A Moment Inequality Approach to Estimating Multinomial Choice Models with Unobserved Consideration Sets*

(Job Market Paper)

Zhentong Lu†

Department of Economics

University of Wisconsin-Madison

zhentong.lu@wisc.edu

December 4, 2014

*The latest version can be found here: https://sites.google.com/site/zhentonglu/

†I am indebted to my advisors Amit Gandhi and Xiaoxia Shi for invaluable guidance and support on this project. I would also like to thank Ken Hendricks and Alan Sorensen for insightful discussions and detailed comments. Conversations with Jean-Pierre Dubé, Matt Gentry, Han Hong, Francesca Molinari, Yanqin Fan, Marc Rysman have improved the paper. I am grateful to the participants at the 2014 North American Summer Meeting of the Econometric Society, as well as the seminar participants at Wisconsin-Madison, Wisconsin-Whitewater for their many helpful comments and questions.

Abstract

This paper proposes a moment inequality approach to estimating random utility models when consumers' consideration sets are unobservable to econometricians. I show that, without relying on a specific model of consideration set formation, the random utility model can be identified and estimated via a system of conditional moment inequalities derived from the utility maximization assumption. Monte Carlo experiments demonstrate that the method can correct the bias caused by misspecified consideration set generation.

I apply the moment inequality approach to study whether attention inertia can explain some of the observed persistence in consumers' brand choices, as opposed to alternative explanations in terms of preference, e.g., state-dependent utilities. The estimation results, obtained using household scanner data, show that up to 20% of the observed persistence, in terms of re-purchase probability, can be attributed to the fact that previous purchase of a brand increases its present consideration probability.

Keywords: Discrete Choice; Consideration Set; Moment Inequality; Persistence; Brand Choice

JEL: C14, C50, L00, M30.

1 Introduction

This paper concerns demand estimation for differentiated products when consumers are allowed to have heterogeneous consideration sets that are unobservable to econometricians. Exploiting the consumer utility maximization assumption, I develop a moment inequality approach to estimating random utility models without specifying a complete model for consideration set formation. I apply the proposed approach to household purchase panel data to quantify the importance of attention inertia as a source for explaining the observed persistence in brand choice. The results show that a considerable amount of the persistence is driven by the fact that previous purchase of a brand increases its chance of entering consumers' consideration sets.

Understanding consumers' intrinsic preferences from revealed choices is a basic issue in economics. Originated from Marschak (1960) and subsequently developed by McFadden (1974), Berry, Levinsohn, and Pakes (1995), etc., the random utility model has become the standard econometric tool for eliciting preferences using data on consumer characteristics and purchase decisions. The central premise of the random utility model is utility maximization, i.e., a consumer chooses the product that gives her the highest utility from the consideration set. Both utility function and consideration set are essential in determining a consumer's final choice. The traditional literature on discrete choice model (see Train (2009) for a comprehensive coverage on the topic) focuses on modelling the utility function (preference) and treat a consumer's consideration set as exogenously given; in practice, empirical researchers have to impute a consideration set, which usually include all the products available in a given market, to estimate the model.

But the imputed set may not be the underlying true consideration set of each consumer. As a basic consensus of the large literature in marketing and psychology concerning the behavioral foundation and empirical measurement of consideration sets, see, e.g., Gilbride and Allenby (2006), Gilbride and Allenby (2004), Hauser and Wernerfelt (1990), Hauser, Toubia, Evgeniou, Befurt, and Dzyabura (2010), Jedidi and Kohli (2005), Liu and Arora

(2011), Roberts and Lattin (1991) and Shocker, Ben-Akiva, Boccara, and Nedungadi (1991), consumers' consideration sets are endogenously chosen and typically small comparing to the whole set of products available in the market due to cognitive capacity limitations. Both theoretical analysis, see, e.g., Masatlioglu, Nakajima, and Ozbay (2012); Manzini and Mariotti (2014); Eliaz and Spiegler (2011), and empirical evidence, see, e.g., Goeree (2008), Draganska and Klapper (2010), suggest that ignoring the consideration set heterogeneity and endogeneity has important consequences for understanding the preferences from consumer choice data.

To address this concern, we need to extend the random utility framework in a way that both consideration set formation and preference can be jointly modeled. However, the fundamental challenge of modelling the consideration set generation is that the number of possible consideration sets, which grows exponentially as the number of products in the market, is impractically large in many empirical applications. To circumvent this difficulty, the existing literature imposes detailed parametric structures and identification assumptions on the consideration set formation, e.g., Ben-Akiva and Boccara (1995), Chiang, Chib, and Narasimhan (1999), Goeree (2008), Bruno and Vilcassim (2008), Gentry (2011), Conlon and Mortimer (2008), Mehta, Rajiv, and Srinivasan (2003), Hortacsu and Syverson (2004), Santos, Hortacsu, and Wildenbeest (2012), Van Nierop, Bronnenberg, Paap, Wedel, and Franses (2010), Paola and Marco (2013) and Pires (2012). These different specifications of consideration set formation are plausible in certain applications. But this is not always the case - these structures and assumptions can be restrictive for some empirical applications and thus increase the risk of misspecification. Furthermore, even with a well specified model of consideration set generation, researchers have to use simulation and sampling techniques to integrate over all the possible consideration sets when estimating the model, see e.g., Chiang, Chib, and Narasimhan (1999), Goeree (2008), Moraga-González, Sándor, and Wildenbeest (2009) and Bruno and Vilcassim (2008), which can be computationally demanding in many empirical applications.

The moment inequality approach proposed in this paper gets around the dimensionality issue while placing very limited restrictions on the consideration set formation. The main contribution is to show that, when consumers' consideration sets are unobservable (to econometricians) and/or endogenously determined, we can still estimate the random utility model based on a system of conditional moment inequalities without modelling many details of the consideration set formation. The construction of the moment inequalities is based on (1) prior restrictions defining the set of possible consideration sets, (2) the monotonicity of choice function with respect to consideration set: if an alternative x is chosen from a set T, and x is also an element of a subset $S \subseteq T$, then the x must be chosen from S, which is a immediate implication of utility maximization assumption and also known as Chernoff's condition (Chernoff, 1954) or Sen's property α (Sen, 1971) in choice theory. With the monotonicity property, the choice probability for a product is bounded below (above) by the counterfactual choice probability if the superset (subset) of all possible consideration sets containing this product were the true consideration set whenever the product is considered. The bounds, when matched with the observed choice probabilities, imply a system of moment inequalities that can be used to estimate the parameters in the model.

The moment inequality approach imposes minimal restrictions on the consideration set distribution and thus is consistent with a general class of underlying consideration set generating processes. Such generality makes the approach immune to potential misspecifications of the consideration set formation. In addition, this approach is very easy to implement even when the number of products is large because for each product, only two consideration sets (the superset and subset) are needed to construct the bounds - there is no need to simulate each consumer's consideration sets in the estimation process as typically required when using a complete model of consideration set generation. Finally, I show that the moment inequalities can be informative enough to point identify the parameters under certain conditions. These conditions clarify the requirements on the variations of the data and necessary restrictions on the model.

I apply the proposed moment inequality approach to the IRI household panel and store scanner data, which is a leading source for demand information of consumer packaged goods, see Bronnenberg, Kruger, and Mela (2008) for an introduction of this data set. As pointed out by Keane (2013), the most salient feature of scanner panel data is that consumers' brand choice exhibits strong persistence. The literature on explaining the persistence has been focusing on the potential sources from preference, e.g., state dependence and persistent preference heterogeneity discussed intensively in Keane (1997), Seetharaman, Ainslie, and Chintagunta (1999), Dubé, Hitsch, and Rossi (2010) and Sudhir and Yang (2014). However, focusing on the preference ignores a potentially important channel through which the past purchase experiences affect the current decision: previous choice of a brand increases its chance of getting into the consideration set for the current purchase. Both marketing and economics literature provide evidence that previous choice plays an important role in the formation of consumers' consideration sets, see, e.g., Ratneshwar and Shocker (1991), Nedungadi (1990), Kardes, Kalyanaram, Chandrashekaran, and Dornoff (1993) and Chiang, Chib, and Narasimhan (1999).

The empirical question of this paper is whether the consideration set formation can explain some of the observed persistence. To address this question, I set up a model in which the previous choice enters both utility function and consideration set generation; the identification is aided by the in-store display promotion that is assumed to shift consideration set formation but be excluded from utility function. The estimation results shows that up to 20% of the observed persistence in terms of re-purchase probability are driven by the effect of previous choice on the formation of current consideration set.

The rest of the paper is organized as follows. Section 2 sets up the random utility model, derives the moment inequalities, introduces the sufficient conditions for identification and comments on the connections to the consideration set generation models in the literature. Section 3 presents a consistent point estimator based on the conditional moment inequalities. Sections 4 and 5 report the results of Monte Carlo experiments and the empirical applications.

Section 6 concludes.

2 Model and Identification

2.1 Setup

Consider a generic market consisting of a universal set of differentiated products $\mathcal{J} = \{0, 1, ..., J\}$ and a population of independent and ex ante identical consumers, indexed by $i = 1, ..., N$. The product labeled by 0 is the "outside option" and those labeled by $j = 1, \ldots, J$ are the "inside goods". Each consumer i is associated with a set of consumer-/product-specific observables $X_i \equiv (X_{i1}, ..., X_{iJ})$. Typically X_i includes product characteristics, e.g., price, consumer attributes, e.g., income, as well as the interactions between them. Each consumer i's purchase decision is observable and designated as $d_i \equiv (d_{i0}, d_{i1}..., d_{iJ})$, where $d_{ij} = 1$ when consumer i chooses product j.

Consumer preference is represented by a random utility model; the indirect utility to consumer i from choosing j is

$$u_{ij} = U\left(X_{ij}; \lambda\right) + \varepsilon_{ij}, \tag{2.1}$$

where

$$\varepsilon_i \equiv (\varepsilon_{i0}, ..., \varepsilon_{iJ}) \sim F_\varepsilon\left(\cdot \,|X_i; \rho\right)$$

represents an idiosyncratic preference shock, λ and ρ are finite dimensional vectors of parameters. Allowing X to enter F_ε, the specification nests the usual random coefficients logit model (see, e.g., Berry, Levinsohn, and Pakes, 2004)

$$\begin{aligned} U_j\left(X_i, \varepsilon_i; \lambda\right) &= X_{ij}\left(\lambda + \nu_i\right) + \upsilon_{ij} \\ &= X_{ij}\lambda + \underbrace{\left(X_{ij}\nu_i + \upsilon_{ij}\right)}_{\varepsilon_{ij}} \end{aligned} \tag{2.2}$$

where the random coefficient $\nu_i \sim N\left(0, \Sigma\left(\rho\right)\right)$, υ_{ij} follows the standard Gumbel distribution

and is i.i.d. across j.

Econometricians can not observe each consumer i's consideration set and thus regard it as a discrete random variable taking subsets of $\mathcal{J}$ as realizations. Assume that economic theory and institutional knowledge allow researchers to restrict the set of possible consideration sets, denoted as $\mathcal{C}(X_i)$, as a strict subset of the power set $2^{\mathcal{J}}$. This assumption is not particularly restrictive because researchers almost always have to define the set of possible consideration sets as a preliminary step before setting up any consideration set generation model in empirical applications. In the following, I show some examples of $\mathcal{C}(X_i)$.

Example. Restrictions on possible consideration sets.

1. A simple and commonly used restriction is that outside option is always in each consumer i's consideration set, which implies that

$$\mathcal{C}(X_i) = \left\{\mathcal{I} \in 2^{\mathcal{J}} : \ 0 \in \mathcal{I}\right\}.$$

   This assumption is typically maintained in empirical applications as long as an outside option is defined, see among others, Goeree (2008) and Moraga-González, Sándor, and Wildenbeest (2009).

2. In the empirical application of this paper, I assume that (1) all the products that are on display[1] in a store are automatically considered by the consumers who visits the store, (2) each consumer i only considers the products that are available at the time/store she makes the purchase. Let $\mathcal{D}_i$ and $\mathcal{S}_i$ denote the display set and the available set respectively at the time/store that i makes the purchase. Then the assumptions imply

$$\mathcal{C}(X_i) = \left\{\mathcal{I} \in 2^{\mathcal{J}} : \ \mathcal{D}_i \subseteq \mathcal{I} \subseteq \mathcal{S}_i\right\}.$$

[1] According to wikipedia, "In-store displays are promotional fixtures in retail stores. Variations of in-store displays include Point-of-Sale Displays, which are located near cash registers to encourage impulse buying; Floor Stickers, or advertisements for products on the aisle of a store; Feature Displays, which can be located at the end of an aisle to draw attention to a product; and Special Racks, or manipulation of a store shelf to make more space available for a product or bring attention to the promoted product. In-store Displays can be perceived as more visually appealing to consumers than product alone on a retail shelf."

Gentry (2011) also imposes these assumptions in his empirical application with store scanner data.

3. Conlon and Mortimer (2008) study a stock-out problem, in which econometricians observe the product availability at the beginning and the end of a time interval, as well as consumer i's purchase during this time period. In this case, consumer i's consideration set is unobservable because researchers do not know the product availability at the time that consumer i makes the purchase, however, consumer i's consideration set can be safely assumed to be a subset of products available at the beginning (C_i^{begin}) and a superset at the end (C_i^{end}) of the time interval, i.e.,

$$\mathcal{C}(X_i) = \left\{ \mathcal{I} \in 2^{\mathcal{J}} : C_i^{end} \subseteq \mathcal{I} \subseteq C_i^{begin} \right\}.$$

2.2 Moment Inequalities

Given the well-defined random utility model (2.1) and the set of possible consideration sets $\mathcal{C}(X_i)$, the utility maximization assumption is sufficient for constructing nontrivial bounds on the choice probabilities. The first step of the construction is defining the bounds on consideration sets in terms of inclusion order. For any product $j \in \mathcal{J}$, let

$$C_j^{sub}(X_i) = \bigcap_{\mathcal{I} \in \mathcal{C}_j(X_i)} \mathcal{I}$$

and

$$C_j^{sup}(X_i) = \bigcup_{\mathcal{I} \in \mathcal{C}_j(X_i)} \mathcal{I},$$

where $\mathcal{C}_j(X_i) \equiv \{\mathcal{I} \in \mathcal{C}(X_i) : j \in \mathcal{I}\}$ is the set of possible consideration sets that include j. It follows immediately that for any $j \in \mathcal{J}$ and $\mathcal{I} \in \mathcal{C}_j(X_i)$,

$$C_j^{sub}(X_i) \subseteq \mathcal{I} \subseteq C_j^{sup}(X_i).$$

The following example shows the bounds on the consideration set for the cases in Example 2.1, respectively.

Example. Bounds on the consideration sets for the cases in Example 2.1.

1. $C_j^{sub}(X_i) = \{0, j\}$ and $C_j^{sup}(X_i) = \mathcal{J}$, for $j \in \mathcal{J}$
2. $C_j^{sub}(X_i) = \{j\} \bigcup \mathcal{D}_i$ and $C_j^{sup}(X_i) = \mathcal{S}_i$, for $j \in \mathcal{S}_i$.
3. $C_j^{sub}(X_i) = \{j\} \bigcup C_i^{end}$ and $C_j^{sup}(X_i) = C_i^{begin}$, for $j \in C_i^{begin}$.

Combining the bounds on the consideration set with consumer utility maximization assumption, we have that for any $j \in \mathcal{J}$ and $\mathcal{I} \in \mathcal{C}_j(X_i)$,

$$\begin{aligned} & 1\left(j = \arg \max_{k \in C_j^{sup}(X_i)} \{U(X_{ik}; \lambda) + \varepsilon_{ik}\}\right) \\ \leq & 1\left(j = \arg \max_{k \in \mathcal{I}} \{U(X_{ik}; \lambda) + \varepsilon_{ik}\}\right) \\ \leq & 1\left(j = \arg \max_{k \in C_j^{sub}(X_i)} \{U(X_{ik}; \lambda) + \varepsilon_{ik}\}\right). \end{aligned} \tag{2.3}$$

This monotonicity property, implied by the utility maximization assumption, is also known as Chernoff's condition (Chernoff, 1954) or Sen's property α (Sen, 1971) of the choice function: if an alternative x is chosen from a set T, and x is also an element of a subset S of T, then the x must be chosen from S.

To proceed with the derivation, we need to define the conditional distribution of consideration set given the preference shock ε and covariates X. Let

$$\{P(\mathcal{I} | \varepsilon_i, X_i) : \mathcal{I} \in \mathcal{C}(X_i)\} \tag{2.4}$$

denote the probability measures on all possible consideration sets of consumer i. And by the definition of $\mathcal{C}(X_i)$,

$$P(\mathcal{I} | \varepsilon_i, X_i) = 0 \text{ for } \mathcal{I} \notin \mathcal{C}(X_i).$$

Two comments on the distribution of consideration set are necessary before continuing with the derivation.

1. The distribution in (2.4) is characterized by $2^{|\mathcal{C}(X_i)|}-1$ $\left(\approx O\left(2^J\right)\right)$ unknown functions. So it is in general intractable to treat them as nonparametric parameters and estimate them even when J is only moderately large. Many consideration set generation models in the literature can be viewed as different ways of reducing the dimension by imposing certain structures that are characterized by a small number of parameters, e.g., Goeree (2008), Hortacsu and Syverson (2004). The moment inequality approach presented below provides an alternative way of getting around this dimensionality issue without imposing too much structure on the consideration set formation.

2. Allowing the distribution of consideration sets to depend on the preference shock, the specification can capture possible correlations between preference and consideration set generation given the observables X. The correlations are natural implications of many search models. For example, Santos, Hortacsu, and Wildenbeest (2012) assume that consumers learn ε before search and thus the consideration set (the set of searched products) naturally depends on the realization of ε. Another example is the sequential search model (see, e.g, Hortacsu and Syverson (2004)), in which each consumer searches for an additional product as long as the expected marginal benefit is greater than the search cost. Since the realized preference shocks of the products that have been already searched affect the current decision, the probability of a consideration set thus depends on a subvector or the whole vector of preference shocks.

Next, for a given $j \in \mathcal{J}$, we can aggregate the inequalities (2.3) over $\mathcal{I} \in \mathcal{C}_j(X_i)$ to obtain

$$\begin{aligned}
&\sum_{\mathcal{I}\in\mathcal{C}_j(X_i)} 1\left(j = \arg\max_{k\in C_j^{sup}(X_i)} \{U(X_{ik};\lambda) + \varepsilon_{ik}\}\right) P(\mathcal{I}\,|\varepsilon_i, X_i) \\
\leq &\sum_{\mathcal{I}\in\mathcal{C}_j(X_i)} 1\left(j = \arg\max_{k\in\mathcal{I}} \{U(X_{ik};\lambda) + \varepsilon_{ik}\}\right) P(\mathcal{I}\,|\varepsilon_i, X_i) \\
\leq &\sum_{\mathcal{I}\in\mathcal{C}_j(X_i)} 1\left(j = \arg\max_{k\in C_j^{sub}(X_i)} \{U(X_{ik};\lambda) + \varepsilon_{ik}\}\right) P(\mathcal{I}\,|\varepsilon_i, X_i).
\end{aligned}$$

Observing that $C_j^{sub}(X_i)$ and $C_j^{sub}(X_i)$ are invariant across the index of the summation, we have

$$\begin{aligned}
&1\left(j = \arg\max_{k\in C_j^{sup}(X_i)} \{U(X_{ik};\lambda) + \varepsilon_{ik}\}\right) \sum_{\mathcal{I}\in\mathcal{C}_j(X_i)} P(\mathcal{I}\,|\varepsilon_i, X_i) \\
\leq &\sum_{\mathcal{I}\in\mathcal{C}_j(X_i)} 1\left(j = \arg\max_{k\in\mathcal{I}} \{U(X_{ik};\lambda) + \varepsilon_{ik}\}\right) P(\mathcal{I}\,|\varepsilon_i, X_i) \\
\leq &1\left(j = \arg\max_{k\in C_j^{sub}(X_i)} \{U(X_{ik};\lambda) + \varepsilon_{ik}\}\right) \sum_{\mathcal{I}\in\mathcal{C}_j(X_i)} P(\mathcal{I}\,|\varepsilon_i, X_i). \qquad (2.5)
\end{aligned}$$

Furthermore, the sum of probabilities of consideration sets that include product j is equal to the probability of j being considered, i.e.,

$$\sum_{\mathcal{I}\in\mathcal{C}_j(X_i)} P(\mathcal{I}\,|\varepsilon_i, X_i) = \Pr(j \in \mathcal{I}\,|\varepsilon_i, X_i), \qquad (2.6)$$

where $\mathcal{I}$ follows the distribution (2.4). Letting

$$\mu_{ij} \equiv \Pr(j \in \mathcal{I}\,|\varepsilon_i, X_i) \qquad (2.7)$$

and

$$\sigma_j(X_i, \varepsilon_i\,|\mathcal{I};\lambda) = 1\left(j = \arg\max_{k\in\mathcal{I}} \{U(X_{ik};\lambda) + \varepsilon_{ik}\}\right),$$

we can rewrite (2.5) as

$$\begin{aligned}\sigma_j\left(X_i,\varepsilon_i\left|C_j^{sup}\left(X_i\right);\lambda\right.\right)\mu_{ij} &\leq \sum_{\mathcal{I}\in\mathcal{C}_j(X_i)}\sigma_j\left(X_i,\varepsilon_i\left|\mathcal{I};\lambda\right.\right)P\left(\mathcal{I}\left|\varepsilon_i,X_i\right.\right)\\ &\leq \sigma_j\left(X_i,\varepsilon_i\left|C_j^{sub}\left(X_i\right);\lambda\right.\right)\mu_{ij}. \end{aligned} \tag{2.8}$$

Finally, we can integrate out ε in (2.8) and obtain

$$\begin{aligned}&\int\sigma_j\left(X_i,\varepsilon_i\left|C_j^{sup}\left(X_i\right);\lambda\right.\right)\mu_{ij}dF_\varepsilon\left(\varepsilon_i\left|X_i;\rho\right.\right)\\ \leq &\int\sum_{\mathcal{I}\in\mathcal{C}_j(X_i)}\sigma_j\left(X_i,\varepsilon_i\left|\mathcal{I};\lambda\right.\right)P\left(\mathcal{I}\left|\varepsilon_i,X_i\right.\right)dF_\varepsilon\left(\varepsilon_i\left|X_i;\rho\right.\right)\\ \leq &\int\sigma_j\left(X_i,\varepsilon_i\left|C_j^{sub}\left(X_i\right);\lambda\right.\right)\mu_{ij}dF_\varepsilon\left(\varepsilon_i\left|X_i;\rho\right.\right). \end{aligned} \tag{2.9}$$

Observe that consumer i's choice probability of product j is

$$\begin{aligned}\Pr\left(d_{ij}=1\left|X_i\right.\right) &= \int\sum_{\mathcal{I}\in\mathcal{C}(X_i)}\sigma_j\left(X_i,\varepsilon_i\left|\mathcal{I};\lambda_0\right.\right)P\left(\mathcal{I}\left|\varepsilon_i,X_i\right.\right)dF_\varepsilon\left(\varepsilon_i\left|X_i;\rho_0\right.\right)\\ &= \int\sum_{\mathcal{I}\in\mathcal{C}_j(X_i)}\sigma_j\left(X_i,\varepsilon_i\left|\mathcal{I};\lambda_0\right.\right)P\left(\mathcal{I}\left|\varepsilon_i,X_i\right.\right)dF_\varepsilon\left(\varepsilon_i\left|X_i;\rho_0\right.\right), \end{aligned}$$

where (λ_0,ρ_0) is the true value of the parameter vector, the second equality holds because $\sigma_j\left(X_i,\varepsilon_i\left|\mathcal{I};\lambda_0\right.\right)=0$ for any $\mathcal{I}\notin\mathcal{C}_j\left(X_i\right)$. Thus we have the following results.

Theorem 1. *For each consumer i and product $j\in\mathcal{J}$,*

$$\begin{aligned}&\int\sigma_j\left(X_i,\varepsilon_i\left|C_j^{sup}\left(X_i\right);\lambda_0\right.\right)\mu_{ij}dF_\varepsilon\left(\varepsilon_i\left|X_i;\rho_0\right.\right)\\ \leq &\Pr\left(d_{ij}=1\left|X_i\right.\right)\\ \leq &\int\sigma_j\left(X_i,\varepsilon_i\left|C_j^{sub}\left(X_i\right);\lambda_0\right.\right)\mu_{ij}dF_\varepsilon\left(\varepsilon_i\left|X_i;\rho_0\right.\right). \end{aligned} \tag{2.10}$$

Theorem 1 states that the choice probability $\Pr\left(d_{ij}=1\left|X_i\right.\right)$, which can be directly recovered from consumer choice data, is bounded above (below) by the counterfactual choice

probability predicted by the random utility model when $C_j^{sub}(X_i)$ ($C_j^{sup}(X_i)$) is the only element in $\mathcal{C}_j(X_i)$. To see why, note that if $\mathcal{C}_j(X_i)$ only contains one element $C_j^{sub}(X_i)$, then the consideration probability of j equals to the probability of consideration set $C_j^{sub}(X_i)$, i.e.,

$$\mu_{ij} = P\left(C_j^{sub}(X_i) \,|\varepsilon_i, X_i\right).$$

And thus the upper bound in (2.10) becomes the choice probability of j if consumer i has $C_j^{sub}(X_i)$ as her consideration set whenever she considers product j. The lower bound has the same interpretation but with $C_j^{sub}(X_i)$ replaced by $C_j^{sup}(X_i)$.

To construct moment inequalities based on Theorem (1), we need to model the consideration probability of j, μ_{ij}. Since μ_{ij} is a function of ε_i and X_i, we can write it as

$$\mu_{ij} = h_j(\varepsilon_i, X_i; \varphi_0), \tag{2.11}$$

where φ_0 is the true value of a parameter vector φ. The specification of h_j can be very flexible and even nonparametric (φ becomes infinite dimensional) if necessary, however, too flexible specifications may reduce statistical efficiency and cause difficulties for identification; I shall discuss this issue in the Subsection 2.3.

Back to the dimensionality issue of modelling (2.4), now we only need to model one consideration probability (2.11) for each product instead of modelling $|\mathcal{C}(X_i)|$ conditional probability functions $\{P(\mathcal{I}\,|\varepsilon_i, X_i) : \mathcal{I} \in \mathcal{C}(X_i)\}$, i.e., we reduce the dimension of the problem from $O\left(2^J\right)$ to $O(J)$.

With the parametrization (2.11), we can re-write the inequalities (2.10) as a system of

conditional moment inequalities,

$$\begin{aligned}
&\mathbf{E}\left[d_{ij}-\int\sigma_j\left(X_i,\varepsilon_i\left|C_j^{sup}\left(X_i\right);\lambda_0\right.\right)h_j\left(\varepsilon_i,X_i;\varphi_0\right)dF_\varepsilon\left(\varepsilon_i\left|X_i;\rho_0\right.\right)\right|X\Big]\geq 0\\
&\mathbf{E}\left[\int\sigma_j\left(X_i,\varepsilon_i\left|C_j^{sub}\left(X_i\right);\lambda_0\right.\right)h_j\left(\varepsilon_i,X_i;\varphi_0\right)dF_\varepsilon\left(\varepsilon_i\left|X_i;\rho_0\right.\right)-d_{ij}\right|X\Big]\geq 0\ \forall j\ a.s.\left[X_i\right]
\end{aligned}\tag{2.12}$$

A natural question is whether the moment inequalities (2.12) characterize the sharp identified set of the parameters, i.e., whether they exhaust all the information in the model. The answer is no because the bounds in (2.12) can be attained for a given product but typically not for all products simultaneously. To see this, consider the case where there are two products j and $j^{'}$ such that $C_k^{sub}\left(X_i\right)=\{0,k\}$ and $\mu_{ik}=.8$ for $k=j,j^{'}$. The upper bounds on choice probability of j and $j^{'}$ can not be attained at the same time because $\mu_{ij}+\mu_{ij^{'}}>1$, i.e., consumer i must have some positive probability on consideration sets that include both j and $j^{'}$; thus her possible consideration sets can not be just $\{0,j\}$ and $\left\{0,j^{'}\right\}$.

We can in principle apply the techniques of Beresteanu and Molinari (2008); Beresteanu, Molchanov, and Molinari (2011) to exploit all the information in the model to characterize the sharp identified set of the parameters, as shown in Appendix (C). However, the computational burden becomes prohibitive when there are more than a handful products. On the contrary, the moment inequalities (2.12) does not suffer from the dimensionality issue of the power set and is thus computationally very easy to work with. Certainly, a potential concern is that (2.12) may not be very informative about parameters because they entails information loss from the non-sharpness. To address this concern, the next subsection explores the conditions under which (2.12) have substantial identification power and point identify the parameters in the model.

2.3 Identification

Although the parameters of interest are only restricted by moment inequalities, it is still possible to find sufficient conditions for point identification. These conditions clarify the requirements on the data and assumptions needed for identification. Also, under point identification, we can obtain point estimates using standard numerical optimization techniques and thus avoid the need for finding the set estimate, which can be challenging in many empirical applications when the number of parameters is large.

Following the ideas of Powell (1984) and Kahn and Tamer (2009), I provide a sufficient condition for identification which guarantees that for any $(\lambda, \rho, \varphi) \neq (\lambda_0, \rho_0, \varphi_0)$, some inequalities are violated. To introduce the condition, I first define the following two quantities. The first one is a measure of the change in bounds on the choice probabilities when the parameter vector deviates from the true value, i.e.,

$$\begin{aligned} R_j^m\left(X_i;\theta,\theta_0\right) &= \int \sigma_j\left(X_i,\varepsilon_i\left|C_j^m\left(X_i\right);\lambda\right.\right) h_j\left(\varepsilon_i,X_i;\varphi\right) dF_\varepsilon\left(\varepsilon_i\left|X_i;\rho\right.\right) \\ &- \int \sigma_j\left(X_i,\varepsilon_i\left|C_j^m\left(X_i\right);\lambda_0\right.\right) h_j\left(\varepsilon_i,X_i;\varphi_0\right) dF_\varepsilon\left(\varepsilon_i\left|X_i;\rho_0\right.\right), \ m \in \{sub, sup\} \end{aligned}$$

where $\theta \equiv (\lambda, \rho, \varphi)$ denote the vector of all the parameters in the model. The other one is a measure of the "slackness" of the bounds, $C_j^{sub}(X_i)$ and $C_j^{sup}(X_i)$,

$$\Delta_j^m\left(X_i;\theta_0\right) = \int \left[h_j\left(\varepsilon_i,X_i;\varphi_0\right) - P\left(C_j^m\left(X_i\right)\left|\varepsilon_i,X_i\right.\right)\right] dF_\varepsilon\left(\varepsilon_i\left|X_i;\rho_0\right.\right), \ m \in \{sub, sup\}.$$

Recall that the upper bound of (2.10) is attained when

$$\mu_{ij} = P\left(C_j^{sub}\left(X_i\right)\left|\varepsilon_i,X_i\right.\right)$$

so $\Delta_j^{sub}(X_i;\theta_0)$ can be interpreted as the departure from this boundary case. The same interpretation applies to $\Delta_j^{sup}(X_i;\theta_0)$.

Let S_X denote the support of X_i and the following result provides a sufficient condition

for point identification.

Theorem 2 (Point Identification)**.** *If for any $\theta \neq \theta_0$, there exists some $j \in \mathcal{J}$ such that*

$$\left\{X_i \in S_X : R_j^{sup}\left(X_i; \theta, \theta_0\right) > \Delta_j^{sup}\left(X_i; \theta_0\right)\right\} \tag{2.13}$$

or

$$\left\{X_i \in S_X : R_j^{sub}\left(X_i; \theta, \theta_0\right) < -\Delta_j^{sub}\left(X_i; \theta_0\right)\right\} \tag{2.14}$$

has positive measure, then θ_0 is point identified.

Proof. See AppendixA. □

Intuitively, Theorem (2) states that if for any $\theta \neq \theta_0$, there is a positive measure of consumers (we do not need to know who) whose consideration set distributions assign high probabilities to either $C_j^{sub}\left(X_i\right)$ or $C_j^{sup}\left(X_i\right)$ for some product j so that $\Delta_j^{sub}\left(X_i; \theta_0\right)$ or $\Delta_j^{sup}\left(X_i; \theta_0\right)$ is small for some j; and, for these consumers, θ shifts the bounds on the choice probabilities sufficiently so that $R_j^{sup}\left(X_i; \theta, \theta_0\right)$ is a large positive number or $R_j^{sub}\left(X_i; \theta, \theta_0\right)$ is a large negative number for the corresponding j, then those inequalities indexed by j have a good chance of being violated and thus the point identification is achieved.

It is also instructive to consider the special case in which the true consideration set of each consumer in the market is $\mathcal{J}$. Suppose econometricians know that $\mathcal{J}$ is the true consideration set and let

$$C_j^{sub}\left(X_i\right) = C_j^{sup}\left(X_i\right) = \mathcal{J}.$$

Also, by definition,

$$\mu_{ij} = h_j\left(\varepsilon_i, X_i; \varphi_0\right) = 1.$$

It follows immediately that

$$\Delta_j^{sub}\left(X_i; \theta_0\right) = \Delta_j^{sup}\left(X_i; \theta_0\right) = 0$$

Thus the identification conditions (2.13) and (2.14) becomes

$$\left| \int \sigma_j \left(X_i, \varepsilon_i \left| \mathcal{J}; \lambda \right. \right) dF_\varepsilon \left(\varepsilon_i \left| X_i; \rho \right. \right) - \int \sigma_j \left(X_i, \varepsilon_i \left| \mathcal{J}; \lambda_0 \right. \right) dF_\varepsilon \left(\varepsilon_i \left| X_i; \rho_0 \right. \right) \right| > 0$$

with positive probaiblity, which is the standard identification condition in traditional discrete choice models with the full consideration set assumption.

The identification condition (2.13) is easy to satisfy when $R_j^{sup}\left(X_i;\theta,\theta_0\right)$ $\left(-R_j^{sub}\left(X_i;\theta,\theta_0\right)\right)$ is large or $\Delta_j^{sup}\left(X_i;\theta_0\right)$ $\left(\Delta_j^{sub}\left(X_i;\theta_0\right)\right)$ is small. To make $\Delta_j^{sup}\left(X_i;\theta_0\right)$ $\left(\Delta_j^{sub}\left(X_i;\theta_0\right)\right)$ small, $C_j^{sup}\left(X_i\right)$ $\left(C_j^{sub}\left(X_i\right)\right)$ must be informative in the sense that there is a positive measure of consumers that have $C_j^{sup}\left(X_i\right)$ $\left(C_j^{sub}\left(X_i\right)\right)$ as their consideration sets. To make $R_j^{sup}\left(X_i;\theta,\theta_0\right)$ $\left(-R_j^{sub}\left(X_i;\theta,\theta_0\right)\right)$ large, we need large variations in X_i and could impose some restrictions on the model that are sensible in specific applications.

One commonly used assumption is that the preference shock and consideration set are conditionally independent given the observables, see, e.g., Manski (1977), Ben-Akiva and Boccara (1995), Chiang, Chib, and Narasimhan (1999), Goeree (2008), Bruno and Vilcassim (2008), Conlon and Mortimer (2008). Using the notation in this paper, this assumptions means $P\left(\mathcal{I}\left|\varepsilon_i, X_i\right.\right) = P\left(\mathcal{I}\left|X_i\right.\right)$ for any $\mathcal{I} \in \mathcal{C}\left(X_i\right)$, which implies that the consideration probability does not depend on ε, i.e.,

$$\mu_{ij} = h_j\left(X_i;\varphi_0\right). \tag{2.15}$$

The conditional independence assumption can also be satisfied in the fixed sample size search models (see, e.g., Santos, Hortacsu, and Wildenbeest (2012) and Moraga-González, Sándor, and Wildenbeest (2009)), provided that consumers do not know ε before the search, i.e., the search process does not depend on the realization of ε.

Another type of assumptions are exclusion restrictions, i.e., some variables are assumed to only affect the consideration set formation and are excluded from utility function. For example, in Goeree (2008)'s empirical application, advertising expenditure is assumed to be

excluded from utility function; Gentry (2011) makes a similar assumption on in-store display promotion.

Functional form restrictions are also useful for identification. For example, we can assume the consideration probability function (2.7) are symmetric across products, i.e.,

$$h_j\left(\varepsilon_i, X_i; \varphi_0\right) = h\left(\varepsilon_{ij}, X_{ij}, \varepsilon_{i,-j}, X_{i,-j}; \varphi_0\right),$$

where $\varepsilon_{i,-j}$ and $X_{i,-j}$ refers to the preference shocks and characteristics of all the products other than j. Or we may further assume that the consideration probability of product j only depends its own X and ε, i.e.,

$$h_j\left(\varepsilon_i, X_i; \varphi_0\right) = h\left(\varepsilon_{ij}, X_{ij}; \varphi_0\right).$$

2.4 Connections to Consideration Set Generation Models

In this subsection, I present several models of consideration set generation appeared in the literature and comment on the connections to the moment inequality approach proposed in this paper.

Example 1 (Independent Consideration Model)**.** As in Ben-Akiva and Boccara (1995), Andrews and Srinivasan (1995), Goeree (2008), etc., after imposing the conditional independence between preference and consideration set (so that (2.15) holds), the model generates the consideration set distribution based on parametrized consideration probability for each product and the independence assumption, i.e.,

$$P\left(\mathcal{I} \left| X_i \right.\right) = \prod_{l \in \mathcal{I}} h_l\left(X_i; \varphi\right) \prod_{m \notin \mathcal{I}} \left[1 - h_m\left(X_i; \varphi\right)\right]. \tag{2.16}$$

The independent consideration model provide a simple way to generate consideration set distribution from the consideration probabilities of products, however, the assumption

of independent consideration can be violated if, say, the consideration probabilities of two products are affected by some common unobservable. As an example, consider the supermarket environment, the shelf position of a product is unobservable and hard to measure. If two products are next to each other, then their consideration probabilities are naturally correlated, which breaks the independence. This assumption can also be violated in various search models presented in the next few examples (the violation can be easily verified by checking (2.16) in these models).

Suppose (2.15) holds, then my moment inequality approach only specifies $h_j(X_i;\varphi)$ for $j = 1, ..., J$ without imposing any dependence structure among these consideration probabilities. So in this special case, the moment inequality approach becomes useful when researchers are skeptical about the independence assumption and need a more robust method to estimate the model.

Example 2 (Fixed Sample Size Search)**.** Let

$$M_{\mathcal{I}}(X_i, \varepsilon_i) = E_p\left[\max_{k\in\mathcal{I}} u_{ik}\right] - \kappa(X_i)\cdot|\mathcal{I}|$$

be the expected payoff (expected utility minus search cost) of searching choice set $\mathcal{I}$, where $\kappa(X_i)$ is the cost of searching one product. The optimal search set is generated by the following maximization problem

$$\max_{\mathcal{I}\in\mathcal{C}(X_i)} \{M_{\mathcal{I}}(X_i, \varepsilon_i) + \epsilon_{\mathcal{I}}\},$$

where $\epsilon_{\mathcal{I}}$ is a standard Gumbel distributed disturbance term that is used mainly for smoothing the probabilities for search sets. It follows that

$$P(\mathcal{I}\,|\varepsilon_i, X_i) = \frac{\exp[M_{\mathcal{I}}(X_i, \varepsilon_i)]}{\sum_{\mathcal{I}'\in\mathcal{C}}\exp[M_{\mathcal{I}'}(X_i, \varepsilon_i)]}.$$

Similar models have been used in Mehta, Rajiv, and Srinivasan (2003), Santos, Hortacsu,

and Wildenbeest (2012), Moraga-González, Sándor, and Wildenbeest (2009), etc.

Example 3 (Sequential Search Models)**.** Consider a sequential search model in which consumer i samples products with replacement with a given success probability vector $(p_1, ..., p_J)$. Assume there is an underlying search cost distribution that determines the probability of searching M products, $q_M(X_i, \varepsilon_i) \quad \forall M = 0, 1, ..., J$. Let n_j denote the times that j is drawn. We can obtain the probability of search set $\mathcal{I} \in \mathcal{C}(X_i)$ as

$$P(\mathcal{I}|\varepsilon_i, X_i) = q_M(X_i, \varepsilon_i) \sum_{N=|\mathcal{I}|}^{\infty} \sum_{\substack{n_0 + ... + n_J = N \\ n_k \geq 1 \ \forall k \in \mathcal{I} \\ n_m = 0 \ \forall k \notin \mathcal{I}}} \left[\frac{N!}{n_1! \cdots n_J!} (p_1)^{n_1} \cdots (p_J)^{n_J} \right].$$

This sampling process has been used in Carlson and McAfee (1983), Hortacsu and Syverson (2004), etc.

In each of the two examples, we can write down the implied consideration probability for each product, using (2.6), given the expressions for probabilities of consideration sets. Also, it is easy to verify that these consideration probabilities are not independent, i.e., they have a specific dependent structure implied by the search mechanism. Since the moment inequality approach allows for arbitrary correlation between the consideration probabilities, it can be regarded as a reduced-form approximation to a wide range of search models. [2]If the objective is to estimate the utility function and researchers would like to control the consideration set heterogeneity and endogeneity in a general way, then the moment inequality approach may be a good option.

[2]However, if the object of interests is the underlying structure of the search, e.g., search cost distribution, then the moment inequality approach may not be very useful because it does not specify these structures.

3 Estimation

The estimation is based on the conditional moment inequalities (2.10), which can be written as

$$\begin{aligned} \mathbf{E}\left[m_j^{sub}\left(X_i, d_i; \theta\right) \middle| X_i\right] &\geq 0 \\ \mathbf{E}\left[m_j^{sup}\left(X_i, d_i; \theta\right) \middle| X_i\right] &\geq 0 \; \forall j \in \mathcal{J} \; a.s. \left[X_i\right], \end{aligned} \tag{3.1}$$

where

$$m_j^{sup}\left(X_i, d_i; \theta\right) \equiv d_{ij} - \int \sigma_j\left(X_i, \varepsilon_i \middle| C_j^{sup}\left(X_i\right); \lambda_0\right) h_j\left(\varepsilon_i, X_i; \varphi_0\right) dF_\varepsilon\left(\varepsilon_i \middle| X_i; \rho_0\right)$$

and

$$m_j^{sub}\left(X_i, d_i; \theta\right) \equiv \int \sigma_j\left(X_i, \varepsilon_i \middle| C_j^{sub}\left(X_i\right); \lambda_0\right) h_j\left(\varepsilon_i, X_i; \varphi_0\right) dF_\varepsilon\left(\varepsilon_i \middle| X_i; \rho_0\right) - d_{ij}.$$

Following Andrews and Shi (2013), the conditional moment inequalities (3.1) can be transformed into a set of unconditional moment inequalities without information loss, i.e.,

$$\begin{aligned} \mathbf{E}\left[m_j^{sub}\left(X_i, d_i; \theta\right) g\left(X_i\right)\right] &\geq 0 \\ \mathbf{E}\left[m_j^{sup}\left(X_i, d_i; \theta\right) g\left(X_i\right)\right] &\geq 0 \; \forall j \in \mathcal{J} \; \forall g \in \mathcal{G} \; a.s. \left[X_i\right], \end{aligned} \tag{3.2}$$

where $\mathcal{G}$ is some space of instrumental functions. See Andrews and Shi (2013) for an extensive discussion about different types of instrumental functions. Then a simple analog estimator, in the spirit of Andrews and Shi (2013), can be defined as

$$\hat{\theta} = \arg\min_{\theta \in \Theta} \int \sum_{j=1}^{J} \left\{ \left(\widehat{\mathbf{E}}\left[m_j^{sub}\left(X_i, d_i; \theta\right) g\left(X_i\right)\right] \right)_-^2 + \left(\widehat{\mathbf{E}}\left[m_j^{sup}\left(X_i, d_i; \theta\right) g\left(X_i\right)\right] \right)_-^2 \right\} d\mu\left(g\right), \tag{3.3}$$

where $\widehat{\mathbf{E}}\left[\cdot\right] \equiv \frac{1}{N}\sum_{i=1}^{N}\left[\cdot\right]$, $\left(x\right)_- = \min\left(0, x\right)$, $\mu\left(\cdot\right)$ is some measure over $\mathcal{G}$. Note that if the

point identification assumption is not desirable in some applications, we can extend this procedure to obtain a consistent estimate of the identified set by constructing a level set of the objective function. For the current analysis, I focus on the point identification case. The following result shows the consistency of the estimator.

Proposition 1 (Consistency)**.** *Suppose the conditions in Theorem 2 hold. Then, under regularity conditions, the estimator* (3.3) *is consistent.*

Proof. See Appendix B. □

In Appendix B, I present a more general version of the consistency result that allows for unknown functions (nonparametric components) in the moment conditions; the unknown functions are approximated by sieve methods. This can be useful when researchers would like to specify a very flexible functional form for the consideration probability for each product.

4 Monte Carlo Simulations

In this section, I present a set of Monte Carlo experiments to illustrate the performance of the moment inequality approach. Assume there are $J = 10$ inside goods and an outside option, labeled by 0, in the market. There are $N = 1,000$ consumers in the market and the utility of consumer i from an inside product j is

$$u_{ij} = \alpha + \beta X_{ij} + \epsilon_{ij}$$

and that from the outside option is

$$u_{i0} = \epsilon_{i0},$$

where

$$X_{ij} \sim \begin{cases} N(\mu_X, 1) & \text{for } j = 1, ..., 5 \\ N(0, 1) & \text{for } j = 6, ..., 10 \end{cases}$$

and

$$\epsilon_{ij} \sim \text{i.i.d. Standard Gumbel distribution.}$$

The consideration set of each consumer i is generated as follows. The products $1, ..., 5$ and the outside option are automatically in each consumer i's consideration set. For each product $j = 6, ..., 10$, each consumer i draws a $w_{ij} \sim U[0, 1]$ and decide to include j in the consideration set if $w_{ij} < h_1$ for $i = 1, ..., 500$ or $w_{ij} < h_2$ for $i = 500, ..., 1,000$. For each consumer i, let C_i denote her realized consideration set and then her choice is generated by

$$d_{ij} = \begin{cases} 1\left(j = \arg\max_{k \in C_i} u_{ik}\right) & \text{for } j \in C_i \\ 0 & \text{otherwise.} \end{cases}$$

I consider three data generating processes by varying μ_X, h_1 and h_2, which are summarized in the following table. For each DGP, I replicate to produce 1,000 samples, implement several alternative estimation strategies on each sample, and then report the root of mean square error (RMSE), bias, standard deviation (SD) in Table 1.

DGP	μ_X	h_1	h_2
I	1	.5	.5
II	1	0	.5
III	-1	0	.5

The first strategy is estimating the standard logit model via maximum likelihood method, assuming all the 10 products are in the consideration set of each consumer. Specifically, the log-likelihood function is

$$\sum_{i=1}^{N} \left\{ \sum_{j=1}^{J} d_{ij} \log \left[\frac{\exp(\alpha + \beta X_{ij})}{1 + \sum_{k=1}^{J} \exp(\alpha + \beta X_{ik})} \right] + d_{i0} \log \left[\frac{1}{1 + \sum_{k=1}^{J} \exp(\alpha + \beta X_{ik})} \right] \right\}.$$

The results are labeled as "Full C.S. (Consideration Set)" in Table 1.

The second approach is to assume that the first 5 products are automatically considered

and the remaining products are considered independently with the consideration probability h, which becomes a parameter to be estimated. To implement this estimation strategy, I simulate each consumer's consideration set using the same procedure as in the data generating process, i.e., the first 5 products and the outside option are always in the consideration set; draw a uniform random variable for each product $j = 6, ..., 10$ and include the product if the random draw is less than h. To eliminate the simulation error, I use the same set of random draws as in the data generating process. With the simulated consideration set for each consumer i, $C_i(h)$, the log-likelihood function can be written as

$$\sum_{i=1}^{N} \left\{ \sum_{j \in C_i(h)} d_{ij} \log \left[\frac{\exp(\alpha + \beta X_{ij})}{1 + \sum_{k \in C_i(h)} \exp(\alpha + \beta X_{ik})} \right] + d_{i0} \log \left[\frac{1}{1 + \sum_{k \in C_i(h)} \exp(\alpha + \beta X_{ik})} \right] \right\}.$$

The results are reported in the panel labeled as "Ind. Consid. (Independent Consideration)" in Table 1.

Finally, I implement the moment inequality approach proposed in this paper. Given the first 5 products are always considered, the set of possible consideration sets for each consumer is

$$\mathcal{C} = \left\{ \mathcal{I} \in 2^{\mathcal{J}} : \ \{0, 1, ..., 5\} \subseteq \mathcal{I} \right\},$$

and thus the bounds on the consideration set that includes j is

$$C_j^{sub} = \{j\} \bigcup \{0, 1, ..., 5\}$$

and

$$C_j^{sup} = \mathcal{J}.$$

Table 1: Monte Carlo Results

Parameter		β			h		
DGP	Estimator	RMSE	Bias	SD	RMSE	Bias	SD
I	Full C.S.	.117	.110	.040	-	-	-
	Ind. Consid.	.049	-.003	.049	.014	-.007	.012
	Mom. Ineq.	.068	-.041	.055	.043	-.010	.042
II	Full C.S.	.193	.189	.041	-	-	-
	Ind. Consid.	.059	.015	.057	.129	.048	.120
	Mom. Ineq.	.065	-.005	.065	.028	.003	.028
III	Full C.S.	.472	-.471	.036	-	-	-
	Ind. Consid.	.313	-.309	.049	.199	.180	.085
	Mom. Ineq.	.084	.047	.069	.030	-.026	.015

Note: The numbers in the table are based on 1,000 replications.

The moment inequalities in this case can be written as

$$\begin{aligned} \mathbf{E}\left[d_{ij} - \frac{h \cdot \exp\left(\alpha + \beta X_{ij}\right)}{\sum_{k \in C_j^{sup}} \exp\left(\alpha + \beta X_{ik}\right)} \middle| X_i \right] &\geq 0 \\ \mathbf{E}\left[\frac{h \cdot \exp\left(\alpha + \beta X_{ij}\right)}{\sum_{k \in C_j^{sub}} \exp\left(\alpha + \beta X_{ik}\right)} - d_{ij} \middle| X_i \right] &\geq 0 \; \forall j \; a.s. \left[X_i\right], \end{aligned} \tag{4.1}$$

where h is imputed as 1 for $j = 1, ..., 5$ and treated as an unknown parameter for $j = 6, ..., 10$. In constructing the g functions, I follow Gandhi, Lu, and Shi (2013) to define $\mathcal{G}$ as

$$\left\{ g_{a,r}\left(X\right) = 1\left(X \in B_{a,r}\right) : B_{a,r} \in \left(\frac{a-1}{2r}, \frac{a}{2r} \right], a_u \in \{1, 2, ..., 2r\}, r = r_0, ..., \bar{r} \right\}$$

and choose $\mu\left(\cdot\right)$ such that

$$\mu\left(g_{a,r}\right) \propto \left(100 + r\right)^{-2} \left(2r\right)^{-1} \text{ for } g \in \mathcal{G}. \tag{4.2}$$

I use $r_0 = 2$ and $\bar{r} = 15$ for the reported results. I also tried several other values of $\bar{r}$ and the results are not sensitive to the its choice. The results are presented in in Table 1, labeled as "Mom. Ineq. (Moment Inequalities)".

For all the three DGPs, the full consideration set approach yields pretty large biases for

the slope parameter β. To get some intuition for the direction of the biases, let us consider a simple regression model

$$y_i = \beta x_i + e_i.$$

Ignoring the consideration set heterogeneity is akin to the omitted variable problem in this simple model. If the consideration probability of a product is positively correlated with its X, as in DGP I and II, then, in terms of this simple regression, the omitted variable is positively correlated with x, implying that $\text{cov}(x_i, e_i) > 0$ and thus β is overestimated. We can see the same pattern in the results of Table (1). This intuition carries over to the negative correlation case so we can see a downward bias of β in DGP III.

In DGP I, all the consumers consider each product $j = 6, ..., 10$ independently with probability .5. In this case the independent consideration model is correctly specified and, as expected, has very small bias and standard deviations. Comparing to this benchmark case, the moment inequality approach performs reasonably well and the slight larger biases and standard deviations comes from the fact that it does not use all the information in the model, which is the independent consideration assumption in this case.

In DGP II and III, the first 500 consumers do not consider the last 5 products at all but the last 500 consumers consider each $j = 6, ..., 10$ independently with probability .5. So for the first 500 consumers, the consideration of the last 5 products are perfectly correlated. So imposing independent consideration assumption for all the consumers leads to misspecification, which shows up as biased estimates from the independent consideration model in the table. In contrast, the moment inequality approach, which does not impose any dependence structure among the consideration of products, exhibits very small biases and standard deviations in this case.

5 Empirical Application

As pointed out by Keane (2013), the most salient feature of scanner panel data is that consumer's brand choice exhibits strong persistence, i.e., a consumer has a higher probability of purchasing the brand that she bought previously. A large literature has been devoted to understanding the sources of the persistence, e.g., state dependence and unobserved heterogeneity, in terms of consumer preference under the random utility framework, see, e.g., Keane (1997), Seetharaman, Ainslie, and Chintagunta (1999), Dubé, Hitsch, and Rossi (2010). However, when the consideration set formation is taken into account, another potentially important channel through which past purchase affects the current decision emerges: previous purchase of a brand increases the probability of it being considered, which in turn increases its present choice probability[3]. As shown by Ratneshwar and Shocker (1991), Nedungadi (1990), Kardes, Kalyanaram, Chandrashekaran, and Dornoff (1993) and Chiang, Chib, and Narasimhan (1999), the purchase history plays an important role in shaping consumers' consideration sets. The behavioral explanation is that consumers' attention exhibits inertia, i.e., they tend to consider the product that was bought previously more than those were not. The empirical question in this paper is whether the consideration inertia can explain some of the observed persistence in brand choice. To answer this question, I set up a brand choice model in which last purchase of a product affects both its utility and consideration probability, and apply it to study the persistence in consumer purchase of potato chips using a household scanner data set.

5.1 Data Description

I obtain the household purchase panel and store data from Information Resources Inc. (IRI). See Bronnenberg, Kruger, and Mela (2008) for an overview of the IRI data set. The household panel data keep track of the purchase history of products in 30 categories for a sample of

[3]The two potential effects of previous purchase on the current choice, raising utility level and incerasing the chance of consideration, discussed here are closely related to Mehta, Rajiv, and Srinivasan (2001)'s definitions of "active" versus "passive" state dependence.

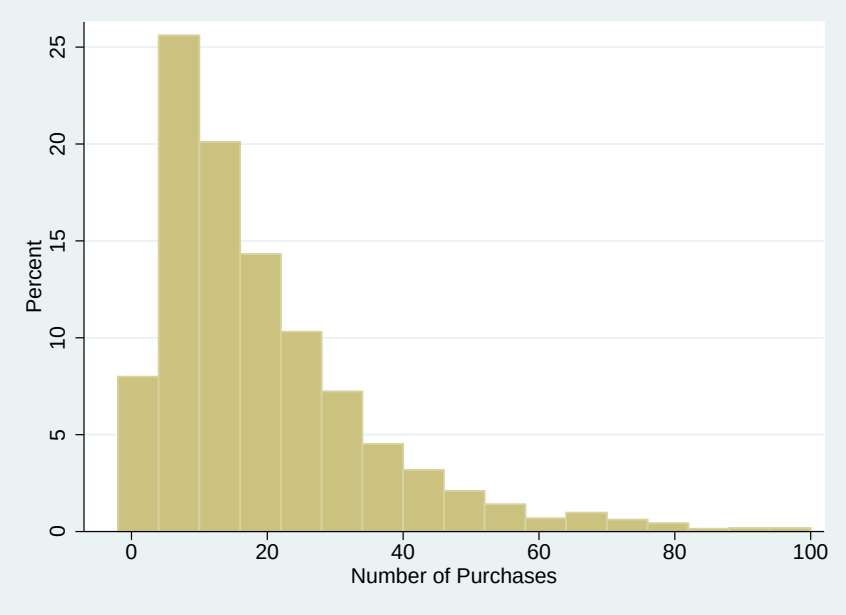
households in two cities, Eau Claire, WI and Pittsfield, MA over the years 2001-2011. For each household in the sample, the data records the timing, location and products bought on every shopping trips to the set of stores that IRI data set covers (about 80% of all the stores in the two cities). The store data contain information on quantity, price and marketing mix at store/week/UPC level. Combining household panel with store data, we can obtain the price, marketing mix for the set of UPCs that are available on shelf for the store/week at which each purchase occurred in the household panel data.

For the current empirical analysis, I focus on the household purchases of potato chips in 2010 and 2011. Potato chips come with different brands, flavors and package sizes; usually there is a large varieties of potato chips UPCs sold in a given store/week. To make the analysis tractable, I consider the consumer choice at brand level and restrict attention to the products with packages sizes between 8 and 12 ounces, which makes up more than 90% of all the purchases of potato chips observed in the sample. The inside products are defined as the top 10 brands and the outside option is the aggregation of potato chips brands outside top 10 and all other products in the broad salty snacks category, including tortilla chips, etc. With the definition of outside option, I drop the purchase occasions in which neither inside nor outside goods are purchased. And this leaves 2,782 households with 53,885 purchases in total. Table 2 and Figure 1 present some demographical statistics and a histogram of number of purchases, respectively, for the sample of households used in the subsequent analysis.

Table 2: Summary Statistics for the Sample of Households and Purchases

HH Income (dollars per year)	Percentage (%)	HH Size	Percentage (%)
0 to 19,999	18.76	1	18.73
20,000 to 34,999	20.20	2	47.05
35,000 to 64,999	31.06	3	14.56
65,000 to 99,999	19.74	4	12.40
$\geq$10,000	10.24	$\geq$5	7.07
Total Number of Households		2,782	

Figure 1: Distribution of Number of Purchases



For the store scanner data, I aggregate UPC level variables into brand level for a given week/store pair, e.g., price of a brand is a sale-weighted average (across UPCs) price index and display indicator for a brand is on if at least one UPC in it is on display. Then the store data are merged with the household purchases to re-construct the choice environment of each purchase, including the price and marketing mix of all the products available at the store/week of the purchase occasion. To sum up, for each purchase occasion in the sample, we observe the following information:

1. Price and marketing mix for the set of products available at the week/store of the purchase;

2. Product purchased;

3. Demographics and purchase history of the household that makes the current purchase.

Table 3 presents some summary statistics for the purchases of the top 10 brands. In the average price ("Avg. Price") and average display ("Avg. Display") columns, "All" refers to the average over all the store/weeks pairs in the sample whereas "Purch." means the average

conditional on purchases. The display is classified into two types by location: major includes lobby and end-of-aisle; minor usually refers to other less important spots. We can see that the price level conditional on purchase is in general lower than the unconditional one, which implies that consumers have higher probabilities of purchasing when there is price cut. For the display advertising, consumers are more likely to buy a brand on major display than the ones that are not, indicating major display is pretty effective in stimulating consumer demand. There are rich variations across brands in terms of price cut and display activities, e.g., 60% of Lays Natural purchases are associated with major display while only 14% of Cape Cod purchases are.

The "Avail. Rate" measures the percent of store/week pairs in which the brand is available on the shelf. The "Purch. Prob." column presents the choice probabilities of the top 10 brands. The re-purchase probability, showed in the last column, for a given brand is the choice probability of this brand conditional on purchasing the same brand in the last purchase. We can see that the re-purchase probabilities are generally much higher than the unconditional choice probabilities, which resembles the consumer brand choice persistence documented in the literature; see Dubé, Hitsch, and Rossi (2010) and the references therein.

Table 3: Summary Statistics for Purchases by Brands

	Avg. Price (per 16oz)		Avg. Display				Avail. Rate	Purch. Prob.	Re-Purch. Prob.
			All		Purch.				
	All	Purch.	Major	Minor	Major	Minor	(%)	(%)	(%)
Lays	4.97	4.39	.43	.21	.55	.17	100	29.59	48.01
Wavy Lays	5.07	4.42	.40	.23	.51	.20	100	19.03	39.30
Lays Natural	5.91	4.59	.40	.28	.61	.25	85.91	16.83	48.89
Cape Cod	6.13	5.80	.08	.26	.14	.21	55.93	7.17	53.57
Ruffles	6.28	5.30	.11	.05	.36	.15	99.18	5.94	24.84
Old Dutch	4.68	3.78	.32	.05	.45	.15	46.14	5.04	25.96
UTZ	5.57	5.51	.14	.07	.23	.08	23.00	4.98	56.14
Lays Kettle	5.80	5.37	.38	.22	.51	.26	81.53	4.47	36.72
Ripples	5.11	3.74	.35	.05	.78	.05	45.92	3.96	26.61
Rachels	4.54	3.84	.25	.12	.50	.18	44.51	3.00	29.68

5.2 Model Specification

In each time period (defined as a purchase occasion), every consumer forms a consideration set and then make a purchase decision. Let i and t index consumer and purchase occasion, respectively. The utility to consumer i of buying j at t is

$$u_{ijt} = \delta_j + \beta_i p_{jt} + \varphi \cdot 1\left(d_{ij,t-1} = 1\right) + \varepsilon_{ijt}.$$

Here, δ_j denotes the brand level effect for j, p_{jt} is the sale-weighted price index for brand j at t. The price coefficient

$$\beta_i = \beta + \beta^S HHSize_{it} + \beta^I HHInc_{it}$$

vary across groups of consumers divided by household size and income level. The state-dependent utility is represented by the common Markov form of state dependence, see, e.g., Dubé, Hitsch, and Rossi (2010). More general forms of the state dependence term has been adopted in the literature; for example, Guadagni and Little (1983) and Keane (1997) use the exponentially smoothed weighted average of past purchases, with an additional smoothing parameter to be chosen or estimated. The preference shock ε_{ijt} is the Standard Gumbel distribution that is i.i.d. across i,j, t. This distributional assumption brings computational advantages comparing to the probit model because conditioning on previous purchase, ε_{ijt} is independent across t and thus the current choice probability has the usual logit probability form, see Erdem (1996) for a similar specification.

With this specification, the coefficient φ on the last purchase may capture a mix effect of true state dependence and spurious state dependence generated by autocorrelated unobserved heterogeneity. Since the goal of this empirical exercise is not to disentangle true state dependence and unobserved heterogeneity, I interpret $\varphi \cdot 1\left(d_{ij,t-1} = 1\right)$ as a reduced form term of state dependence capturing the overall effect of last purchase on the utility of current

purchase.

Now let us move to the specification of consideration set distribution. To define the "largest" and "smallest" consideration sets, C_{ijt}^{super} and C_{ijt}^{sub}, I assume that each consumer i at t automatically consider the outside option and the products that are on major display at t. This assumption implies a natural lower bound on i's consideration set. Thus, we have

$$C_{ijt}^{sub} = \begin{cases} \mathcal{D}_{it}^{major} \bigcup \{0\} & \text{if } j \in \mathcal{D}_{it}^{major} \\ \mathcal{D}_{it}^{major} \bigcup \{0\} \bigcup \{j\} & \text{if } j \notin \mathcal{D}_{it}^{major}, \end{cases}$$

where $\mathcal{D}_{it}^{major}$ is the set of brands that are on major display at t for consumer i. Also, assume that C_{ijt}^{super} is the set of brands available on the shelf at the store/week at which the purchase occasion t occurs. This rules out the possibility that consumers may substitute to another store/week if certain brands are not available at the current store/week. Since the potato chips are the so-called frequently-purchased-products and not expensive, this assumption seems sensible in this setting and likely to hold for most consumers in the sample. These assumptions on the consideration sets are also imposed in Gentry (2011).

To model the consideration probabilities, I assume that display and last time purchase affect the consideration of all the brands in a symmetric way. Also, for simplicity, I assume the consideration probability of j only depends on the display and purchase history of j, not other brands. Thus, consumer i's consideration probabilities of brand j at purchase occasion t, denoted as h_{ijt}, can be summarized by the following table,

	$d_{ij,t-1} = 0$	$d_{ij,t-1} = 1$
No Disp.	r_b	r_s
Minor Disp.	r_d	r_{ds}
Major Disp.	1	1

where $r_k \in [0, 1]$ $k \in \{b, s, d, ds\}$ are parameters to be estimated.

Given the specified random utility model and consideration probabilities, we can write

down the moment inequalities as

$$\mathbf{E}\left[d_{ijt} - \frac{h_{ijt}\cdot\exp\left[\delta_j+\beta_i p_{jt}+\varphi\cdot 1\left(d_{ij,t-1}=1\right)\right]}{1+\sum_{k\in C_{ijt}^{sup}}\exp\left[\delta_k+\beta_i p_{kt}+\varphi\cdot 1\left(d_{ik,t-1}=1\right)\right]}\middle| X_{it}, d_{i,t-1}\right]\geq 0$$
$$\mathbf{E}\left[\frac{h_{ijt}\cdot\exp\left[\delta_j+\beta_i p_{jt}+\varphi\cdot 1\left(d_{ij,t-1}=1\right)\right]}{1+\sum_{k\in C_{ijt}^{sub}}\exp\left[\delta_k+\beta_i p_{kt}+\varphi\cdot 1\left(d_{ik,t-1}=1\right)\right]} - d_{ijt}\middle| X_{it}, d_{i,t-1}\right]\geq 0\ \forall j\ a.s.\left[X_i, d_{i,t-1}\right], \tag{5.1}$$

where $d_{i,t-1}\equiv\left(d_{i1,t-1},...,d_{iJ,t-1}\right)$, X_{it} includes brand dummies, the prices and display variables of all available brands for i at t, household income and family size. The close-form expression of choice probabilities follows from the fact that, conditioning on last purchase, the error term ε is i.i.d. across t, see Chapter 6 of Train (2009) for more explanations.

I implement the estimator (3.3) based on the moment inequalities (5.1). The set of instrumental functions consists of brand dummies, last purchase $d_{i,t-1}$, indicator variables for different groups of households classified by income and family size, indicator variable for the store/week pair that each purchase happens, as well as some flexible interactions between them. Also, I use a uniform weight across all the instrumental functions. The estimation results are fairly robust to different choices of the instrumental functions and their weights.

5.3 Estimation Results and Counterfactual Experiments

For comparison, I include results from two alternative approaches: one is the standard random utility model with the assumption that each consumer considers all the brands available at the store/week. The other one takes the consideration set generation into account but assumes that the probabilities of consideration are independent across products.

The coefficients on the state-dependence utility term estimated using moment inequalities approaches are significantly smaller than those from alternative estimation strategies, which resembles the finding in Chiang, Chib, and Narasimhan (1999). This means a substantial part of the effect of last purchase on current choice is due to the fact that it makes the brands

purchased previously easier to be recalled and thus considered.

Also, comparing to the full consideration set model, both independent consideration and moment inequality models imply weaker price sensitivity: the independent consideration has smaller price coefficient; the moment inequality approach has larger coefficients on the interactions terms between price and demographics. This pattern confirms the findings in Goeree (2008) and Gentry (2011). The intuitive interpretation is that if consumers' true consideration sets are very small but researchers wrongly impose a large consideration set for them, then in the model each product is forced to compete with all the other products and thus appears to be more sensitive to price change.

Finally, assuming full consideration or independent consideration may overestimate the state dependence utility term. Also, note that the estimate of r_s and r_{ds} are 1 meaning if a brand was purchased last time by a consumer, then she will consider this brand on the current purchase trip with probability one. In addition, r_d is bigger than r_b shows that display increases the probability of consideration for a given product, which suggests display is useful in getting a product into consumers consideration sets.

Table 4: Estimation Results

	Full Consid.	Ind. Consid	Mom. Ineq.
Utility Param.			
Last Purch.	1.36	1.60	0.94
	(.01)	(.02)	(.01)
Price	-.60	-.38	-.60
	(.01)	(.01)	(.00)
Price×HHInc	.002	.002	.023
	(.00)	(.00)	(.00)
Price×HHSize	.017	.016	.035
	(.01)	(.01)	(.01)
Minor Disp.	.56	-	-
	(.02)		
Major Disp.	.45	-	-
	(.01)		
Consid. Prob.			
No Disp., No Last Purch.	-	.71	.25
		(.01)	(.00)
No Disp., Last Purch.	-	1.00	1.00
		(.02)	(.00)
Minor Disp. No Last Purch	-	.99	.75
		(.02)	(.00)
Minor Disp. Last Purch	-	.69	1.00
		(.01)	(.00)

Note: 1. The estimates of brand level effects are omitted.
2. The standard errors of the estimates from the moment inequality approach are obtained using a simple nonparametric bootstrap with 1,000 bootstrap samples of households.

The estimation results show that last time purchase plays a role in formation of the current consideration set. If we take this effect into account, the effect of last purchase on the utility function becomes smaller. In the following, I shall run a simple counterfactual experiment to quantify the contribution of consideration set formation, driven by the last purchase, to the persistence. Table 5 presents the results.

The first column, labeled as "Actual", contains the empirical re-purchase probabilities observed in the data. The second column, "Model Prediction", includes the predicted bounds on the re-purchase probabilities by the model evaluated at the point estimate. We can see that the model does a decent job in fitting the data as the predicted bounds cover the actual

values for most of the brands. For the results in "No Consid. Eff.", I remove the last purchase from consideration set formation and use the estimated model to predict re-purchase probabilities. The difference between these probabilities and the "Model Prediction" column provides a measure of how much the last purchase can affect the persistence through consideration set formation. The "Consid. Eff." measures this difference by taking the percentage change of the mid-points in "No Consid. Eff." with respect to those of "Model Prediction". So roughly 2% to 20% of the re-purchase probability, also known as purchase carry-over effect, are driven by that the last purchase of a brand increases its current consideration probability. Analogously, I compute effect of last purchase on the persistence through utility function and the results are in the last column labeled as "State Depend.". These results suggests that both effects of last purchase on utility and consideration probability are important, of which the former is slightly stronger.

Table 5: Counterfactual Re-purchase Probabilities (%)

	Actual	Model Prediction	No Consid. Eff.	No State Depend.	Consid. Eff.	State Depend.
Lays	48.01	[39.77, 57.81]	[30.94, 42.92]	[22.18, 38.04]	11.57	18.23
Wavy Lays	39.30	[28.46, 46.26]	[20.98, 31.38]	[15.15, 29.02]	8.35	11.41
Lays Natural	48.89	[31.58, 53.27]	[26.65, 43.18]	[16.41, 33.84]	6.37	14.68
Cape Cod	53.57	[33.86, 49.79]	[16.31, 23.39]	[17.25, 30.23]	18.38	15.13
Ruffles	24.84	[23.10, 39.80]	[14.18, 24.39]	[11.04, 22.64]	7.65	9.19
Old Dutch	25.96	[23.95, 42.86]	[15.14, 26.46]	[11.30, 24.57]	8.42	10.34
UTZ	56.14	[35.05, 51.20]	[16.77, 24.14]	[18.19, 31.54]	19.55	15.75
Lays Kettle Cooked	36.72	[14.34, 29.00]	[10.89, 21.01]	[6.34, 15.00]	2.48	4.77
Old Dutch Ripples	26.61	[23.83, 39.57]	[21.53, 35.18]	[11.23, 21.75]	2.12	9.65
Rachels	29.68	[17.38, 32.17]	[12.42, 21.00]	[7.81, 16.63]	4.00	6.22

As discussed in Mehta, Rajiv, and Srinivasan (2001), disentanglling the two effects that previous purchase on current decision, state-dependent utility versus attention inertia may have some managerial implications. For example, if a manager (of a company or retail store) would like to launch a new product and knows that the consumers' "loyalty" to a brand mostly comes from restricted attention, then she can display (or use other marketing strategies) the new product heavily with a shallow price discount because in this case the key

to break the "loyalty" is to attract consumers' attention instead of discounting to increase the utility level; on the contrary, if the "loyalty" mainly comes from state dependent utility, then she should moderately advertise the new product with deep price discount to make the product more competitive.

6 Concluding Remarks

In this paper, I present a simple moment inequality approach to estimating random utility model when consumers' consideration sets are unobserved and heterogeneous. Comparing to the various consideration set generation models in the literature, the moment inequality approach imposes weaker assumptions on the consideration set distribution and is very easy to implement. This method is applied to household panel and store scanner data to study the following empirical question: can attention inertia explain some of the observed persistence in consumer brand choice? The empirical results shows that a substantial amount of the persistence is driven by the effect of previous purchase on consumers' consideration set formation.

In the main text, I develop the moment inequality approach in the context of individual choice data. But the consideration set heterogeneity can also be a problem for aggregate data. In Appendix D, I consider the extensions of the current framework to handle aggregate data and price endogeneity.

Future research that applies the proposed approach to empirical applications to the markets with many choices, like books, hotels, air tickets, etc., may be fruitful. Since consumers in these markets usually do not consider all the available options, the moment inequality approach is useful in understanding and estimating the demand in these markets.

A Proof of Theorem 2

Proof. For any $\theta \neq \theta_0$, if

$$R_j^{sup}\left(X_i;\theta,\theta_0\right) > \Delta_j^{sup}\left(X_i;\theta_0\right),$$

then

$$\begin{aligned}
&\int \sigma_j\left(X_i,\varepsilon_i\left|C_j^{sup}\left(X_i\right);\lambda\right.\right) h_j\left(\varepsilon_i,X_i;\varphi\right) dF_\varepsilon\left(\varepsilon_i\left|X_i;\rho\right.\right) \\
=&R_j^{sup}\left(X_i;\theta,\theta_0\right) + \int \sigma_j\left(X_i,\varepsilon_i\left|C_j^{sup}\left(X_i\right);\lambda_0\right.\right) h_j\left(\varepsilon_i,X_i;\varphi_0\right) dF_\varepsilon\left(\varepsilon_i\left|X_i;\rho_0\right.\right) \\
=&R_j^{sup}\left(X_i;\theta,\theta_0\right) + \mathbf{E}\left[d_i\left|X_i\right.\right] \\
+&\left\{\int \sigma_j\left(X_i,\varepsilon_i\left|C_j^{sup}\left(X_i\right);\lambda_0\right.\right) h_j\left(\varepsilon_i,X_i;\varphi_0\right) dF_\varepsilon\left(\varepsilon_i\left|X_i;\rho_0\right.\right) - \mathbf{E}\left[d_i\left|X_i\right.\right]\right\} \\
=&R_j^{sup}\left(X_i;\theta,\theta_0\right) + \mathbf{E}\left[d_i\left|X_i\right.\right] \\
+&\int \sigma_j\left(X_i,\varepsilon_i\left|C_j^{sup}\left(X_i\right);\lambda_0\right.\right) h_j\left(\varepsilon_i,X_i;\varphi_0\right) dF_\varepsilon\left(\varepsilon_i\left|X_i;\rho_0\right.\right) \\
-&\int \sigma_j\left(X_i,\varepsilon_i\left|C_j^{sup}\left(X_i\right);\lambda_0\right.\right) P\left(C_j^{sup}\left(X_i\right)\left|\varepsilon_i,X_i\right.\right) dF_\varepsilon\left(\varepsilon_i\left|X_i;\rho_0\right.\right) \\
-&\int \sum_{C\in C_j(X_i)\backslash C_j^{sup}(X_i)} \sigma_j\left(X_i,\varepsilon_i\left|C;\lambda_0\right.\right) P\left(C\left|\varepsilon_i,X_i\right.\right) dF_\varepsilon\left(\varepsilon_i\left|X_i;\rho_0\right.\right) \\
\geq&R_j^{sup}\left(X_i;\theta,\theta_0\right) + \mathbf{E}\left[d_i\left|X_i\right.\right] \\
+&\int \sigma_j\left(X_i,\varepsilon_i\left|C_j^{sup}\left(X_i\right);\lambda_0\right.\right)\left[h_j\left(\varepsilon_i,X_i;\varphi_0\right) - P\left(C_j^{sup}\left(X_i\right)\left|\varepsilon_i,X_i\right.\right)\right] dF_\varepsilon\left(\varepsilon_i\left|X_i;\rho_0\right.\right) \\
-&\int \sigma_j\left(X_i,\varepsilon_i\left|C_j^{sub}\left(X_i\right);\lambda_0\right.\right)\left[h_j\left(\varepsilon_i,X_i;\varphi_0\right) - P\left(C_j^{sup}\left(X_i\right)\left|\varepsilon_i,X_i\right.\right)\right] dF_\varepsilon\left(\varepsilon_i\left|X_i;\rho_0\right.\right) \\
=&R_j^{sup}\left(X_i;\theta,\theta_0\right) + \mathbf{E}\left[d_i\left|X_i\right.\right] \\
+&\int \left[\sigma_j\left(X_i,\varepsilon_i\left|C_j^{sup}\left(X_i\right);\lambda_0\right.\right) - \sigma_j\left(X_i,\varepsilon_i\left|C_j^{sub}\left(X_i\right);\lambda_0\right.\right)\right] \\
&\left[h_j\left(\varepsilon_i,X_i;\varphi_0\right) - P\left(C_j^{sup}\left(X_i\right)\left|\varepsilon_i,X_i\right.\right)\right] dF_\varepsilon\left(\varepsilon_i\left|X_i;\rho_0\right.\right) \\
\geq&R_j^{sup}\left(X_i;\theta,\theta_0\right) + \mathbf{E}\left[d_i\left|X_i\right.\right] - \Delta_j^{sup}\left(X_i;\theta_0\right) \\
>&\mathbf{E}\left[d_i\left|X_i\right.\right].
\end{aligned}$$

Analogously, if

$$R_j^{sup}\left(X_i;\theta,\theta_0\right) < -\Delta_j^{sup}\left(X_i;\theta_0\right),$$

then

$$\begin{aligned}
&\int \sigma_j\left(X_i,\varepsilon_i\left|C_j^{sub}\left(X_i\right);\lambda\right.\right) h_j\left(\varepsilon_i,X_i;\varphi\right) dF_\varepsilon\left(\varepsilon_i\left|X_i;\rho\right.\right)\\
=&R_j^{sub}\left(X_i;\theta,\theta_0\right) + \int \sigma_j\left(X_i,\varepsilon_i\left|C_j^{sub}\left(X_i\right);\lambda_0\right.\right) h_j\left(\varepsilon_i,X_i;\varphi_0\right) dF_\varepsilon\left(\varepsilon_i\left|X_i;\rho_0\right.\right)\\
=&R_j^{sub}\left(X_i;\theta,\theta_0\right) + \mathbf{E}\left[d_i\left|X_i\right.\right]\\
+&\left\{\int \sigma_j\left(X_i,\varepsilon_i\left|C_j^{sub}\left(X_i\right);\lambda_0\right.\right) h_j\left(\varepsilon_i,X_i;\varphi_0\right) dF_\varepsilon\left(\varepsilon_i\left|X_i;\rho_0\right.\right) - \mathbf{E}\left[d_i\left|X_i\right.\right]\right\}\\
=&R_j^{sub}\left(X_i;\theta,\theta_0\right) + \mathbf{E}\left[d_i\left|X_i\right.\right]\\
+&\int \sigma_j\left(X_i,\varepsilon_i\left|C_j^{sub}\left(X_i\right);\lambda_0\right.\right) h_j\left(\varepsilon_i,X_i;\varphi_0\right) dF_\varepsilon\left(\varepsilon_i\left|X_i;\rho_0\right.\right)\\
-&\int \sigma_j\left(X_i,\varepsilon_i\left|C_j^{sub}\left(X_i\right);\lambda_0\right.\right) P\left(C_j^{sub}\left(X_i\right)\left|\varepsilon_i,X_i\right.\right) dF_\varepsilon\left(\varepsilon_i\left|X_i;\rho_0\right.\right)\\
-&\int \sum_{C\in C_j(X_i)\setminus C_j^{sub}(X_i)} \sigma_j\left(X_i,\varepsilon_i\left|C;\lambda_0\right.\right) P\left(C\left|\varepsilon_i,X_i\right.\right) dF_\varepsilon\left(\varepsilon_i\left|X_i;\rho_0\right.\right)\\
\leq&R_j^{sub}\left(X_i;\theta,\theta_0\right) + \mathbf{E}\left[d_i\left|X_i\right.\right]\\
+&\int \sigma_j\left(X_i,\varepsilon_i\left|C_j^{sub}\left(X_i\right);\lambda_0\right.\right) \left[h_j\left(\varepsilon_i,X_i;\varphi_0\right) - P\left(C_j^{sub}\left(X_i\right)\left|\varepsilon_i,X_i\right.\right)\right] dF_\varepsilon\left(\varepsilon_i\left|X_i;\rho_0\right.\right)\\
-&\int \sigma_j\left(X_i,\varepsilon_i\left|C_j^{sup}\left(X_i\right);\lambda_0\right.\right) \left[h_j\left(\varepsilon_i,X_i;\varphi_0\right) - P\left(C_j^{sub}\left(X_i\right)\left|\varepsilon_i,X_i\right.\right)\right] dF_\varepsilon\left(\varepsilon_i\left|X_i;\rho_0\right.\right)\\
=&R_j^{sub}\left(X_i;\theta,\theta_0\right) + \mathbf{E}\left[d_i\left|X_i\right.\right]\\
+&\int \left[\sigma_j\left(X_i,\varepsilon_i\left|C_j^{sub}\left(X_i\right);\lambda_0\right.\right) - \sigma_j\left(X_i,\varepsilon_i\left|C_j^{sup}\left(X_i\right);\lambda_0\right.\right)\right]\\
&\left[h_j\left(\varepsilon_i,X_i;\varphi_0\right) - P\left(C_j^{sub}\left(X_i\right)\left|\varepsilon_i,X_i\right.\right)\right] dF_\varepsilon\left(\varepsilon_i\left|X_i;\rho_0\right.\right)\\
\leq&R_j^{sub}\left(X_i;\theta,\theta_0\right) + \mathbf{E}\left[d_i\left|X_i\right.\right] + \Delta_j^{sub}\left(X_i;\theta_0\right)\\
<&\mathbf{E}\left[d_i\left|X_i\right.\right].
\end{aligned}$$

□

B Proof of Consistency

In the following, I present a consistency theorem that allows for unknown functions in the generic moment inequalities model

$$\mathbf{E}\left[m\left(W_i;\theta,\tau\right)|\,X_i\right]\geq 0\;\forall\tau\in\mathcal{T}\quad a.s.\left[X_i\right], \tag{B.1}$$

where $\theta\left(\equiv(\lambda,h)\right)\in\Theta\left(\equiv\Lambda\times\prod_{j=1}^{J}\mathcal{H}_j\right)$ and $\mathcal{T}$ is an index set that can be finite or infinite. This general setting is useful when empirical researchers would like to employ a nonparameteric specification for the consideration probabilities μ_{ij}. Suppose θ includes nonparametric components $h\left(\cdot\right)\equiv\left(h_1\left(\cdot\right),...,h_J\left(\cdot\right)\right)$ and I shall use sieve method to approximate the unknown functions h. To keep things simple, I restrict the attention to the smooth class of functions as the parameter space for h. Specifically, define a Hölder ball with smoothness α as

$$\mathbf{H}_M^{\alpha}\left(\mathcal{X}\right)=\left\{f\in C^k\left(\mathcal{X}\right):\sup_{k\leq\underline{\alpha}}\sup_{X\in\mathcal{X}}\left|D^k f\left(X\right)\right|\leq M,\sup_{k=\underline{\alpha}}\sup_{X\neq X'}\frac{\left|D^k f\left(X\right)-D^k f\left(X'\right)\right|}{\left\|X-X'\right\|_E^{\alpha-\underline{\alpha}}}\leq M\right\},$$

where $\underline{\alpha}$ is the greatest integer smaller than α, $\left\|\cdot\right\|_E$ is the Euclidean norm, $q\in(0,1]$, $\mathcal{X}$ denotes the support of X_i, $C^k\left(\mathcal{X}\right)$ is the space of all m-times continuously differentiable real-valued functions on $\mathcal{X}$, $k\equiv k_1+k_2+...+k_{d_X}$ and $D^{\alpha}\equiv\frac{\partial^{[\alpha]}}{\partial X_1^{\alpha_1}...\partial X_{d_X}^{\alpha d_X}}$ denotes the differential operator. Define a metric on Θ, $d\left(\theta,\theta'\right)=\left\|\lambda-\lambda'\right\|_E+\sum_{j=1}^{J}\left\|h_j-h_j'\right\|_H$, where $\left\|\cdot\right\|_{\mathcal{H}}$ denotes the usual sup-norm or L_2-norm.

B.1 Assumptions

Assumption 1. *The data* $\{W_i\}_{i=1}^n$ *is i.i.d.*

Assumption 2. *The parameter space of the finite dimensional parameters is compact and* $\mathcal{H}_j=\mathbf{H}_M^{\alpha}\left(\mathcal{X}\right)$ *for all* $j=1,...,J$.

Assumption 3. $\mathbf{E}\left[\sup_{\theta\in\Theta}\sup_{(\tau,g)\in\mathcal{T}\times\mathcal{G}}\left|m\left(W_i;\theta,\tau,g\right)\right|\right]<\infty$ *and* $\mathbf{Var}\left[\sup_{(\tau,g)\in\mathcal{T}\times\mathcal{G}}\left|m\left(W_i;\theta_0,\tau,g\right)\right|\right]<\infty$.

Assumption 4. *There exist a constant* $c>0$ *and a random variable* $U_1\left(W_i\right)$ *satisfying* $\mathbf{E}\left[U_1\left(W_i\right)\right]^2<\infty$ *such that for any* θ,$\theta'\in\Theta$,

$$\sup_{(\tau,g)\in\mathcal{T}\times\mathcal{G}}\left|m\left(W_i;\theta,\tau,g\right)-m\left(W_i;\theta',\tau,g\right)\right|\leq U_1\left(W_i\right)\cdot\left[d\left(\theta,\theta'\right)\right]^c.$$

Assumption 5. *For any* $\theta\in\Theta$, *there exists* $\Pi_n\theta\in\Theta_n$ *such that* $d\left(\Pi_n\theta,\theta\right)=o\left(1\right)$;

Remark 1. Assumption 1 rules out interactions among consumers but is typically imposed in the discrete choice literature. Assumption 2 implies that Θ is compact under $d\left(\cdot,\cdot\right)$, which is a common requirement in the non-/semi-parametric literature, see, e.g., Newey and Powell (2003); Ai and Chen (2003). Allowing for non-compact parameter space is technically more involved (see, e.g., Chen and Pouzo (2012)) and seems not necessary for currently application. Assumption 3 and 4 are some basic requirements on moment functions and are typically imposed even in the nonlinear parametric estimation literature. Assumption 5 simply defines the sieve space, see, e.g., Newey and Powell (2003); Ai and Chen (2003); Chen (2007) for more explanations.

Theorem 3 (Consistency)**.** *Suppose the conditions in Theorem* (2) *hold. If Assumption 1, 2, 3, 4, 5 hold, then* $d\left(\hat{\theta},\theta_0\right)=o_p\left(1\right)$.

B.2 Proof of Theorem (3)

Let $Q\left(\theta\right)=S\left[m\left(\theta\right)\right]$, where $m\left(\theta\right)=\left\{\mathbf{E}\left[m\left(W_i;\theta,\tau,g\right)\right]:(\tau,g)\in\mathcal{T}\times\mathcal{G}\right\}$. Let $Q_n\left(\theta\right)=S\left[m_n\left(\theta\right)\right]$ and thus the estimator can be re-written as $\hat{\theta}\equiv\arg\min_{\theta\in\Theta_n}Q_n\left(\theta\right)$. I first present a proposition that will be used to show consistency. This proposition combines and rephrases the Remark 3.1 (4) in Chen (2007) and Lemma A1 of Newey and Powell (2003).

Proposition 2. *Suppose the following conditions hold: (i) Θ is compact under $d(\cdot,\cdot)$; (ii) $Q(\theta)$ is lower semicontinuous on Θ under $d(\cdot,\cdot)$; (iii) $Q(\theta)$ has a unique minimum on Θ at $\theta_0 \in \Theta$ and $Q(\theta_0) < \infty$; (iv) the sieve spaces $\{\Theta_n : n \geq 1\}$ are compact subsets of Θ under $d(\cdot,\cdot)$ and for each $\theta \in \Theta$, there exists $\Pi_n\theta \in \Theta_n$ such that $d(\Pi_n\theta, \theta) = o(1)$; (v) $\sup_{\theta\in\Theta} |Q_n(\theta) - Q(\theta)| = o_p(1)$. Then $d\left(\hat{\theta}, \theta_0\right) = o_p(1)$.*

The following lemma shows (v) of Proposition 2, i.e., uniform convergence of $Q_n(\cdot)$.

Lemma 1. *If Assumptions 1, 2, 3, 4 hold, then $\sup_{\theta\in\Theta} |Q_n(\theta) - Q_F(\theta)| = o_p(1)$.*

Proof. For any $\delta > 0$, since Θ is compact by Assumption 2, we can construct a finite cover of Θ, $B(\theta^i, \delta) = \{\theta \in \Theta : d(\theta, \theta^i) \leq \delta\}$ $i = 1, ..., N_\delta$. For each $B(\theta^i, \delta)$, we have

$$
\begin{aligned}
&\sup_{\theta\in B(\theta^i,\delta)} |Q_n(\theta) - Q_F(\theta)| \\
&\leq \sup_{\theta\in B(\theta^i,\delta)} \sup_{(\tau,g)\in\mathcal{T}\times\mathcal{G}} \left| [m_n(\theta,\tau,g)]_-^2 - [m_F(\theta,\tau,g)]_-^2 \right| \\
&\leq \sup_{\theta\in B(\theta^i,\delta)} \sup_{(\tau,g)\in\mathcal{T}\times\mathcal{G}} \left\{ |m_n(\theta,\tau,g) - m_F(\theta,\tau,g)|^2 + 2|m_F(\theta,\tau,g)|\,|m_n(\theta,\tau,g) - m_F(\theta,\tau,g)| \right\} \\
&\leq \left[\sup_{\theta\in B(\theta^i,\delta)} \sup_{(\tau,g)\in\mathcal{T}\times\mathcal{G}} |m_n(\theta,\tau,g) - m_F(\theta,\tau,g)| \right]^2 \\
&+2 \left[\sup_{\theta\in B(\theta^i,\delta)} \sup_{(\tau,g)\in\mathcal{T}\times\mathcal{G}} |m_F(\theta,\tau,g)| \right] \left[\sup_{\theta\in B(\theta^i,\delta)} \sup_{(\tau,g)\in\mathcal{T}\times\mathcal{G}} |m_n(\theta,\tau,g) - m_F(\theta,\tau,g)| \right].
\end{aligned}
$$

Next, note that

$$
\begin{aligned}
&\sup_{\theta\in B(\theta^i,\delta)}\sup_{(\tau,g)\in\mathcal{T}\times\mathcal{G}}\left|m_n(\theta,\tau,g)-m_F(\theta,\tau,g)\right| \\
\leq&\sup_{\theta\in B(\theta^i,\delta)}\sup_{(\tau,g)\in\mathcal{T}\times\mathcal{G}}\left|m_n(\theta,\tau,g)-m_n\left(\theta^i,\tau,g\right)\right|+\sup_{(\tau,g)\in\mathcal{T}\times\mathcal{G}}\left|m_n\left(\theta^i,\tau,g\right)-m_F\left(\theta^i,\tau,g\right)\right| \\
+&\sup_{\theta\in B(\theta^i,\delta)}\sup_{(\tau,g)\in\mathcal{T}\times\mathcal{G}}\left|m_F\left(\theta^i,\tau,g\right)-m_F(\theta,\tau,g)\right| \\
\leq&\frac{1}{n}\sum_{i=1}^{n}\sup_{\theta\in B(\theta^i,\delta)}\sup_{(\tau,g)\in\mathcal{T}\times\mathcal{G}}\left|m(W_i;\theta,\tau,g)-m\left(W_i;\theta^i,\tau,g\right)\right| \\
+&\sup_{(\tau,g)\in\mathcal{T}\times\mathcal{G}}\left|m_n\left(\theta^i,\tau,g\right)-m_F\left(\theta^i,\tau,g\right)\right|+C_\delta\cdot\delta^c \\
\leq&\frac{1}{n}\sum_{i=1}^{n}U_1(W_i)\cdot\delta^c+\sup_{(\tau,g)\in\mathcal{T}\times\mathcal{G}}\left|m_n\left(\theta^i,\tau,g\right)-m_F\left(\theta^i,\tau,g\right)\right|+C_\delta\cdot\delta^c \\
\leq&\delta^c\left|\frac{1}{n}\sum_{i=1}^{n}U(W_i)-\mathbf{E}_F\left[U(W_i)\right]\right|+C'+\sup_{(\tau,g)\in\mathcal{T}\times\mathcal{G}}\left|m_n\left(\theta^i,\tau,g\right)-m_F\left(\theta^i,\tau,g\right)\right|+C_\delta\cdot\delta^c \\
=&o_p(1)+C_\delta\cdot\delta^c+C'
\end{aligned}
$$

for some finite $C', C_\delta > 0$, where the second and third inequalities follow by Assumption 4, the last equality follows from the i.i.d. assumption and Assumption 3. Also, by Assumption 3,

$$
\sup_{\theta\in B(\theta^i,\delta)}\sup_{(\tau,g)\in\mathcal{T}\times\mathcal{G}}\left|m_F(\theta,\tau,g)\right|\leq\mathbf{E}_F\left[\sup_{\theta\in\Theta}\sup_{(\tau,g)\in\mathcal{T}\times\mathcal{G}}\left|m(W_i;\theta,\tau,g)\right|\right]<M_\delta
$$

for some finite $M_\delta > 0$. Hence,

$$
\begin{aligned}
\sup_{\theta\in\Theta}\left|Q_n(\theta)-Q_F(\theta)\right| &\leq \max_i\sup_{\theta\in B(\theta^i,\delta)}\left|Q_n(\theta)-Q_F(\theta)\right| \\
&= \left(C_\delta\delta^c+C'\right)^2+2M_\delta\left(C_\delta\delta^c+C'\right)+o_p(1) \\
&\equiv R(\delta)+o_p(1).
\end{aligned}
$$

Finally, for any $\epsilon > 0$, there exists $\delta \equiv R^{-1}(\epsilon) > 0$ such that

$$\sup_{\theta \in \Theta} |Q_n(\theta) - Q_F(\theta)| \leq \epsilon + o_p(1).$$

The conclusion follows by noting that ϵ is arbitrary. □

Proof of Theorem 3

The proof amounts to verifying the conditions of Lemma 2. Condition (i) is implied by Assumption 2. To show condition (ii), observe that

$$\begin{aligned}
&\left|Q_F(\theta) - Q_F\left(\theta'\right)\right| \\
\leq &\sup_{(\tau,g)\in\mathcal{T}\times\mathcal{G}} \left|[m_F(\theta,\tau,g)]_-^2 - \left[m_F\left(\theta',\tau,g\right)\right]_-^2\right| \\
\leq &\sup_{(\tau,g)\in\mathcal{T}\times\mathcal{G}} \left\{\left|m_F(\theta,\tau,g) - m_F\left(\theta',\tau,g\right)\right|^2 + 2\left|m_F(\theta,\tau,g)\right|\left|m_F(\theta,\tau,g) - m_F\left(\theta',\tau,g\right)\right|\right\} \\
\leq &\left\{\sup_{(\tau,g)\in\mathcal{T}\times\mathcal{G}} \left|m_F(\theta,\tau,g) - m_F\left(\theta',\tau,g\right)\right| + 2\left[\sup_{(\tau,g)\in\mathcal{T}\times\mathcal{G}} |m_F(\theta,\tau,g)|\right]\right\} \\
&\cdot \left[\sup_{(\tau,g)\in\mathcal{T}\times\mathcal{G}} \left|m_F(\theta,\tau,g) - m_F\left(\theta',\tau,g\right)\right|\right] \\
\leq &C \cdot \left\{C\left[d\left(\theta,\theta'\right)\right]^c + 2C\left[d(\theta,\theta_0)\right]^c + 2\sup_{(\tau,g)\in\mathcal{T}\times\mathcal{G}} |m_F(\theta_0,\tau,g)|\right\}\left[d\left(\theta,\theta'\right)\right]^c \\
\leq &B \cdot \left[d\left(\theta,\theta'\right)\right]^c,
\end{aligned}$$

where the second inequality follows from the fact the for all $m_1, m_2 \in \mathbb{R}$,

$$\left|[m_1]_-^2 - [m_2]_-^2\right| = \left|[m_1]_- + [m_2]_-\right|\left|[m_1]_- - [m_2]_-\right| \leq \left(|m_1 - m_2| + 2|m_2|\right)|m_1 - m_2|,$$

the fourth inequality holds by Assumption 4, the last inequality is a consequence of compactness of Θ and Assumption 3. Condition (iii) is guaranteed by Theorem 2 and definition of $Q_F(\theta)$. Condition (iv) is a consequence of Assumption 5. Finally, Lemma 1 shows condition

(v).

C Sharp Identified Set

Given the set of possible consideration sets $\mathcal{C}(X_i)$ and parametrized consideration probability (2.11), the consideration set distribution satisfies the following constraints

$$\sum_{\mathcal{I}\in\mathcal{C}(X_i)} P(\mathcal{I}\,|\varepsilon_i, X_i) = h_j(\varepsilon_i, X_i; \varphi_0) \;\; \forall j = 1, ..., J, \tag{C.1}$$

$$P(\mathcal{I}\,|\varepsilon_i, X_i) \geq 0 \;\forall \mathcal{I} \in \mathcal{C}(X_i), \tag{C.2}$$

$$\sum_{\mathcal{I}\in\mathcal{C}(X_i)} P(\mathcal{I}\,|\varepsilon_i, X_i) = 1, \tag{C.3}$$

where constraint (C.1) follows from the definition of the consideration probability $h_j(\varepsilon_i, X_i; \varphi_0)$, (C.2) and (C.3) are the basic requirements of a well-defined probability distribution.

These constraints define a set of consideration set distributions. Using matrix notation, we can write the set as

$$\mathbf{Y}(\varepsilon_i, X_i; \varphi_0) = \left\{ y \in \triangle^{K^*-1} : \Omega \cdot y = h(\varepsilon_i, X_i; \varphi_0) \right\}, \tag{C.4}$$

where $K^*$is the dimension of $\mathcal{C}(X_i)$, $\triangle^{K^*-1}$ is the standard K^*-simplex, $h(\cdot) = (h_1(\cdot), ..., h_J(\cdot))$, and

$$\Omega \equiv \begin{bmatrix} 1(1 \in \mathcal{I}_1) & ... & 1(1 \in \mathcal{I}_{K^*}) \\ \vdots & \ddots & \vdots \\ 1(J \in \mathcal{I}_1) & ... & 1(J \in \mathcal{I}_{K^*}) \end{bmatrix}$$

is a membership matrix with $[\Omega]_{jk}$ indicating whether product j belongs to consideration set k.

The next step is to combine (C.4) with the random utility model (2.1) to derive the

implications for the choice probabilities. For a given value of (λ, ρ, φ) and (X_i, ε_i), the set of consideration set distributions can be translated to a set of choice probability vectors, i.e.

$$\Lambda\left(X_i, \varepsilon_i; \lambda, \varphi\right) = \left\{\Sigma\left(X_i, \varepsilon_i; \lambda\right) \cdot y : y \in \mathbf{Y}\left(\varepsilon_i, X_i; \varphi\right)\right\},$$

where

$$\Sigma\left(X_i, \varepsilon_i; \lambda\right) = \begin{bmatrix} \sigma_1\left(X_i, \varepsilon_i \left| \mathcal{I}_1; \lambda\right.\right) & \cdots & \sigma_1\left(X_i, \varepsilon_i \left| \mathcal{I}_K; \lambda\right.\right) \\ \vdots & \ddots & \vdots \\ \sigma_J\left(X_i, \varepsilon_i \left| \mathcal{I}_1; \lambda\right.\right) & \ldots & \sigma_J\left(X_i, \varepsilon_i \left| \mathcal{I}_K; \lambda\right.\right) \end{bmatrix}.$$

Then, following Beresteanu and Molinari (2008); Beresteanu, Molchanov, and Molinari (2011), we can apply the notion of Aumann expectation to derive the empirical implication of the predicted set of choice probabilities, i.e.,

$$\mathbf{E}\left[d_i \left| X_i\right.\right] \in \mathbb{E}_\varepsilon\left[\Lambda\left(X_i, \varepsilon_i; \lambda_0, \varphi_0\right) \left| X_i\right.\right] \ a.s.\left[X_i\right], \tag{C.5}$$

where $\mathbb{E}_\varepsilon\left[\cdot \left| X_i\right.\right]$ is the conditional Aumann expectation taken with respect to ε using $F_\varepsilon\left(\varepsilon_i \left| X_i; \rho_0\right.\right)$. Since the choice probability vector $\mathbf{E}\left[d \left| X\right.\right]$ can be recovered from data, the sharp identified set of $(\lambda_0, \rho_0, \varphi_0)$ is made up of all the parameter values that satisfy (C.5).

D Extensions: Price Endogeneity and Market Level Data

D.1 Price Endogeneity

To deal with the potential price endogeneity issue that stems from the possible correlation between unobserved product-level characteristic and price, I follow the idea of Berry,

Levinsohn, and Pakes (2004); Goolsbee and Petrin (2004) to include product dummies in X and obtain estimates of their coefficients, denoted as $\delta_j \ \forall j$ using the approach proposed in this paper, and then decompose the δ as $\delta_j = x_j\beta_0 + \xi_j$, where x_j is a vector of observable characteristics (typically including price) and $\xi_j \in \mathbb{R}$ denotes the unobservable (to the econometricians) attribute. With a mean independence assumption $\mathbf{E}\left[\xi_j \left| z_j \right.\right] = 0$, where z_j is a set of instruments for price, we can estimate β_0 using standard IV regression based on

$$\mathbf{E}\left[\delta_j - x_j\beta_0 \left| z_j \right.\right] = 0 \quad \forall j = 1, ..., J. \tag{D.1}$$

To obtain precise estimate of β_0, we typically need market level variations, i.e., data from multiple markets, see Berry, Levinsohn, and Pakes (2004) for detailed discussions on this issue.

D.2 Market Level Data

Suppose we only have data on aggregate consumer choices, i.e., market shares $s = (s_0, s_1, ..., s_J)$, and product-level characteristics x_j. We can still apply the idea in this paper to estimate demand with an incomplete model of choice set generation. The key is to apply the moment restriction D.1, which in turn relies on inverting the demand system to get δ's. With the incomplete model, we can not uniquely recover δ's but can still characterize the set of δ's that are consistent with data.

For simplicity, consider a simple random utility model without random coefficients

$$u_{ij} = \delta_j + \varepsilon_{ij},$$

where $\delta_j = x_j\beta_0 + \xi_j$ and ε_{ij} follows a known distribution. Following the procedure in Section C, we can obtain the set of predicted market shares as $\Lambda\left(\delta; h\right)$ for given (δ, h). Thus, given h, there is a set δ satisfying $s \in \Lambda\left(\delta; \lambda, h\right)$, following Beresteanu, Molchanov, and Molinari

(2011), this set can be characterized by the equation

$$\max_{b \in \mathbb{B}^J} \left(b^{\mathrm{T}} s - \sup_{y \in \Lambda(\delta;h)} b^{\mathrm{T}} y \right) = 0, \tag{D.2}$$

where $\mathbb{B}^J$ is the J-dimensional unit ball. If we use the approach in Subsection 2.2, we can get bounds on the market shares

$$\sigma_{j,L}(\delta;h) \leq s_j \leq \sigma_{j,U}(\delta;h), \tag{D.3}$$

where $\sigma_{j,r}(\delta;h) = \sigma_{j|\mathcal{I}^{super}(j)}(\delta) h_j(x) \quad r = U, L$. By the inverse isotone property of $\sigma_{j,L}(\delta;\lambda,h)$ and $\sigma_{j,U}(\delta;\lambda,h)$ (see Berry, Gandhi, and Haile, 2012), we can invert the bounds in (D.3) to obtain bounds on δ, i.e.,

$$\sigma_{j,U}^{-1}(s;h) \leq \delta_j \leq \sigma_{j,L}^{-1}(s;h) \quad j = 1, ..., J. \tag{D.4}$$

With restrictions (D.2) or (D.4) on δ's, we can define an estimator based on following constrained optimization problem

$$\min_{\delta,\beta,h} Q(\delta, \beta, h)$$
$$s.t. \text{(D.2) or (D.4)},$$

where $Q(\cdot)$ is a GMM objective function constructed from the moment conditions (D.1).

References

Ai, C., and X. Chen (2003): "Efficient Estimation of Models with Conditional Moment Restrictions Containing Unknown Functions," *Econometrica*, 71(6), 1795–1843.

Andrews, D. W. K., and X. Shi (2013): "Inference Based on Conditional Moment

Inequalities," *Econometrica*, 81, 609–666.

Andrews, R. L., and T. C. Srinivasan (1995): "Studying Consideration Effects in Empirical Choice Models Using Scanner Panel Data," *Journal of Marketing Research*, 32(1), 30–41.

Ben-Akiva, M., and B. Boccara (1995): "Discrete choice models with latent choice sets," *International Journal of Research in Marketing*, 12, 9–24.

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

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

Berry, S., A. Gandhi, and P. Haile (2012): "Connected Substitutes and Invertibility of Demand," Discussion paper, National Bureau of Economic Research.

Berry, S., J. Levinsohn, and A. Pakes (1995): "Automobile prices in market equilibrium," *Econometrica: Journal of the Econometric Society*, pp. 841–890.

Berry, S., J. Levinsohn, and A. Pakes (2004): "Differentiated Products Demand Systems from a Combination of Micro and Macro Data: The New Vehicle Market," *Journal of Political Economy*, 112, 68–104.

Bronnenberg, B. J., M. W. Kruger, and C. F. Mela (2008): "Database paper-The IRI marketing data set," *Marketing Science*, 27(4), 745–748.

Bruno, H., and N. Vilcassim (2008): "Structural demand estimation with varying product availability," *Marketing Science*, 27, 1126–1131.

Carlson, J. A., and R. P. McAfee (1983): "Discrete equilibrium price dispersion," *The Journal of Political Economy*, pp. 480–493.

Chen, X. (2007): "Large Sample Sieve Estimation of Semi-Nonparametric Models," in *Handbook of Econometrics*, ed. by J. Heckman, and E. Leamer, vol. 6 of *Handbook of Econometrics*, chap. 76. Elsevier.

Chen, X., and D. Pouzo (2012): "Estimation of Nonparametric Conditional Moment Models With Possibly Nonsmooth Generalized Residuals," *Econometrica*, 80(1), 277–321.

Chernoff, H. (1954): "Rational selection of decision functions," *Econometrica: journal of the Econometric Society*, pp. 422–443.

Chiang, J., S. Chib, and C. Narasimhan (1999): "Markov Chain Monte Carlo and Models of Consideration Set and Parameter Heterogeneity," *Journal of Econometrics*, 89, 223–248.

Conlon, C. T., and J. H. Mortimer (2008): "Demand estimation under incomplete product availability," Discussion paper, National Bureau of Economic Research.

Draganska, M., and D. Klapper (2010): "Choice Set Heterogeneity and the Role of Advertising: An Analysis with Micro and Macro Data," Working Paper.

Dubé, J.-P., G. J. Hitsch, and P. E. Rossi (2010): "State dependence and alternative explanations for consumer inertia," *The RAND Journal of Economics*, 41(3), 417–445.

Eliaz, K., and R. Spiegler (2011): "Consideration sets and competitive marketing," *The Review of Economic Studies*, 78(1), 235–262.

Erdem, T. (1996): "A dynamic analysis of market structure based on panel data," *Marketing Science*, 15(4), 359–378.

Gandhi, A., Z. Lu, and X. Shi (2013): "Estimating Demand for Differentiated Products with Error in Market Shares," .

Gentry, M. (2011): "Displays, Sales, and In-Store Search in Retail Markets," Discussion paper, Working Paper.

Gilbride, T. J., and G. M. Allenby (2004): "A choice model with conjunctive, disjunctive, and compensatory screening rules," *Marketing Science*, 23(3), 391–406.

——— (2006): "Estimating heterogeneous EBA and economic screening rule choice models," *Marketing Science*, 25(5), 494–509.

Goeree, M. S. (2008): "Limited Information and Advertising in the U.S. Personal Computer Industry.," *Econometrica*, 76, 1017–1074.

Goolsbee, A., and A. Petrin (2004): "The consumer gains from direct broadcast satellites and the competition with cable TV," *Econometrica*, 72(2), 351–381.

Guadagni, P. M., and J. D. Little (1983): "A logit model of brand choice calibrated on scanner data," *Marketing science*, 2(3), 203–238.

Hauser, J. R., O. Toubia, T. Evgeniou, R. Befurt, and D. Dzyabura (2010): "Disjunctions of conjunctions, cognitive simplicity, and consideration sets," *Journal of Marketing Research*, 47(3), 485–496.

Hauser, J. R., and B. Wernerfelt (1990): "An evaluation cost model of consideration sets," *Journal of consumer research*, pp. 393–408.

Hortacsu, A., and C. Syverson (2004): "Product Differentiation, Search Costs, and Competition in the Mutual Fund Industry: A Case Study of S&P 500 Index Funds," *The Quarterly Journal of Economics*, 119(2), pp. 403–456.

Jedidi, K., and R. Kohli (2005): "Probabilistic subset-conjunctive models for heterogeneous consumers," *Journal of Marketing Research*, 42(4), 483–494.

Kahn, S., and E. Tamer (2009): "Inference on Randomly Censored Regression Models Using Conditional Moment Inequalities," *Journal of Econometrics*, 152, 104–119.

Kardes, F. R., G. Kalyanaram, M. Chandrashekaran, and R. J. Dornoff (1993): "Brand retrieval, consideration set composition, consumer choice, and the pioneering advantage," *Journal of Consumer Research*, pp. 62–75.

Keane, M. P. (1997): "Modeling heterogeneity and state dependence in consumer choice behavior," *Journal of Business & Economic Statistics*, 15(3), 310–327.

——— (2013): "Panel data discrete choice models of consumer demand," Discussion paper, Economics Group, Nuffield College, University of Oxford.

Liu, Q., and N. Arora (2011): "Efficient choice designs for a consider-then-choose model," *Marketing Science*, 30(2), 321–338.

Manski, C. F. (1977): "The structure of random utility models," *Theory and Decision*, 8, 229–254.

Manzini, P., and M. Mariotti (2014): "Stochastic Choice and Consideration Sets," *Econometrica*, 82(3), 1153–1176.

Marschak, J. (1960): "Binary-choice constraints and random utility indicators," in *Proceedings of a Symposium on Mathematical Methods in the Social Sciences*, vol. 7, pp. 19–38.

Masatlioglu, Y., D. Nakajima, and E. Y. Ozbay (2012): "Revealed attention," *The American Economic Review*, 102(5), 2183–2205.

McFadden, D. L. (1974): *Frontiers in Econometrics*chap. 4 Conditional Logit Analysis of Qualitative Choice Behavior, pp. 105–142. Academic Press: New York.

Mehta, N., S. Rajiv, and K. Srinivasan (2001): "Active versus passive loyalty: A structural model of consideration set formation," *Review of Marketing Science*.

——— (2003): "Price uncertainty and consumer search: A structural model of consideration set formation," *Marketing science*, 22(1), 58–84.

Moraga-González, J. L., Z. Sándor, and M. R. Wildenbeest (2009): "Consumer search and prices in the automobile market," .

Nedungadi, P. (1990): "Recall and consumer consideration sets: Influencing choice without altering brand evaluations," *Journal of consumer research*, pp. 263–276.

Newey, W. K., and J. L. Powell (2003): "Instrumental Variable Estimation of Nonparametric Models," *Econometrica*, 71(5), 1565–1578.

Paola, M., and M. Marco (2013): "Stochastic Choice and Consideration Sets," .

Pires, T. (2012): "Consideration Sets in Storable Goods Markets," *working paper, Northwestern University*.

Powell, J. L. (1984): "Least absolute deviations estimation for the censored regression model," *Journal of Econometrics*, 25(3), 303–325.

Ratneshwar, S., and A. D. Shocker (1991): "Substitution in use and the role of usage context in product category structures," *Journal of Marketing Research*, pp. 281–295.

Roberts, J. H., and J. M. Lattin (1991): "Development and Testing of a Model of Consideration Set Composition.," *Journal of Marketing Research (JMR)*, 28(4).

Santos, B. D. L., A. Hortacsu, and M. R. Wildenbeest (2012): "Testing Models of Consumer Search Using Data on Web Browsing and Purchasing Behavior," *American Economic Review*, 106, 2955–2980.

Seetharaman, P., A. Ainslie, and P. K. Chintagunta (1999): "Investigating household state dependence effects across categories," *Journal of Marketing Research*, pp. 488–500.

Sen, A. K. (1971): "Choice functions and revealed preference," *The Review of Economic Studies*, pp. 307–317.

Shocker, A. D., M. Ben-Akiva, B. Boccara, and P. Nedungadi (1991): "Consideration set influences on consumer decision-making and choice: Issues, models, and suggestions," *Marketing letters*, 2(3), 181–197.

Sudhir, K., and N. Yang (2014): "Exploiting the Choice-Consumption Mismatch: A New Approach to Disentangle State Dependence and Heterogeneity1," Discussion paper, Cowles Foundation for Research in Economics, Yale University.

Train, K. E. (2009): *Discrete choice methods with simulation*. Cambridge university press.

Van Nierop, E., B. Bronnenberg, R. Paap, M. Wedel, and P. H. Franses (2010): "Retrieving unobserved consideration sets from household panel data," *Journal of Marketing Research*, 47(1), 63–74.