is not to document the effects of this acquisition, or to make welfare statements. Rather, we study this setting because it helps us shine a brighter light on certain decisions that can be opaque when addressed without an empirical example in mind. Our aim is to guide practitioners on how to use moment inequalities to answer questions such as this one. We do not claim that any estimates below are informative for antitrust policy.

Throughout the rest of this section and the next, we treat the bottled water and fruit drinks market as the relevant product market for the analysis. This implies that the we will limit ourselves to estimating demand and supply equations for products included in Nielsen's Bottled Water, Fruit Drinks - Canned, and Fruit Drinks - Other Container modules. For the merging parties, these modules contain the products sold under the Dasani, Powerade, Glaceau Smartwater and Vitaminwater brands. This decision implies that we will not model substitution and pricing spillovers to other markets where Coca Cola operates, such as carbonated soft drinks and fruit juices. This market definition allows us to keep demand and supply estimation fairly simple, and is not out of line with typical practice in industrial organization. For example, recent papers using the Nielsen data that define markets in a similar fashion include Atalay et al. (2020); Brand (2021) and Döpper et al. (2022).

7.1 Data Construction

We work with price and quantity data obtained from the NielsenIQ Retail Scanner Dataset. This dataset provides scanner data from over 30,000 grocery, drug, and mass merchandise stores in 205 designated market areas (DMAs) throughout the United States. For each universal product code (UPC), NielsenIQ provides sales at the store-week level along with the average price at which the product was sold. We also observe a number of product characteristics, such as size and flavor. We restrict our analysis period to 2006 and the pre-merger months in 2007, a window of around 1.5 years.

We restrict our attention to the main products in the product market, as estimating demand and supply models including all products is likely intractable. This is a common practice when working in markets with a long tail of products with small shares - for example, Nevo (2001); Miller and Weinberg (2017) and Miravete et al. (2018) all make similar restrictions. In particular, we restrict attention to UPCs that have a market share of at least 1% in at least one DMA-month during our sample period. This leaves us with 212 UPCs and 52 brands, which are owned by 34 firms. Across the DMA-months in our sample, the median fraction of sales that is attributed to these $J = 212$ products is 69%, with the 10th percentile being 60%. We obtain ownership information at the monthly level for each of these products from Euromonitor Passport. Table A.1 in Appendix A presents summary statistics for the 10 best-selling UPCs in our sample, as well as for the next 5 top-selling Coca Cola and Energy Brands products.

7.2 Estimation of Variable Profit

In order to evaluate the moment functions in (27) and (28), we need estimated variable profit differentials $\Delta_j \hat{r}_{s,i}(O_i)$, a value for the bounds on the structural error component $\bar{V}$, and the parameter of interest θ_s. In this section we discuss the most relevant considerations behind the estimation of variable profit differentials, including statements of some additional assumptions that are required to properly estimate demand and recover variables affecting demand and marginal costs that are unobserved to the econometrician.

Further details are presented in Appendix B.1. We leave the discussion on the choice of $\bar{V}$ for Section 8, when we move to issues that are specific to moment inequalities.

To estimate variable profit differentials we rely on the standard supply and demand estimation framework in industrial organization. That is, we posit a demand system, make a conduct assumption regarding how firms set prices, and invert first order conditions of the pricing game to recover marginal costs. One of the features of this approach is that it requires the existence of product-market level demand and cost unobservables; see Berry and Haile (2021) for a discussion of the critical role these unobservables play. This, in turn, means that in order to compute counter-factual values of variable profit (and thus obtain the differentials) we need to introduce a few additional assumptions to be consistent with Assumption 3.1. Below we introduce these assumptions formally and discuss why they are needed.

Assumption 7.1. *Variable profits admit the representation*

$$r_{s,i}(O_i) = \sum_{j \in \mathcal{J}_s} M_i D_{i,j}(p_{i,j} - c_{i,j}(\omega_{i,j}))\mathrm{s}_{i,j}(p_i, D_i, X_i, \xi_i) \ , \tag{52}$$

where M_i is the market size for market i, $p_{i,j}$ is the price of good j in market i, $c_{i,j}(\omega_{i,j})$ is the marginal cost of selling good j in market i given a cost shock $\omega_{i,j}$, and $\mathrm{s}_{i,j}(p_i, D_i, \xi_i)$ is the market share of good j in market i given a price vector $p_i \equiv (p_{i,j} : j \in \mathcal{J})$, and vectors of product characteristics that are observed and unobserved to the econometrician, $X_i = (X_{i,j} : j \in \mathcal{J})$ and $\xi_i = (\xi_{i,j} : j \in \mathcal{J})$, respectively.

Assumption 7.2. *Firms make product offering and pricing decisions in a two-stage game. In the first stage, firms decide on product offerings with common knowledge regarding market and product characteristics X_i, the demand function $\mathrm{s}_i(\cdot)$ for any product offering portfolio, the distribution of marginal cost shocks $\omega_{i,j}$, and the distribution of product characteristics.*

First stage*: firms choose a product portfolio to maximize*

$$E[\pi_{s,i}(O_i)] = E\left[\sum_{j \in \mathcal{J}_s} M_i D_{i,j}(p_{i,j} - c_{i,j}(\omega_{i,j}))\mathrm{s}_{i,j}(p_i, D_i, X_i, \xi_i) - D_{i,j} e_{i,j}(\theta) \mid I_s\right] .$$

Second stage*: firms observe product offerings and the realizations of $\xi_{i,j}$ and $\omega_{i,j}$ for each $j \in \mathcal{J}$ that is offered in market $i \in \mathcal{N}$. Provided products are offered, $\xi_{i,j}$ and $\omega_{i,j}$ are common knowledge across firms. Firms then play a Nash-Bertrand pricing game, setting prices in each market i to solve*

$$\max_{\{p_{i,j} : j \in \mathcal{J}_s\}} \sum_{j \in \mathcal{J}_s} M_i D_{i,j}(p_{i,j} - c_{i,j}(\omega_{i,j}))\mathrm{s}_{i,j}(p_i, D_i, X_i, \xi_i) \ . \tag{53}$$

Assumption 7.1 imposes that marginal costs are constant and defines how variable profits are calculated. Assumption 7.2 is more nuanced. In the first stage of the game ξ and ω have not yet been realized, but their distributions are assumed to be known by the firms and unknown to the econometrician. In the second stage the realizations of ξ and ω are only known by firms for products that are offered, and this leads to a situation where the econometrician can only identify, or recover estimates of, these unobservables for products that are offered. This introduces a challenge, as computing one product deviations that add products to a market requires estimates of these unobservables for products that are not offered. Given the assumption that these unobservables are not realized in the first stage, we obtain that firms cannot select on them when making product variety decisions. It then follows that the estimated ω and ξ conditional on products being offered are also informative of ω and ξ for products that are not offered.

Remark 7.1. The fact that unobservables are not known for out-of-sample products is a common issue in the industrial organization literature; see Nevo (2003) for a discussion. In general, what is required is either a model for out of sample unobservables, as we do next, or a modification of Assumption 3.1 that turns $U_{i,j}$ into a structural unobservable, i.e., where the expectation of $U_{i,j}$ conditional on the product offering decision is not equal to the unconditional expectation. ■

Given these assumptions, we can estimate demand and recover the vector of realized $\xi_{i,j}$ for each product $j \in \mathcal{J}$ that is offered in market $i \in \mathcal{N}$. While this step can be done in a variety of ways, we use a nested logit specification as described in Appendix B.1. The first order conditions for the pricing game can then be inverted to recover marginal cost realizations $c_{i,j}(\omega_{i,j})$ and, given these realizations, the vector of realized $\omega_{i,j}$ can be recovered; see Appendix B.2 for details. Armed with these realizations, we assume the following.

Assumption 7.3. *A market i is a combination of a DMA a and a month t, so that $i = (a, t)$. The model for product characteristics that are unobserved by the econometrician is*

$$\xi_{i,j} = \xi_{j,a} + \xi_{a,t,j} \quad \textit{where} \quad E[\xi_{a,t,j}|p_{a,t,j}, \xi_{j,a}] = 0 \quad \textit{and} \quad \xi_{a,t,j} \perp D_i \ .$$

The model for marginal cost realizations is

$$\ln(c_{a,t,j}) = \omega_{j,a} + \omega_{a,t,j} \quad \textit{where} \quad E[\omega_{a,t,j}|\omega_{j,a}] = 0 \quad \textit{and} \quad \omega_{a,t,j} \perp D_i \ .$$

The restrictions on the distribution of these unobservables allows us to use the empirical distribution of $\xi_{a,t,j}$ and $\omega_{a,t,j}$ for offered products as the empirical distribution for products that are not offered. More precisely, recall that Assumption 3.1 requires $\Delta_j \hat{r}_{s,i}(O_i) = E[\Delta_j r_{s,i}(D_i)|I_s] + U_{i,j}$, with $E[U_{i,j}|D_{i,j}] = E[U_{i,j}|I_s] = 0$. Since $E[\Delta_j r_{s,i}(D_i)|I_s] = E[r_{s,i}(\partial_j D_i)|I_s] - E[r_{s,i}(D_i)|I_s]$, to calculate expected variable profits for a given product portfolio we draw from the empirical distribution of $\xi_{j,a} + \xi_{a,t,j}$ and of $\omega_{j,a} + \omega_{a,t,j}$, solve for optimal prices given these draws, and compute variable profits following (52). Appendix B.3 contains all remaining details.

Remark 7.2. Assumption 7.3 is internally consistent with the rest of the model only if firms do not observe $\xi_{a,t,j}$ and $\omega_{a,t,j}$ before deciding on product offerings. Otherwise, these decisions will be a function of the values of these unobservables, and the conditional distribution will differ from the unconditional. ■

8 Empirical Application: Moment Inequalities

We now return to the moment functions defined in (27)-(28), which we re-state here for readability:

$$
\begin{aligned}
m^l_j(W_i,\theta) &= \big(\big(\Delta_j \hat{r}_{s,i}(O_i) - \theta_s\big)(1 - D_{i,j}) - \bar{V} D_{i,j}\big)\, h(Z_{i,j}) \\
m^u_j(W_i,\theta) &= \big(\big(\Delta_j \hat{r}_{s,i}(O_i) + \theta_s\big) D_{i,j} - \bar{V}(1 - D_{i,j})\big)\, h(Z_{i,j}) \ .
\end{aligned}
$$

The previous section explained how to compute variable profit differentials $\Delta_j \hat{r}_{s,i}(O_i)$. This section describes our approach to dealing with $\bar{V}$ and discusses the role of instruments $Z_{i,j}$. We then describe and discuss alternative approaches.

Before moving on, it is important to note that the parameter θ_s only enters into moments associated with the $\mathcal{J}_s$ products offered by firm s and that these moments do not depend on $\theta_{s'}$ or $\mathcal{J}_{s'}$ for $s' \neq s$. Since our goal is to compare the expected sunk costs of offering a product in a market for Coca-Cola and Energy Brands, denoted by θ_1 and θ_2 respectively, this implies that we do not need to estimate θ_s for $s > 2$. As a consequence, the total number of relevant products becomes 31, which amounts to the products offered by these two firms, and the total number of moment inequalities is $62 = 31 \times 2$ when $h(Z_{i,j}) = 1$, i.e., there are no instruments. From here on, $J = 31$ and $\mathcal{J} = \mathcal{J}_1 \cup \mathcal{J}_2$. This contrasts with the 212 products that we used to estimate $\Delta_j \hat{r}_{s,i}(O_i)$. Note, however, that the same methodology could be used to estimate expected sunk costs for any other firm in the market.

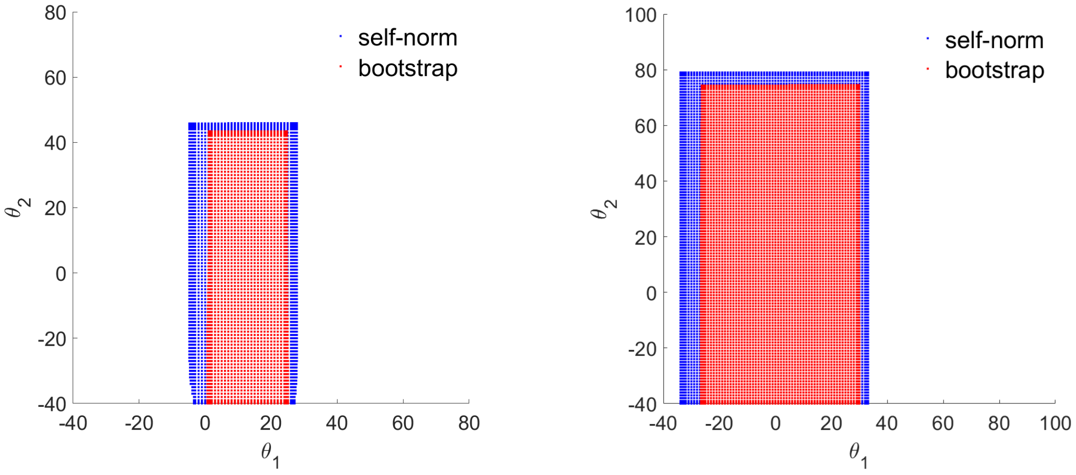
8.1 Main Specification and Results

Our parameters of interest are θ_1 and θ_2, the expected sunk costs of offering a product in a market for Coca-Cola and Energy Brands, respectively. Since there are two moment inequality functions for each product, this leads in theory to $62 \times N_Z$ moment inequalities, where we denote the number of instruments by N_Z. In practice, however, there are fewer moment inequalities since some Coca Cola products are offered in all markets and so $m^l_j(W_i,\theta) = 0$ for all $i \in \mathcal{N}$ for such products. A product that is always offered is a product whose expected marginal contribution to variable profit always covers the expected sunk cost. Thus, it delivers information about an upper bound for sunk cost but is uninformative about a lower bound. Because of this, we ignore lower bound moments for products that are always offered. This leaves us with 41 moments for Coca Cola and 14 moments for Energy Brands, for a total of $55 \times N_Z$ moment inequalities.

We construct confidence intervals for θ_1 and θ_2 following the steps described in Section 6, using projections of the 2-dimensional confidence region for $\theta = (\theta_1, \theta_2)$ into each of its coordinates, as well as using subvector inference for each coordinate separately. In the case of projections, we operationalize a grid search over θ in two steps. In the first step, we use a grid between $-40,000$ and $100,000$ dollars in steps of $1,000$ for each coordinate and evaluate the firm's moment functions for each value in the grid. This leads to 141^2 evaluation points. In the second step, we refine the grid in steps of 100 dollars around the bounds obtained in the first step. This adds an additional $40,000$ evaluation points. For each value in the grid, we compute the test statistic in equation (39), and compare it to the critical value $c_n(1-\alpha, \theta)$ defined in equation (42), with $\alpha = 0.05$. We also compute confidence regions using the bootstrap variant of the test statistic introduced in Section 5.2. One convenient feature of the setting we consider in this application is that the expected variable profit differentials are not a function of θ and, as a result, they can be computed once and saved, rather than re-computed for different θ values. In the case of subvector inference, we exploit another convenient feature of our setting that simplifies the inference problem. Each moment is only a function of a particular firm's θ, but not of both. This implies that one can separate the problem into two: the problem of estimating Coca-Cola's expected sunk costs and the problem of estimating Energy Brand's expected sunk costs. We then operationalize the grid search over each θ_s using a grid between $-40,000$ and $100,000$ dollars in steps of 100, leading to 1401 evaluation points for each θ_s; a substantial reduction relative to the projection approach. This approach, which we label "partitioning", is analogous to the subvector inference approach we discussed in Section 6.1.

Table 1 presents results for Coca Cola's and Energy Brands' sunk costs for two values of $\bar{V}$: $500,000$ dollars and 1 million dollars per product-month. Panel A presents results for projections and Panel B presents results for partitioning. As expected, projections into individual coordinates are expected to be conservative and so the results in Panel B show narrower confidence intervals relative to those in Panel A, but not by much. The computational gains of exploiting the partitioning feature present in our setting are remarkable, cutting down computing time by a factor of 10. Under the first value of $\bar{V}$, for the self-normalized version of the critical value (as defined in (42)), we find a confidence interval for Coca Cola's sunk cost of offering a product in a given city in a particular month of between $-3,200$ and $27,000$ (Panel B). Using the bootstrap version of the critical value (as defined in (49)) yields a slightly narrower interval of between $4,300$ and $23,300$ (Panel B). The cost of this improvement is a 2.5 times longer run time. However, we can still recover these bounds in roughly 4 seconds, which illustrates the convenience of having separability of moments in θ and working with a model where the dimensionality of θ is small. In general, researchers working with relatively low-dimensional models would benefit from using the bootstrap version of the critical value. The confidence region is much less informative for Energy Brands, where the expected sunk cost is between $-40,000$ and $41,900$ (Panel B). Since $-40,000$ is the lower bound of our grid, we do not find an informative lower bound on Energy Brand's sunk cost. Turning to results when $\bar{V} = 1,000,000$, we find that the upper bound of the confidence region for Coca Cola products is slightly higher. However, the lower bound for Coca Cola products falls significantly, and the upper bound for Energy Brands products increases substantially. This illustrates that Assumption 4.2 plays an important role and highlights that dealing with structural unobservables in these types of models requires making assumptions that can drastically affect empirical estimates. Thus, empirical researchers using these tools should state their assumptions clearly, discuss why they are sensible, and document how results change as these assumptions are relaxed. The last column in Table 1 reports computational time in seconds, accounting for the fact that we carried all our computations using Northwestern's High-Performance Computing (HPC) cluster, Quest.[3]

Figure 1 presents the same results in Panel A of Table 1 in graphical form. In this figure, blue points denote the 95% confidence set obtained using self-normalized critical values, while red points denote the set obtained using bootstrap critical values.



(a) $\bar{V} = 500$

(b) $\bar{V} = 1000$

Figure 1: 95%-confidence regions for θ_1 and θ_2 as reported in Panel A of Table 1.

[3] Concretely, each row of each table were computed in one Intel Xeon Gold 6132 2.6 GHz node with 28 cores (3.4 GB GB memory per node); see exact specifications at https://www.it.northwestern.edu/departments/it-services-support/research/computing/quest/specs.html.

	Crit. Value	θ_1: Coca-Cola	θ_2: Energy Brands	Comp. Time
		Panel A: Projections		
$\bar{V}$=500	self-norm	[-4.8 , 27.8]	[-40.0 , 45.9]	16.8
	bootstrap	[0.7 , 25.1]	[-40.0 , 43.4]	49.4
$\bar{V}$=1000	self-norm	[-33.9 , 33.1]	[-40.0 , 79.0]	19.3
	bootstrap	[-26.6 , 30.0]	[-40.0 , 74.5]	58.7
		Panel B: Partitioning/Subvector		
$\bar{V}$=500	self-norm	[-3.2 , 27.0]	[-40.0 , 41.9]	1.6
	bootstrap	[4.3 , 23.3]	[-40.0 , 39.4]	4.0
$\bar{V}$=1000	self-norm	[-31.8 , 32.2]	[-40.0 , 71.9]	1.7
	bootstrap	[-23.3 , 28.9]	[-40.0 , 67.5]	3.9

Table 1: 95%-confidence intervals for θ_1 and θ_2. The self-normalized and bootstrap critical values are defined in (42) and in Section 5.2, respectively. The parameter space for both parameters is $[-40, 100]$ where units are in thousands of US dollars. Computational time is presented in seconds. Panel A presents results by projecting a 2-dimensional confidence region and Panel B presents results by partitioning moments for each θ_s.

Remark 8.1. In many settings, economic theory provides restrictions on the parameter space Θ that should be accounted for when computing confidence sets. For example, in this application it is natural to restrict θ to be non-negative. We, however, decided not to impose this restriction for pedagogical reasons, as it allows us to better illustrate how our results change as $\bar{V}$ changes and as instruments are brought in. ■

8.2 Why? Alternative Specifications

Having established baseline results, we now highlight three topics for further discussion. First, the bounds obtained in the previous section are quite large, and it is natural to ask whether instruments can be used to tighten them. We explore how instruments change our results, and highlight some common issues in the first subsection. Second, we explore the implications of ignoring selection bias in the structural unobservable by setting $\bar{V} = 0$, as well as ways to deal with misspecification by re-centering test statistics or by using more recent misspecification robust methods. Finally, we discuss challenges that arise when estimating more complex models where the specification for the expected sunk cost of offering a product includes covariates. In order to keep tables concise, we only present results for the case where we partition the moments for each θ_s, so the results below are comparable with those in Panel B of Table 1.

8.2.1 Instrumental Variables

Equations (27) and (28) highlight that instruments operate as weights in this setting, increasing the importance of the value of the moments for some observations relative to others. This may lead to tighter bounds. In applications similar to the one we consider here, the natural instruments are demand shifters such as market demographics (Ho, 2009; Pakes et al., 2015), as one can argue that these shift expected variable profit without shifting sunk costs. Other commonly used instruments are product characteristics, either from the same firm or its competitors (Ho, 2009; Wollmann, 2018). The rationale behind such instruments is that sunk costs of offering a product do not vary as a function of the product's characteristics, or of the characteristics of what rivals are offering, while variable profits do. Naturally, whether these arguments are reasonable or not is specific to the setting of interest. Finally, shifters of marginal cost that do not shift the structural unobservable also serve as valid instruments but, again, whether or not a given candidate variable would satisfy these condition is context-dependent. In this section we present results based on market demographics and then discuss other alternatives briefly.

Let $Z_{i,j} = (Z_i^{(1)}, Z_i^{(2)}, Z_i^{(3)})'$ be a vector of the following three random variables at the market level: (a) employment rate $Z_i^{(1)}$, (b) average income in market $Z_i^{(2)}$, and (c) median income $Z_i^{(3)}$. The four instruments we use in this section are the following:

$$h(Z_{i,j}) = \{\text{constant}, Z_i^{(1)}, I\{Z_i^{(2)} > \text{median}(Z_i^{(2)})\}, I\{Z_i^{(3)} > \text{median}(Z_i^{(3)})\}\} \ . \qquad (54)$$

We work with binary instruments in the cases of $Z_i^{(2)}$ and $Z_i^{(3)}$ following standard practice, e.g., Ho (2009); Holmes (2011); Eizenberg (2014); Wollmann (2018); Houde et al. (2023). There are three features of working with indicator variables as instruments that are worth highlighting. First, in the case of binary instruments working with unconditional expectations (or moments) is equivalent to working with conditional ones, up to scalar multiplication. To see this, suppose that $m(W,\theta) = g(O,\theta)I\{Z \in A\}$ for some event A and note that

$$E[m(W,\theta)] = E[g(O,\theta)|Z \in A]P\{Z \in A\} \propto E[g(O,\theta)|Z \in A] \ .$$

This means that binary instruments are isomorphic to models that do not use instruments but instead take expectations of the moment functions conditional on subsets of the data. For example, taking expectations over equations (27) and (28) separately for markets with average income above/below the national median is identical to using indicators for average income above/below the national median as instruments. The flip-side of this argument implies that the common practice of computing averages using only certain subsets of the data is equivalent to using indicators for such subsets as instruments and thus should be followed by a discussion of whether such an instrument should be expected to be valid or not.

Second, binary instruments lead to simple intuition for their relevance (or power) but may trap researchers into focusing too much on instrument relevance while putting instrument validity at risk. To see this point, let $\mathcal{N}_{a,j} \equiv \{i \in \mathcal{N} : D_{i,j} = a\}$ and manipulate (27) to see that the binary instrument $I\{Z \in A\}$ produces a tighter lower bound whenever

$$\frac{\sum_{i \in \mathcal{N}_{0,j}} \Delta_j \hat{r}_{s,i}(O_i) I\{Z_{i,j} \in A\} - \bar{V} \sum_{i \in \mathcal{N}_{1,j}} I\{Z_{i,j} \in A\}}{\sum_{i \in \mathcal{N}_{0,j}} I\{Z_{i,j} \in A\}} \tag{55}$$

is high. That is, instruments that place higher weight on observations associated with unserved markets with larger estimated variable profit gains from counter-factually adding the product will lead to a tighter lower bound for the parameter θ. This insight may lead to a practice where researchers search for events A that mechanically lead to the desired higher weights, while ignoring the important property in Assumption 4.1 that the instrument, $I\{Z_{i,j} \in A\}$, must be mean independent from the error $U_{i,j}$ associated with the estimation of the variable profit differentials; i.e., $E[U_{i,j}|Z_{i,j} \in A, D_{i,j}] = 0$. These two requirements on the instruments introduce tension and illustrate the challenges in finding instruments that are both informative and valid. An analogous phenomenon arises with the upper bound on θ.

Finally, when dealing with binary instruments of the form $I\{Z_{i,j} \in A\}$, it is advised to report sample sizes for every moment that uses an indicator variable as an instrument, or at least the minimum sample size over all the moments. This is due to the fact that it is not unusual for the events $I\{Z_{i,j} \in A, D_{i,j} = 1\}$ or $I\{Z_{i,j} \in A, D_{i,j} = 0\}$ to happen with low probability for some $j \in J$, and all the formal arguments behind the desirable properties of the inference methods described in Section 5 rely on proper law of large numbers and central limit theorems for each of the moments.

Table 2 presents 95% confidence intervals for θ_1 and θ_2 using the instruments in (54). Similarly to Table 1, we present results for two values of $\bar{V}$ and two critical values (self-normalized and bootstrap). Note that the number of moment inequalities increases linearly with the dimension of the instruments, N_Z, and this means that there are 55 moment inequalities in Table 1 and 220 in Table 2. The immediate consequence is that the least favorable critical value, $\hat{c}^{\text{lf}}_{n,k}(1-\beta,\theta)$ in (41), increases as the number of inequalities, k, gets larger and the self-normalized critical value $\hat{c}^{\text{ts}}_{n}(1-\beta,\theta)$ in (42) (that includes moment selection) also increases due to the increase in $\hat{k}_n$. If the instruments are not very powerful, this means that adding instruments to the baseline specification may not lead to tighter inference, as it is the case in Table 2 here. In the case of Coca-Cola, the confidence interval gets wider for both values of $\bar{V}$ for the self-normalized critical value, whereas for Energy-Brands only the upper bound is affected since the lower bound

	Crit. Value	θ_1: Coca-Cola	θ_2: Energy Brands	Comp. Time
$\bar{V} = 500$	self-norm	[-11.5 , 31.1]	[-40.0 , 45.7]	18.0
	bootstrap	[4.3 , 23.3]	[-40.0 , 39.4]	24.0
$\bar{V} = 1000$	self-norm	[-40.0 , 36.9]	[-40.0 , 78.6]	17.2
	bootstrap	[-23.3 , 28.9]	[-40.0 , 67.5]	23.4

Table 2: 95%-confidence intervals for θ_1 and θ_2 using a constant and indicators variables based on demographics (employment rate, average and median household income) as instruments. The self-normalized and bootstrap critical values are defined in (42) and in Section 5.2, respectively. The parameter space for both parameters is $[-40, 100]$ where units are in thousands of US dollars. Computational time is presented in seconds. The minimum sample size over all the moments that use an indicator variable is 99.

was already uninformative in Table 1 and the additional instruments did not bring additional identification power. When it comes to the bootstrap critical value, Table 2 illustrates how the bootstrap may be less affected by the presence of additional (and mostly uninformative) moments and leads to confidence intervals that are essentially identical to those in Table 1.

While the instruments in (54) do not lead to tighter inference on θ_1 and θ_2, they lead to some important takeaways. First, they highlight that the computational benefits of simple and fast critical values, like the self-normalized ones, usually comes at the cost of conservative inference and so variants like the bootstrap are worth considering when the computational burden is manageable. Second, the instruments $Z_{i,j}$ are required to satisfy Assumptions 4.1 and 4.2, which illustrates the inter-dependence between the list of instruments and the value of $\bar{V}$ in Assumption 4.2.

8.2.2 Setting V to zero and dealing with misspecification

We now consider the results obtained by assuming that there is no structural unobservable V, which is equivalent to assuming $\bar{V} = 0$ in Assumption 4.2. In our application we have assumed that within firm expected sunk costs are homogeneous for all product-city pairs and so we would expect the model with $\bar{V} = 0$ to be misspecified. More generally, such a model would be misspecified whenever firms take into account more information than what has been included in the model for sunk costs. As discussed earlier, if a model with $\bar{V} = 0$ is misspecified then this necessarily means that increasing the dimension of $Z_{i,j}$ would not alleviate the issue, so here we focus on the case where $Z_{i,j}$ is just a constant.

Table 3 presents three alternative ways to present results in models with $\bar{V} = 0$.

Test Stat.	Crit. Value	θ_1: Coca-Cola	θ_2: Energy Brands	Comp. Time
CCK	self-norm	[16.0 , 25.0]	[-1.0 , 17.9]	3.0
RC-CCK	self-norm	[-11.3 , 48.9]	[-1.0 , 17.9]	12.7
RC-CCK	bootstrap	[-11.7 , 45.0]	[-1.2 , 16.5]	14.4
RC-CCK	SPUR1	[-40.0 , 100.0]	[-2.8 , 58.0]	14.6

Table 3: 95%-confidence intervals for θ_1 and θ_2 assuming $\bar{V} = 0$ (no structural unobservable). † reports θ that minimizes the test statistic instead of an empty confidence interval. CCK and RC-CCK are the max test statistic in (39) and its re-centered version, respectively. The self-normalized, bootstrap, and SPUR1 critical values are defined in (42), in Section 5.2, and in Section C, respectively. The parameter space for both parameters is $[-40, 100]$ where units are in thousands of US dollars. Computational time is presented in seconds.

The first row presents results that are analogous to those in Table 1. The model leads to tighter confidence regions for both companies, relative to Table 1. This is consistent with how model misspecification typically manifests in partially identified models, where one of two possible situations arise: (a) misspecification may lead to what is known as spurious precision; i.e., a false sense of precision, as discussed in Andrews and Kwon (2019), or (b) misspecification may lead to a case where the confidence interval for θ_s is empty. When the confidence interval for θ_s is empty, it is not uncommon for applied researchers to simply report the value of θ_s that minimizes the test statistic, though we argue that reporting an "NA" provides a more accurate characterization of an empty confidence set in such situations. Here we did not obtain empty confidence intervals, but rather a situation possibly associated with spurious precision instead. The second and third rows of Table 3 report confidence intervals associated with a re-centered (RC) version of the test statistic, using either the self-normalized or the bootstrap critical values. This is common practice when no parameter value can rationalize the data, as re-centering, by construction, guarantees that at least one solution exists (i.e., the point(s) that minimizes the test statistic). To be concrete, the test in this case becomes

$$\phi_n^{\text{rc}}(\theta) = I\left\{T_n^{\text{rc}}(\theta) > c_n(1-\alpha,\theta)\right\} \quad \text{for} \quad T_n^{\text{rc}}(\theta) \equiv T_n(\theta) - \min_{\theta\in\Theta} T_n(\theta)\ .$$

Importantly, in those situations where the confidence regions without re-centering are empty, using $\phi_n^{\text{rc}}(\theta)$ instead would mechanically lead to a non-empty confidence region.

Table 3 illustrates that inference based on a misspecificed model that imposes $\bar{V} = 0$ and uses re-centering may be less informative than a model where $\bar{V} = 500$, cf. Table 1 in the case of Coca-Cola. Table 3 also illustrates that re-centering should not be expected to alleviate spurious precision in general, as it is the case for Energy Brands (since in this case $\min_{\theta\in\Theta} T_n(\theta) = 0$). This motivated the recent proposal by Andrews and Kwon (2019), who take concerns about poor properties of re-centering as a starting point to develop a method that changes both the test statistic and the critical value. We report the results associated with their misspecification robust method, denoted by SPUR1, in row 4 of Table 3 and present details associated with its implementation in Appendix C. Overall, confidence regions for this approach are noticeably wider than in the re-centered versions, highlighting that spurious precision may still be present after a simple re-centering. Importantly, comparing the first and last row for Energy Brands illustrates that even in settings where the confidence intervals are non-empty and $\min_{\theta\in\Theta} T_n(\theta) = 0$, a misspecification robust method like SPUR1 may widen the confidence intervals, see Remark 8.2 below. Conceptually, in our context this method finds the minimum value of $\bar{V}$ that would make all the inequalities hold for some θ and then corrects critical values to account for the fact that such a value of $\bar{V}$ is data-dependent (which increases the critical values). Since we compute our confidence intervals separately for each θ_s, the resulting adjustment is different for each θ_s and the implicit value of $\bar{V}$ is sufficiently big that the SPUR1 confidence interval for Coca-Cola ends up being uninformative.

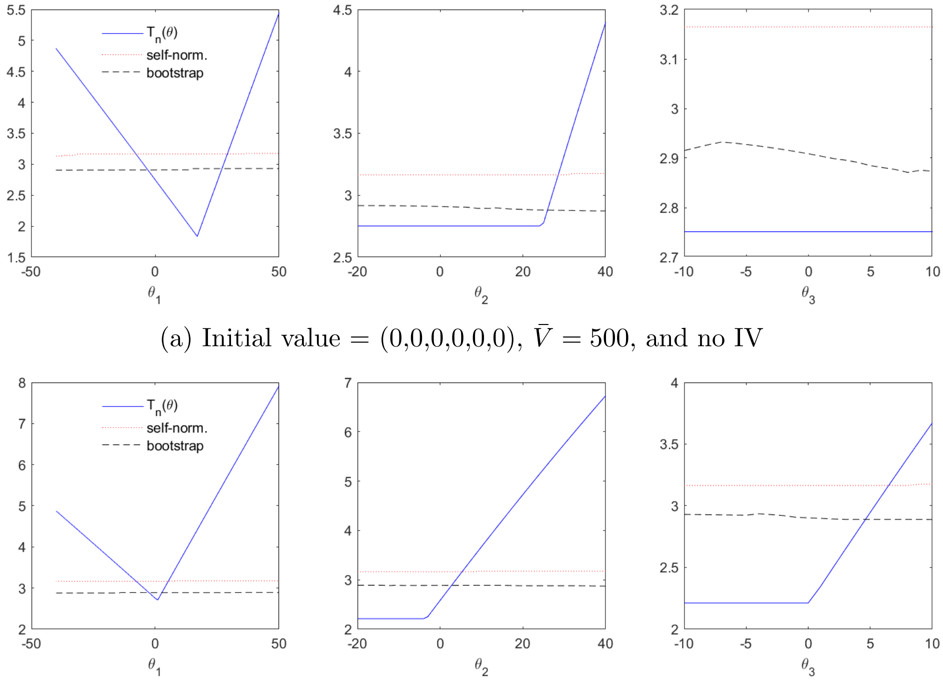
Remark 8.2. There are two features of the method proposed by Andrews and Kwon (2019) that are worth highlighting. First, it is important to understand that SPUR1 delivers a confidence set for the misspecification-robust identified set, which is the minimal enlargement of $\Theta_0(P)$ that makes it non-empty by construction (see (A.14) in Appendix C). When the model is correctly specified, the two identified sets coincide. Second, while the method delivers a data-dependent value of $\bar{V}$, defined to be the minimum value that makes all moment inequalities hold, in sample, for at least one value of θ, such data-dependent value of $\bar{V}$ does not admit a natural economic interpretation and the true model could be one with a smaller or larger value of $\bar{V}$. An immediate consequence of this is that while we have illustrated the approach in a setting with $\bar{V} = 0$, it should be clear to the reader that models with $\bar{V} > 0$ could also be misspecified and that SPUR1 tests could be considered in those settings. In other words, SPUR1 tests could be considered in any instance where there are concerns about misspecification and not limited to cases where the estimated confidence sets are empty. ■

8.2.3 Alternative specification for sunk costs

The baseline model for sunk costs we use is simple relative to the standards of empirical work. This is intentional, as it lets us focus on implementation issues. However, researchers may want to have richer models, either because the question of interest requires it or because dealing with a structural unobservable requires a richer parameterization. In this subsection, we discuss a richer parametrization by considering the following alternative specification

$$e_{i,j}(\theta) = X'_{i,j}\theta_s + V_{i,j} \ ,$$

where $X_{i,j}$ is a vector of the distance between the factory where the product is produced and the designated market area. We consider both, a linear specification $X_{i,j} = (1, d_{i,j})'$, and a quadratic specification $X_{i,j} = (1,\ d_{i,j},\ d^2_{i,j})'$, where distance $d_{i,j}$ is measured in thousands of miles. To illustrate difficulties associated with a larger dimension, in this section we do not partition the problem for each firm and rather work with all the moment inequalities and both firms simultaneously; see the discussion in Section 8.1. This leaves a 4 dimensional parameter in the linear model and a 6 dimensional parameter in the quadratic model. In practice, we stress that researchers should partition their models whenever possible.



(a) Initial value = (0,0,0,0,0,0), $\bar{V} = 500$, and no IV

(b) Initial value = (10,10,10,10,10,10), $\bar{V} = 500$, and no IV

Figure 2: Test statistic and self-normalized and bootstrap critical values as a function of θ.

To find confidence regions for each parameter value, we work with the optimization problems defined in equations (50) and (51) of Section 6. For example, to find bounds for θ_1 in the linear case, c $= (1, 0, 0, 0)'$. The main challenge in this setting is the fact that $T_n(\theta) - c_n(1-\alpha, \theta)$ need not be convex. As a result, minimization techniques geared for convex problems may not work well. We found it useful to plot the behavior of $T_n(\theta)$ and $c_n(1-\alpha, \theta)$ when varying a single dimension of θ and holding the others fixed; see Figure 2. In this setting it appears to be the case that $T_n(\theta)$ is sufficiently well behaved for standard minimization software, such as fmincon or Knitro, to likely find a minimum. The behavior of the bootstrap version of $c_n(1-\alpha, \theta)$ is more erratic, and could lead to local minima. If this is the case, reported confidence regions would be expected to be

		$\bar{V} = 500$		$\bar{V} = 1000$	
	parameter	linear	quadratic	linear	quadratic
	$\theta_{1,1}$	[-8.3 , 50.1]	[-8.3 , 64.6]	[-38.1 , 55.7]	[-38.1 , 70.2]
Coca	$\theta_{1,2}$	[-20.0 , 37.0]	[-20.0 , 50.0]	[-20.0 , 50.0]	[-20.0 , 50.0]
Cola	$\theta_{1,3}$		[-10.0 , 10.0]		[-10.0 , 10.0]
	$\theta_1(\mu)$	[-32.0 , 35.2]	[-45.7 , 37.1]	[-63.3 , 42.8]	[-75.5 , 42.7]
	$\theta_{2,1}$	[-40.0 , 60.5]	[-40.0 , 66.7]	[-40.0 , 94.8]	[-40.0 , 100.0]
Energy	$\theta_{2,2}$	[-20.0 , 50.0]	[-20.0 , 50.0]	[-20.0 , 50.0]	[-20.0 , 50.0]
Brands	$\theta_{2,3}$		[-10.0 , 10.0]		[-10.0 , 10.0]
	$\theta_2(\mu)$	[-55.0 , 50.0]	[-60.6 , 52.8]	[-55.0 , 86.7]	[-60.6 , 87.0]
Comp. time		29.8	30.9	24.3	24.5

Table 4: 95%-confidence intervals for each parameter $\theta_{s,k}$ for $k = 1, 2, 3$ and $s = 1, 2$, and the average entry costs $\theta_s(\mu) \equiv \theta_{s,1} + \theta_{s,2}\mu_s + \theta_{s,3}\mu_s^2$. Confidence intervals are computed by solving optimization problems defined in (50) and (51) using *fmincon* and the least favorable critical value $\hat{c}^{\mathrm{lf}}_{n,k}(1-\alpha,\theta)$ defined in (41) but with α replacing β. The parameter space for the parameters are $\theta_{s,1} \in [-40, 100]$, $\theta_{s,2} \in [-20, 50]$, $\mu_1 \in [0, 3]$ and $\mu_2 \in [0, 2]$, where units are in thousands of US dollars and miles. Computational time is presented in seconds.

too narrow. In practice, we found that trying to solve (50) and (51) using either the bootstrap version of $c_n(1-\alpha,\theta)$ or the self-normalized critical value defined in (42) often led to results that were sensitive to starting points in the minimization procedure, which raises concerns on blind implementation of tools like fmincon or Knitro.

When $T_n(\theta)$ is well-behaved, a simple approach is to solve (50) and (51) using the least favorable critical value $\hat{c}^{\mathrm{lf}}_{n,k}(1-\alpha,\theta)$ defined in (41) but with α replacing β, as this approach, while conservative, delivers a fixed critical value that is not a function of θ. Other alternatives exist in this case, like approximating critical values following Kaido et al. (2019). When $T_n(\theta)$ is not well-behaved, the problem is arguably harder and best practices remain an open question. If it is possible to simplify the model to where a grid search can be operationalized, then solving the problems in (50) and (51) can be bypassed. Alternatively, more recent developments like those in Chen et al. (2018) could be attractive in these situations.

We report estimates obtained by solving the optimization problems defined in equations (50) and (51) using fmincon in Table 4. Overall, confidence regions are large and each marginal confidence interval cannot rule out the hypothesis $\theta_{s,k} = 0$, for each $k = 1, 2, 3$, and $s = 1, 2$. This is likely due to distance affecting marginal costs rather than sunk costs. The table also reports confidence regions for $\theta_s(\mu) \equiv \theta_{s,1} + \theta_{s,2}\mu_s + \theta_{s,3}\mu_s^2$, where μ_s is the average distance to production facilities for company s. We note that it is common in applied work to simply replace μ_s with an estimate $\hat{\mu}_s$ and project the confidence region for $\theta_s = (\theta_{s,1}, \theta_{s,2}, \theta_{s,3})$ into $\theta_{s,1} + \theta_{s,2}\hat{\mu}_s + \theta_{s,3}\hat{\mu}_s^2$, thus not accounting for the additional randomness introduced by estimating μ_s. In this setting, however, accounting for the additional randomness is straightforward as we can simply add two additional moments as a function of $X^d_{s,i,j}$, the distance from market j to firm s's production facilities for product i. The augmented model then includes the additional parameters μ_s for $s \in \mathcal{S}$ and two additional moment inequalities for each of those parameters, i.e., $E[X^d_{s,i,j}] \leq \mu_s$ and $-E[X^d_{s,i,j}] \leq \mu_s$. Confidence regions for $\theta_s(\mu)$ reported in Table 4 are obtained from this augmented model. We find that taking the estimation of μ_s into account widens confidence regions by around $\$3,000 - \$6,800$ for Coca Cola and by about $\$1,900 - \$5,100$ for Energy Brands relative to a simple plug-in approach.

9 Empirical Application: Counter-factuals

Applied researchers are often interested in moving beyond inference on θ in order to study some type of counter-factual analysis, such as simulating equilibrium impacts in alternative settings. While moving to counter-factual analysis introduces certain new challenges, the specifics of their impact may be tightly connected to the details of the particular empirical application under consideration. For example, one could be interested in simulating pricing and product variety decisions for a firm after a potential merger with another firm. This specific goal introduces two challenges. First, inference on the sunk cost parameter θ is obtained by bounding $V_{i,j}$ with $\bar{V}$, but in order to simulate counter-factuals one needs to deal with the fact that $V_{i,j}$ is the object that determines product offering decisions. Second, it is often computationally very costly (or infeasible) to solve for counter-factual product offerings for every value of θ in the confidence region C_n. In Appendix B.4 we discuss ways to address these challenges via a simple algorithm; see Algorithm B.1.

The current practice in applied work often computes counter-factuals evaluated at a single point in the confidence set for θ (say, a middle point) and a particular value of the unobserved error term $V_{i,j}$. The procedure described in Algorithm B.1, on the other hand, checks whether a product assortment belongs to a class of strategies that contains the set of Nash Equilibrium outcomes for any parameter in the confidence region C_n and any value of the unobserved error term $V_{i,j}$. It does so by exploiting the bounding assumption on $V_{i,j}$, and crucially it does not add additional computational steps relative to current practice. In particular, Algorithm B.1 highlights that the computationally difficult part of computing counter-factuals in partially identified models is not dealing with set identification but rather solving for all Nash Equilibria. The main downside is that the algorithm is conservative, and may produce a set of counterfactual outcomes that is larger than those supported under Nash Equilibria.

10 Concluding Remarks

This paper presents a guide for inference in moment inequality models intended to help applied researchers navigate all the decisions required to frame a model as a moment inequality model and then to construct confidence intervals for the parameters of interest. Our main goal is not to provide a comprehensive chronological road-map of all the methods in the literature up to this date, but rather to provide a template that hopefully lowers the entry cost to the literature, both to newcomers and researchers with some exposure to the basic tools. The structure of our guide is divided into “how” and “why” sections, with the why sections discussing the considerations that led to our recommendations as well as other alternatives currently available in the literature. A reader can then choose to focus on the “how”, and learn an established approach to inference in moment inequality models without digging into the overwhelming number of alternatives available at each stage.

A companion Github repository[4] contains all the codes required to replicate the results of this paper in `Matlab`, `Python`, and `R` using simulated data. These codes hopefully also facilitate the way for applied researchers to develop their own code for similar empirical settings. All in all, we expect the combination of the guiding template with the computer codes to provide an easy to digest introduction to inference in moment inequality models that fosters the adoption of such models in empirical research.

References

Andrews, D. W. and Kwon, S. (2019). Misspecified moment inequality models: Inference and diagnostics. *Cowles Foundation Discussion Paper No 2184R2.*

Andrews, D. W. K. and Barwick, P. J. (2012). Inference for parameters defined by moment inequalities: A recommended moment selection procedure. *Econometrica*, **80** 2805–2826.

Andrews, D. W. K. and Guggenberger, P. (2009). Validity of subsampling and “plug-in asymptotic” inference for parameters defined by moment inequalities. *Econometric Theory*, **25** 669–709.

Andrews, D. W. K. and Shi, X. (2013). Inference based on conditional moment inequalities. *Econometrica*, **81** 609–666.

[4] Available at https://github.com/iacanay/guide-inequalities.

Andrews, D. W. K. and Soares, G. (2010). Inference for parameters defined by moment inequalities using generalized moment selection. *Econometrica*, **78** 119–158.

Andrews, I., Roth, J. and Pakes, A. (2019). Inference for linear conditional moment inequalities. Tech. rep., National Bureau of Economic Research.

Armstrong, T. B. (2014a). A note on minimax testing and confidence intervals in moment inequality models. Manuscript. Yale University.

Armstrong, T. B. (2014b). Weighted KS statistics for inference on conditional moment inequalities. *Journal of Econometrics*, **181** 92–116.

Armstrong, T. B. (2015). Asymptotically exact inference in conditional moment inequality models. *Journal of Econometrics*, **186** 51–65.

Armstrong, T. B. and Chan, H. P. (2014). Multiscale adaptive inference on conditional moment inequalities. Manuscript. Yale University.

Atalay, E., Sorensen, A., Sullivan, C. and Zhu, W. (2020). Post-merger product repositioning: An empirical analysis. Mimeo, University of Wisconsin - Madison.

Bai, Y., Santos, A. and Shaikh, A. (2019). A practical method for testing many moment inequalities. *University of Chicago, Becker Friedman Institute for Economics Working Paper.*

Barseghyan, L., Coughlin, M., Molinari, F. and Teitelbaum, J. C. (2021). Heterogeneous choice sets and preferences. *Econometrica*, **89** 2015–2048.

Belloni, A., Bugni, F. and Chernozhukov, V. (2018). Subvector inference in partially identified models with many moment inequalities. arXiv:1806.11466.

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

Berry, S., Levinsohn, J. and Pakes, A. (1995). Automobile prices in market equilibrium. *Econometrica*, **63** 841–890.

Berry, S. T. (1994). Estimating discrete-choice models of product differentiation. *The RAND Journal of Economics*, **25** 242–262.

Berry, S. T. and Haile, P. A. (2021). Foundations of demand estimation. In *Handbook of Industrial Organization, Volume 4* (K. Ho, A. Hortaçsu and A. Lizzeri, eds.), vol. 4 of *Handbook of Industrial Organization*. Elsevier, 1–62.

Bhattacharya, J., Shaikh, A. M. and Vytlacil, E. (2008). Treatment effect bounds under monotonicity assumptions: An application to swan-ganz catheterization. *The American Economic Review* 351–356.

Bhattacharya, J., Shaikh, A. M. and Vytlacil, E. (2012). Treatment effect bounds: An application to swan–ganz catheterization. *Journal of Econometrics*, **168** 223–243.

Blundell, R., Gosling, A., Ichimura, H. and Meghir, C. (2007). Changes in the distribution of male and female wages accounting for employment composition using bounds. *Econometrica*, **75** 323–363.

Brand, J. (2021). Differences in differentiation: Rising variety and markups in retail food stores. Mimeo, University of Texas - Austin.

Bugni, F. A. (2010). Bootstrap inference in partially identified models defined by moment inequalities: Coverage of the identified set. *Econometrica*, **78** 735–753.

Bugni, F. A. (2014). Comparison of inferential methods in partially identified models in terms of error in coverage probability. *Econometric Theory* 1–56.

Bugni, F. A., Canay, I. A. and Shi, X. (2017). Inference for subvectors and other functions of partially identified parameters in moment inequality models. *Quantitative Economics*, **8** 1–38.

Canay, I. A. (2010). El inference for partially identified models: Large deviations optimality and bootstrap validity. *Journal of Econometrics*, **156** 408–425.

Canay, I. A. and Shaikh, A. M. (2017). Practical and theoretical advances for inference in partially identified models. In *Advances in Economics and Econometrics: Volume 2: Eleventh World Congress* (B. Honoré, A. Pakes, M. Piazzesi and L. Samuelson, eds.), vol. 2. Cambridge University Press, 271–306.

Cardell, N. S. (1997). Variance components structures for the extreme-value and logistic distributions with application to models of heterogeneity. *Econometric Theory*, **13** 185–213.

Cattaneo, M. D., Ma, X., Masatlioglu, Y. and Suleymanov, E. (2020). A random attention model. *Journal of Political Economy*, **128** 2796–2836.

Chen, X., Christensen, T. M. and Tamer, E. (2018). Monte carlo confidence sets for identified sets. *Econometrica*, **86** 1965–2018.

Chernozhukov, V., Chetverikov, D. and Kato, K. (2019). Inference on causal and structural parameters using many moment inequalities. *The Review of Economic Studies*, **86** 1867–1900.

Chernozhukov, V., Hong, H. and Tamer, E. (2007). Estimation and confidence regions for parameter sets in econometric models. *Econometrica*, **75** 1243–1284.

Chernozhukov, V., Lee, S. and Rosen, A. M. (2013). Intersection bounds: Estimation and inference. *Econometrica*, **81** 667–737.

Chetverikov, D. (2013). Adaptive test of conditional moment inequalities. *arXiv:1201.0167*.

Cho, J. and Russell, T. M. (2018). Simple inference on functionals of set-identified parameters defined by linear moments. *arXiv preprint arXiv:1810.03180*.

Ciliberto, F., Murry, C. and Tamer, E. (2021). Market structure and competition in airline markets. *Journal of Political Economy*, **129** 2995–3038.

Ciliberto, F. and Tamer, E. (2009). Market structure and multiple equilibria in airline markets. *Econometrica*, **77** 1791–1828.

Cox, G. and Shi, X. (2019). Simple adaptive size-exact testing for full-vector and subvector inference in moment inequality models. *arXiv preprint arXiv:1907.06317*.

Crawford, G. S. and Yurukoglu, A. (2012). The welfare effects of bundling in multichannel television markets. *American Economic Review*, **102** 643–85.

Dickstein, M. J. and Morales, E. (2018). What do Exporters Know?*. *The Quarterly Journal of Economics*, **133** 1753–1801.

Döpper, H., MacKay, A., Miller, N. H. and Stiebale, J. (2022). Rising markups and the role of consumer preferences. Harvard Business School Strategy Unit Working Paper No. 22-025.

Eizenberg, A. (2014). Upstream Innovation and Product Variety in the U.S. Home PC Market. *The Review of Economic Studies*, **81** 1003–1045.

Fan, Y. and Yang, C. (2022). Estimating discrete games with many firms and many decisions: An application to merger and product variety. Working Paper 30146, National Bureau of Economic Research.

Gafarov, B. (2019). Inference in high-dimensional set-identified affine models. *arXiv preprint arXiv:1904.00111*.

Haile, P. A. and Tamer, E. (2003). Inference with an incomplete model of english auctions. *Journal of Political Economy*, **111** 1–51.

Hansen, L. P., Heaton, J. and Luttmer, E. G. J. (1995). Econometric evaluation of asset pricing models. *Review of Financial Studies*, **8** 237–274.

Hansen, L. P. and Jagannathan, R. (1991). Implications of security market data for models of dynamic economies. *Journal of Political Economy* 225–262.

HECKMAN, J. J. and VYTLACIL, E. J. (2001). *Instrumental variables, selection models, and tight bounds on the average treatment effect.* Springer.

HENRY, M. and ONATSKI, A. (2012). Set coverage and robust policy. *Economics Letters*, **115** 256–257.

HO, K. (2009). Insurer-provider networks in the medical care market. *American Economic Review*, **99** 393–430.

HO, K. and PAKES, A. (2014). Hospital choices, hospital prices and financial incentives to physicians. *American Economic Review*, **104** 3841–84.

HO, K. and ROSEN, A. M. (2017). Partial identification in applied research: Benefits and challenges. In *Advances in Economics and Econometrics: Volume 2: Eleventh World Congress*, vol. 2. Cambridge University Press, 307.

HOLMES, T. J. (2011). The diffusion of wal-mart and economies of density. *Econometrica*, **79** 253–302.

HOROWITZ, J. L. and MANSKI, C. F. (1995). Identification and robustness with contaminated and corrupted data. *Econometrica*, **63** 281–302.

HOROWITZ, J. L. and MANSKI, C. F. (1998). Censoring of outcomes and regressors due to survey nonresponse: Identification and estimation using weights and imputations. *Journal of Econometrics*, **84** 37–58.

HOUDE, J.-F., NEWBERRY, P. and SEIM, K. (2023). Nexus tax laws and economies of density in e-commerce: A study of amazon's fulfillment center network. *Econometrica (Forthcoming).*

ILLANES, G. (2016). Switching costs in pension plan choice. Working paper, Northwestern University.

IMBENS, G. and MANSKI, C. F. (2004). Confidence intervals for partially identified parameters. *Econometrica*, **72** 1845–1857.

JONES, D. R. (2001). A taxonomy of global optimization methods based on response surfaces. *Journal of global optimization*, **21** 345–383.

KAIDO, H., MOLINARI, F. and STOYE, J. (2019). Confidence intervals for projections of partially identified parameters. *Econometrica*, **87** 1397–1432.

KLEPPER, S. and LEAMER, E. E. (1984). Consistent sets of estimates for regressions with errors in all variables. *Econometrica: Journal of the Econometric Society* 163–183.

Kline, B., Pakes, A. and Tamer, E. (2020). Moment inequalities and partial identification in industrial. Tech. rep., Harvard University.

Kline, B. and Tamer, E. (2022). Recent developments in partial identification. Working paper, Harvard University.

Kline, P., Santos, A. et al. (2013). Sensitivity to missing data assumptions: Theory and an evaluation of the us wage structure. *Quantitative Economics*, **4** 231–267.

Kline, P. and Tartari, M. (2015). Bounding the labor supply responses to a randomized welfare experiment: A revealed preference approach. Tech. rep., National Bureau of Economic Research.

Maini, L. and Pammolli, F. (2023). Reference pricing as a deterrent to entry: Evidence from the european pharmaceutical market. *American Economic Journal: Microeconomics (Forthcoming).*

Manski, C. F. (1989). Anatomy of the selection problem. *Journal of Human Resources*, **24** 343–360.

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

Manski, C. F. (1994). The selection problem. In *Advances in Econometrics, Sixth World Congress*, vol. 1. 143–70.

Manski, C. F. (1995). *Identification problems in the social sciences*. Harvard University Press.

Manski, C. F. (1997). Monotone treatment response. *Econometrica: Journal of the Econometric Society* 1311–1334.

Manski, C. F. (2003). *Partial Identification of Probability Distributions*. Springer-Verlag, New York.

Manski, C. F. (2007). *Identification for Prediction and Decision*. Harvard University Press.

Manski, C. F. (2013). *Public Policy in an Uncertain World: Analysis and Decisions*. Harvard University Press.

Manski, C. F. and Pepper, J. V. (2000). Monotone instrumental variables: with an application to the returns to schooling. *Econometrica*, **68** 997–1010.

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

Miller, N. H. and Weinberg, M. C. (2017). Understanding the price effects of the millercoors joint venture. *Econometrica*, **85** 1763–1791.

Miravete, E. J., Seim, K. and Thurk, J. (2018). Market power and the laffer curve. *Econometrica*, **86** 1651–1687.

Molinari, F. (2020). Microeconometrics with partial identification. *Handbook of econometrics*, **7** 355–486.

Morales, E., Sheu, G. and Zahler, A. (2019). Extended Gravity. *The Review of Economic Studies*, **86** 2668–2712.

Nevo, A. (2001). Measuring market power in the ready-to-eat cereal industry. *Econometrica*, **69** 307–342.

Nevo, A. (2003). New products, quality changes, and welfare measures computed from estimated demand systems. *The Review of Economics and Statistics*, **85** 266–275.

Nosko, C. (2010). Competition and quality choice in the cpu market. Working Paper 30146, Booth School of Business.

Pakes, A. (2010). Alternative models for moment inequalities. *Econometrica*, **78** 1783–1822.

Pakes, A., Porter, J., Ho, K. and Ishii, J. (2015). Moment inequalities and their application. *Econometrica*, **83** 315–334.

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

Romano, J. P. and Shaikh, A. M. (2010). Inference for the identified set in partially identified econometric models. *Econometrica*, **78** 169–212.

Romano, J. P., Shaikh, A. M. and Wolf, M. (2014). A practical two-step method for testing moment inequalities. *Econometrica*, **82** 1979–2002.

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

Schennach, S. M. (2014). Entropic latent variable integration via simulation. *Econometrica*, **82** 345–385.

Shaikh, A. M. and Vytlacil, E. J. (2011). Partial identification in triangular systems of equations with binary dependent variables. *Econometrica* 949–955.

Stoye, J. (2009). More on confidence intervals for partially identified parameters. *Econometrica*, **77** 1299–1315.

Syrgkanis, V., Tamer, E. and Ziani, J. (2017). Inference on auctions with weak assumptions on information. *arXiv preprint arXiv:1710.03830.*

Tamer, E. (2003). Incomplete simultaneous discrete response model with multiple equilibria. *The Review of Economic Studies*, **70** 147–165.

Tamer, E. (2010). Partial identification in econometrics. *Annual Review of Economics*, **2** 167–195.

Wollmann, T. G. (2018). Trucks without bailouts: Equilibrium product characteristics for commercial vehicles. *American Economic Review*, **108** 1364–1406.

Online Supplemental Appendix

A Additional Tables and Figures

Table A.1 shows summary statistics for the main products in our estimation sample - average prices, average / 25th percentile / median / 75th percentile shares, and the share of markets (month-DMA) in which each product is offered. Panel A presents these statistics for the 10 most popular products in the demand estimation sample. Panel B presents these statistics for the next 5 most popular Coca Cola products, while Panel C does the same for Energy Brands products. Note that some Coca Cola products have perfect coverage during this period - naturally, this will affect estimation of any bounds where the deviation requires adding a product. We discuss this issue in detail in Section 8.1.

		Avg	Shares (in %)				Share of
Firms	Products	price	Avg	p25	p50	p75	markets
Panel A: Top 10 Products							
PepsiCo	Aquafina	5.6	2.28	1.21	1.88	2.85	99.78
Coca-Cola	Dasani	5.9	1.44	0.68	1.12	1.78	99.08
Nestle SA	Arrowhead	4.9	2.41	1.44	2.26	3.25	22.55
Private 1	Independent	4.2	2.59	0.37	1.39	3.99	66.83
Nestle SA	Deer Park	5.5	1.60	0.06	0.94	2.39	33.77
Nestle SA	Poland Spring	5.9	0.58	0.01	0.04	0.19	45.01
PepsiCo	Aquafina 20 Oz	1.3	0.19	0.11	0.16	0.23	100.00
Coca-Cola	Dasani 20 Oz	1.3	0.21	0.11	0.16	0.24	100.00
Private 2	Independent	4.8	1.21	0.06	0.38	1.67	75.91
PepsiCo	Gatorade 32 Oz	1.2	0.14	0.07	0.11	0.18	99.97
Panel B: Next 5 Coca Cola Products							
Coca-Cola	Dasani 12 Oz	4.1	0.15	0.08	0.13	0.19	99.49
	Simply Lemonade	2.7	0.05	0.03	0.04	0.07	85.34
	Dasani 33.8 Oz	1.3	0.08	0.03	0.05	0.09	100.00
	Minute Maid	3.2	0.07	0.03	0.06	0.09	85.64
	Powerade Ion4	1.0	0.14	0.06	0.10	0.17	64.82
Panel C: Next 5 Energy Brands Products							
Energy Brands (VitaminWater)	Dragonfruit	1.3	0.04	0.01	0.03	0.05	97.05
	Fruit Punch	1.3	0.03	0.01	0.02	0.04	96.37
	Tropical Citrus	1.3	0.03	0.01	0.02	0.04	95.85
	Power C	1.3	0.02	0.01	0.02	0.03	93.22
	Essential	1.3	0.02	0.01	0.02	0.03	92.57

Table A.1: Summary statistics for the 10 best-selling UPCs in our sample and for the next 5 top-selling Coca Cola and Energy Brands products.

B Additional Details on Implementation

B.1 Demand Estimation

We estimate demand on pre-merger data using a nested logit model. This is an admittedly simple demand estimation approach but it helps us devote attention to the issues that are more closely related to inference with moment inequalities. Improving the demand estimation step would not affect the construction of confidence regions for the parameters determining sunk costs, though it would certainly affect the actual numbers. Since our goal here is mostly pedagogical, we keep this initial step simple.

Throughout the appendix we define a market i as a combination of DMA a and time t, so that $i = (a, t)$ and interchangeably index variables by i or by (a, t) depending on the context. That is, for any random variable X, $X_{i,j}$ and $X_{a,t,j}$ denote the same random variable.

To estimate the nested logit model, we divide all products into 3 mutually exclusive groups: branded water, private label water, non-water, and the outside option. We assume a consumer c that lives in DMA a in period t and purchases product j belonging to group g receives indirect utility $U_{c,a,t,j}$ according to

$$U_{c,a,t,j} = \gamma_0 + \gamma_1 p_{a,t,j} + \xi_j + \xi_a + \xi_{j,a} + \xi_{a,t,j} + \zeta_{c,g} + (1-\rho)\epsilon_{c,a,t,j} \ , \tag{A.1}$$

where $p_{a,t,j}$ is price, ξ_j, ξ_a, $\xi_{j,a}$ and $\xi_{a,t,j}$ are product attributes that are unobserved to the econometrician, $\epsilon_{c,a,t,j}$ is identically and independently distributed Gumbel error term, $\zeta_{c,g}$ is a group-level error term, and ρ is the nesting parameter. The distribution of $\zeta_{c,g}$ is the unique distribution such that $\zeta_{c,g}+(1-\rho)\cdot\epsilon_{c,a,t,j}$ is also distributed Gumbel (Cardell, 1997). Estimating this model requires calculating market shares for each product and for the outside good. We do so by assuming that market size in DMA a, M_a, is equal to 1.5 times the maximum liters we observe being sold in a over time. With this assumption, one can compute the market share $\mathrm{s}_{a,t,j}$ of good j in market a in period t as the ratio between the number of liters of good j sold in a during t and M_a. The share of the outside good is one minus the sum of shares for each good j being offered in a during t. Having obtained shares, we invert the market share function (Berry, 1994) and estimate

$$\ln(\mathrm{s}_{a,t,j}) - \ln(\mathrm{s}_{a,t,0}) = \gamma_0 + \gamma_1 p_{a,t,j} + \rho \ln(\bar{s}^g_{a,t,j}) + \xi_j + \xi_a + \xi_{j,a} + \xi_{a,t,j} \ , \tag{A.2}$$

where $\mathrm{s}_{a,t,j}$ and $\mathrm{s}_{a,t,0}$ are product j's share and the outside good's share in DMA a and period t, respectively, and $\bar{s}^g_{a,t,j}$ is product j's share within group g in DMA a and period t. We estimate the model in (A.2) by ordinary least squares (OLS) using product, DMA and product-DMA fixed effects. The main identification assumption is that unobserved product-DMA-time product attributes, captured by $\xi_{a,t,j}$, are orthogonal to prices after controlling for product-DMA fixed effects. The most important deviation relative to the frontier of the demand estimation literature is that we do not use random coefficients. This implies that within nest substitution patterns suffer from the issues highlighted in Berry et al. (1995). Again, since our goal is to illustrate how to conduct inference using moment inequalities in this setting, instead of how to best estimate demand for bottled water and juice drinks, we are comfortable making these assumptions. The results are in Table B.1. We find large price elasticities, particularly for Coca Cola products.

The mean price elasticity is 9.2, and the median is 7.2. Given that we are estimating demand for bottled water products at the UPC level, we do not find these magnitudes implausible. We also find that demand for Energy Brands' products is less price-sensitive. This is consistent with our prior, as these products aim to be more differentiated than standard bottled water.

(a) Demand estimation		(b) Price elasticity weighted by shares			
	$\ln(\mathrm{s}_{a,t,j}/\mathrm{s}_{a,t,0})$		All firms	Coca-Cola	Energy Brands
price	-0.331	mean	9.20	7.98	4.34
	[-0.334, -0.327]	q25	4.37	4.13	4.01
$\ln(\bar{\mathrm{s}}^{g}_{a,t,j})$	0.899	q50	7.22	7.74	4.30
	[0.897, 0.900]	q75	12.54	10.83	4.63
Product FE	Y	Num. Prod.	212	Observ.	477,133
DMA FE	Y	Num. DMA	205	R^2	0.945

Table B.1: Demand estimation and price elasticity based on a nested-logit model using monthly pre-merge data.

B.2 Marginal Cost Estimation

Having recovered demand estimates, we now turn to estimating marginal costs. Recall that market i denotes a DMA-period (a,t) combination. Under the assumption that firms compete in prices, each firm's first order conditions in market i are

$$\mathrm{s}_{i,j}(p_i) + \sum_{j' \in \mathcal{J}_s} D_{i,j'}(p_{i,j'} - c_{i,j'}) \frac{\partial \mathrm{s}_{i,j'}(p_i)}{\partial p_{i,j}} = 0 \quad \{\forall j \in \mathcal{J}_s | D_{i,j} = 1\} \ . \tag{A.3}$$

We solve this system of equations and recover estimates $\hat{c}_{i,j}$, or $\hat{c}_{a,t,j}$ when being explicit about $i=(a,t)$. Further, we assume that marginal costs of each product j are constant in output and that marginal cost realizations satisfy

$$\ln(\hat{c}_{a,t,j}) = \omega_j + \omega_a + \omega_{j,a} + \omega_{a,t,j} \ , \tag{A.4}$$

where ω_j, ω_a, ω_{jd} and ω_{jdt} are product attributes that are unobservable to the econometrician. We estimate this model by OLS using product, DMA and product-DMA fixed effects.

B.3 Calculation of Estimated Variable Profit Differentials

This subsection discusses how to move from demand and marginal cost estimates to estimates of variable profit differentials. First, note that we need estimated variable profit differentials for two firms, Coca-Cola and Energy Brands, as we are only interested in computing sunk costs for them. These companies offer $J = 31$ products in our data, so in this section, and we some abuse of notation, we use $\mathcal{J}$ to denote products that are *only* offered by either Coca-Cola or

Energy Brands, without including products offered by other firms. Consistent with this change, each market i has an observed assortment vector $D_i \in \{0,1\}^J$ where $J = 31$ for the purposes of estimating variable profit differentials.

Consider a given product portfolio $\tilde{D}_i \in \{0,1\}^J$, which could be an observed product portfolio or a counter-factual one. The algorithm to compute variable profit differential given this particular product assortment is as follows:

1. For each product j that is offered in market $i = (a,t)$ given the product assortment, i.e. $\tilde{D}_{i,j} = 1$, we draw $B = 200$ values from the empirical distributions of $\hat{\xi}_{j,a} + \hat{\xi}_{a,t,j}$ and $\hat{\omega}_{j,a} + \hat{\omega}_{a,t,j}$ with replacement. These objects are the residuals from equations (A.2) and (A.4), respectively. We denote these values by $\{\hat{\xi}^{(b)}_{j,a} + \hat{\xi}^{(b)}_{a,t,j} : 1 \leq b \leq B\}$ and $\{\hat{\omega}^{(b)}_{j,a} + \hat{\omega}^{(b)}_{a,t,j} : 1 \leq b \leq B\}$.

2. We then add to each draw, $\hat{\xi}^{(b)}_{j,a} + \hat{\xi}^{(b)}_{a,t,j}$, the estimated values of ξ_j and ξ_a, and to every draw, $\hat{\omega}^{(b)}_{j,a} + \hat{\omega}^{(b)}_{a,t,j}$, the estimated values of ω_j and ω_a. For each b and for each j such that $\tilde{D}_{i,j} = 1$, this results in
$$\hat{\xi}^{(b)}_{i,j} = \hat{\xi}^{(b)}_{j,a} + \hat{\xi}^{(b)}_{a,t,j} + \hat{\xi}_j + \hat{\xi}_a \tag{A.5}$$
$$\hat{\omega}^{(b)}_{i,j} = \hat{\omega}^{(b)}_{j,a} + \hat{\omega}^{(b)}_{a,t,j} + \hat{\omega}_j + \hat{\omega}_a \ . \tag{A.6}$$

3. For each b and for each j such that $\tilde{D}_{i,j} = 1$, we compute marginal costs, optimal prices, and market shares. First, we use equation (A.4) to solve for marginal costs using $\hat{\omega}^{(b)}_{i,j}$, i.e. $\hat{c}^{(b)}_{i,j} = \exp(\hat{\omega}^{(b)}_{i,j})$. Second, we solve for optimal prices by solving the non-linear system of equations in prices defined by equation (A.3), and denote them by $\hat{p}^{(b)}_{i,j}$. Finally, we compute nested logit market shares as,
$$\hat{s}^{(b)}_{a,t,j} = \frac{\exp\left(\frac{\hat{\delta}^{(b)}_{a,t,j}}{1-\hat{\rho}}\right)}{(\hat{\mathcal{D}}^{(b)}_g)^{\hat{\rho}}\left(\sum_{g'}(\hat{\mathcal{D}}^{(b)}_{g'})^{1-\hat{\rho}}\right)} \ , \tag{A.7}$$
where $\hat{\mathcal{D}}^{(b)}_g = \sum_{j \in \tilde{\mathcal{J}}_g} \exp\left(\frac{\hat{\delta}^{(b)}_{a,t,j}}{1-\hat{\rho}}\right)$, $\hat{\delta}^{(b)}_{a,t,j} = \hat{\gamma}_0 + \hat{\gamma}_1 p_{a,t,j} + \hat{\xi}_j + \hat{\xi}_a + \hat{\xi}^{(b)}_{j,a} + \hat{\xi}^{(b)}_{a,t,j}$, and $\tilde{\mathcal{J}}_g$ denotes the set of products belonging to group g.

4. For each b and each firm s, we compute variable profit according to (52) in Assumption 7.1, and then compute $\hat{r}_{s,i}(O_i)$ by averaging across the B draws i.e.,
$$\hat{r}_{s,i}(O_i) = \frac{1}{B}\sum_{b=1}^{B}\sum_{j \in \mathcal{J}_s} M_i \tilde{D}_{i,j}(\hat{p}^{(b)}_{i,j} - \hat{c}^{(b)}_{i,j})\hat{\mathrm{s}}^{(b)}_{i,j} \ .$$

5. Variable profit differentials are the difference in these averages between counter-factual product offerings and the observed portfolio, as defined in Section 3, where for each of these alternative assortments we repeat steps 1-4 above.

B.4 Details on Counter-factuals

Following the discussion in Section 9, assume that the counter-factual of interest is predicting pricing and product variety decisions for Coca Cola post-merger. In this counter-factual, Coca Cola (which now includes Energy Brands' products) and all of its competitors will be allowed to adjust pricing and product variety. To perform this analysis, one would need sunk cost estimates θ_s for Coca Cola and *all* of its competitors. This is important since, as we discussed in Section 8, by virtue of the model being separable in θ, it is possible and convenient to conduct inference on θ_s for $s \in \{\text{CocaCola}, \text{EnergyBrands}\}$ without accounting for other competitors. The methodology, however, can easily be expanded to recover sunk cost estimates for other firms. We proceed under the assumption that we have computed confidence regions for sunk costs for *all* firms in the market and so $\mathcal{J}$ and $\mathcal{S}$ now include information on all the firms we used to estimate demand. The algorithm we would propose is the following:

Algorithm B.1. Let $\mathcal{D}_i \equiv \{0,1\}^J$ denote the set of all possible product assortment profiles in market i and consider the following algorithm to compute counter-factuals.

Step 1 Compute expected equilibrium variable profits for each firm $s \in \mathcal{S}$ and for each $D_i \in \mathcal{D}_i$ and denote them by $\hat{r}_{s,i}(D_i)$. In Section 7.2 this is done by solving for equilibrium prices following Appendix B.3. Store the values of objects of interest of the counter-factual - prices, consumer surplus, etc.

Step 2 Let $\underline{\theta}_s$ denote the lowest value of θ in firm s confidence region. For each firm $s \in \mathcal{S}$ and each product portfolio $D_i \in \mathcal{D}_i$, consider the inequality

$$\hat{r}_{s,i}(D_i) - \sum_{j \in \mathcal{J}_s} D_{i,j}(\underline{\theta}_s - \bar{V}) \geq 0 \ . \tag{A.8}$$

Denote by $\mathcal{D}_i^2 \subseteq \mathcal{D}_i$ the set of product offerings D_i for which (A.8) holds for all $s \in \mathcal{S}$.

Step 3 For each firm $s \in \mathcal{S}$ and $D_i \in \mathcal{D}_i^2$, let $\mathcal{D}_s(D_i)$ be the subset of $\mathcal{D}_i^2$ that keeps the assortment decisions of s's competitors fixed relative to D_i, i.e., $\mathcal{D}_s(D_i) \equiv \Lambda_s(D_i) \cap \mathcal{D}_i^2$, where $\Lambda_s(D_i)$ is defined in (13). Then, for each $D_i' \in \mathcal{D}_s(D_i)$ check the condition

$$\hat{r}_{s,i}(D_i) - \hat{r}_{s,i}(D_i') - J_s^{\text{in}}(\underline{\theta}_s - \bar{V}) + J_s^{\text{out}}(\bar{\theta}_s + \bar{V}) \geq 0 \ , \tag{A.9}$$

where J_s^{in} is the number of products offered by s in D_i but not in D_i', and J_s^{out} is the number of products offered by s in D_i' but not in D_i. Let $\mathcal{D}_i^3 \subseteq \mathcal{D}_i^2$ denote the set of product portfolios D_i for which (A.9) holds for all $s \in \mathcal{S}$ and all $D_i' \in \mathcal{D}_s(D_i)$.

Step 4 Construct bounds on the objects of interest (average prices, consumer surplus, etc.) by finding the maximum and minimum values of those objects across all portfolios D_i in $\mathcal{D}_i^3$.

Step 1 is the most computationally demanding step, as it requires computing Nash Equilibrium prices for all possible product offerings. We see this step as the main limitation when performing counter-factual analysis in this class of discrete games. Researchers often will need to restrict strategy space in order to make this feasible. For example, instead of modelling the action space of the product variety game as the decision to offer each of the $J = 212$ products in our market separately, which would require solving for $2^{212} - 1$ Nash Equilibria, one can assume that subsets of products are always offered together, or that they will always be offered. Beyond this, parallelization of step 1 is useful, and a benefit of this algorithm is that step 1 only needs to be done once.

Step 2 checks whether a set of product offerings delivers positive profits to all firms in the most advantageous case possible - when sunk costs are the lowest possible value. Since $e_{i,j} \geq \max[\underline{\theta}_s - \overline{V}, 0]$, D_i cannot be an equilibrium assortment profile if the condition in step 2 does not hold for every firm. This check is straightforward once step 1 is computed and it potentially removes irrelevant assortments to be considered in the next step, which is computationally more intense.

Step 3 checks whether there are unilateral incentives to deviate to any other assortments in the case where deviation incentives are the weakest - when sunk costs for offered products are at the lowest possible value and sunk costs for non offered products are at their highest. It is straightforward to show if D_i is a Nash Equilibrium, it must belong to $\mathcal{D}_i^3$. The converse is not true, however, as in equation (A.9) θ_s differs for offered and un-offered products, and our behavioral model posits a fixed θ_s across products produced by firm s. Therefore, $\mathcal{D}_i^3$ contains all Nash Equilibria, and can be thought of as a conservative approximation to the strategies that are a Nash Equilibrium in the counter-factual.

Claim B.1. Suppose $e_{i,j} \in \left[\underline{\theta}_s - \bar{V}, \bar{\theta}_s + \bar{V}\right]$. If D_i is a Nash Equilibrium, then $D_i \in \mathcal{D}_i^3$.

Proof of Claim B.1: If D_i is a Nash Equilibrium, then any firm s must have positive profits

$$\hat{r}_{s,i}\left(D_i\right) - \sum_{j \in \mathcal{J}_s} D_{i,j} e_{i,j} \geq 0 \ , \tag{A.10}$$

and any firm s has no unilateral incentives to deviate

$$\hat{r}_{s,i}\left(D_i\right) - \hat{r}_{s,i}\left(D_i'\right) - \sum_{j \in \mathcal{J}_s} \left(D_{i,j} - D_{i,j}'\right) e_{i,j} \geq 0 \tag{A.11}$$

for any $D_i' \in \Lambda_s(D_i)$. Since $e_{i,j} \in [\underline{\theta}_s - \bar{V}, \bar{\theta}_s + \bar{V}]$, we have that

$$\hat{r}_{s,i}\left(D_i\right) - \sum_{j \in \mathcal{J}_s} D_{i,j} e_{i,j} \leq \hat{r}_{s,i}\left(D_i\right) - \sum_{j \in \mathcal{J}_s} D_{i,j} \left(\underline{\theta}_s - \bar{V}\right) \ ,$$

which implies $D_i \in \mathcal{D}_i^2$ using (A.10). Similarly, we obtain

$$\hat{r}_{s,i}\left(D_i\right) - \hat{r}_{s,i}\left(D_i'\right) - \sum_{j \in \mathcal{J}_s} \left(D_{i,j} - D_{i,j}'\right) e_{i,j} \leq \hat{r}_{s,i}\left(D_i\right) - \hat{r}_{s,i}\left(D_i'\right) - J_s^{\text{in}}(\underline{\theta}_s - \bar{V}) + J_s^{\text{out}}(\bar{\theta}_s + \bar{V}) \ ,$$

which implies $D_i \in \mathcal{D}_i^3$ using (A.11). ■

C Misspecification Robust CS

Andrews and Kwon (2019) study inference in moment inequality models like those in (1) in settings where the model is misspecified and so $\Theta_0(P)$ in (2) is empty. A potential consequence of misspecification is spurious precision of standard confidence sets for θ, meaning that the coverage probability of the confidence set is less than its nominal level $1-\alpha$ for all parameter values, including any potential pseudo-true value.

Andrews and Kwon (2019) introduce a misspecification index that equals the maximum violation across moment inequalities (normalized by their standard deviations) evaluated at the parameter value that minimizes the maximum violation. In order to describe the method clearly, we introduce some additional notation. Let

$$\sigma_\ell(\theta) = \sqrt{\text{var}[m_\ell(W_i,\theta)]} \tag{A.12}$$

for $1 \leq \ell \leq k$ and define

$$m^*(W_i,\theta) = (m_1^*(W_i,\theta),\ldots,m_k^*(W_i,\theta))' \text{ for } m_\ell^*(W_i,\theta) \equiv \frac{m_\ell(W_i,\theta)}{\sigma_\ell(\theta)} \ . \tag{A.13}$$

$$\Theta_0^*(P) = \{\theta \in \Theta : E_P[m(W_i,\theta)] - r^{\text{inf}} 1_k \leq 0\} \ , \tag{A.14}$$

where 1_k is a k-dimensional vector of ones, and r^{inf} is a scalar given by

$$r^{\text{inf}} = \inf_{\theta\in\Theta} \max_{1\leq\ell\leq k} \max\{E[m_\ell^*(W_i,\theta)],0\} \ . \tag{A.15}$$

Note that r^{inf} equals the maximum violation across moment inequalities (normalized by their standard deviations) evaluated at the parameter value that minimizes the maximum violation. The misspecification-robust identified set is non-empty even under model misspecification.

The SPUR1 test we use in Section 8.2 consists of the following X steps:

1. Compute the sample analog of r^{inf} as follows,

$$\hat{r}_n^{\text{inf}} \equiv \inf_{\theta\in\Theta} \max_{1\leq\ell\leq k} \max\left\{\frac{\bar{m}_{n,\ell}(\theta)}{\hat{\sigma}_{n,\ell}(\theta)}, 0\right\} \ . \tag{A.16}$$

2. Modify the test statistic in (43) to account for $\hat{r}_n^{\text{inf}}$ as follows,

$$T_n^*(\theta) \equiv T\Big(\sqrt{n}(\hat{D}_n^{-1}(\theta)\bar{m}_n(\theta) + \hat{r}_n^{\text{inf}} 1_k), \hat{\Omega}_n(\theta)\Big) \ . \tag{A.17}$$

3. Compute the SPUR1 bootstrap critical, denoted by $\hat{c}_n^{\text{spur}}(1-\alpha,\theta)$, as described in Andrews and Kwon (2019).

4. Reject whenever $T_n^*(\theta) > \hat{c}_n^{\text{spur}}(1-\alpha,\theta)$

Note that the critical value in Step 3, while similar in spirit to the one described in section 5.2, requires substantial modifications to account for the additional randomness introduced by $\hat{r}_n^{\text{inf}}$. We also note that, when the test statistic in (43) equals the max test statistic in (39), $T_n^*(\theta)$ is equivalent to the re-centered version of $T_n(\theta)$ in (39). We refer the reader to Andrews and Kwon (2019) for details or to our companion `Matlab` and `Python` packages.