Preference Estimation with Unobserved Choice Set Heterogeneity using Sufficient Sets*

Gregory S. Crawford† Rachel Griffith‡ Alessandro Iaria§

August 1, 2019

Abstract

In this paper, we provide an introduction to the problems that arise in estimating discrete choice models when choice sets are heterogeneous and unobserved to the econometrician. We discuss the two most popular approaches to tackling these problems and propose the idea of "sufficient sets" motivated by economic theory, which help to practically implement them.

*Griffith gratefully acknowledges financial support from the European Research Council (ERC) under ERC-2009-AdG-249529 and ERC-20150AdG-694822, and the Economic and Social Research Council (ESRC) under RES-544-28-0001 and ES/N011562/1. Iaria gratefully acknowledges financial support from the research grant Labex Ecodec: Investissements d'Avenir (ANR-11-IDEX-0003/Labex Ecodec/ANR-11-LABX-0047).

†Dept. of Economics, University of Zürich and CEPR, gregory.crawford@econ.uzh.ch

‡Dept. of Economics, University of Manchester and IFS, rgriffith@ifs.org.uk

§Dept. of Economics, University of Bristol, alessandro.iaria@bristol.ac.uk

1 Introduction

Discrete choice models are commonly used in applied economics. To estimate a discrete choice model requires that we have information on the set of alternatives over which the consumer is choosing. In many situations, it is trivial to observe consumers' choice sets, for example, Bay Area commuters' mode choice was between car and bus before BART was built. In other cases, researchers can explicitly ask survey respondents about the alternatives that they considered or ask for a ranking (e.g., Train and Winston (2007), Berry et al. (2004)). However, in many situations it can be difficult to formulate choice sets, for example, the discrete choice of housing with the constraint that there is only one house for sale at the specific address the buyer prefers (de Palma et al. (2007)). There are many reasons why consumer choice sets may be heterogeneous and unobserved by econometricians, including limited consumer attention, search, or endogenous product choices by firms. Failing to account for unobserved choice set heterogeneity will generally cause estimators of preference parameters to be inconsistent.

In this paper, we survey the two main empirical approaches to tackling the problem of unobserved choice set heterogeneity: "integrating over" and "differencing out" unobserved choice sets. The two approaches originate from different econometric literatures, started respectively by Manski (1977) and McFadden (1978). While integrating over heterogeneous unobserved choice sets is commonly done in empirical applications, differencing them out appears to be less popular in this context, possibly because the McFadden (1978)'s original motivation was to facilitate estimation with large but *observed* choice sets. We provide a unifying notation for understanding the two approaches and introduce the idea of "sufficient sets," which serve several purposes. First, sufficient sets help clarify that differencing out can also address the problem of unobserved choice sets, and that it is complementary to integrating over them. Second, sufficient sets prove useful to implement both approaches in practice. Third, they help translate economic assumptions derived from the characteristics of a given choice environment into econometric assumptions appropriate for estimation.

To build intuition, we begin by illustrating the well-understood result that, in a multinomial logit (MNL), mistakenly adding alternatives to a consumer's choice set leads to violations of the Independence of Irrelevant Alternatives (IIA) and to inconsistent estimators. This insight forms the basis of the "differencing out" approach proposed by McFadden (1978) for the consistent estimation of MNL models from subsets of true choice sets. Over the decades, the idea of estimating on subsets

of true choice sets has been shown to be effective also in more general models, such as Generalized Extreme Value models (e.g., Train et al. (1987), Bierlaire et al. (2008), and Guevara and Ben-Akiva (2013b)), mixed logit models with discrete distributions of random coefficients (e.g., Bajari et al. (2007), Fox et al. (2011), and Fox et al. (2016)), semi-parametric models (e.g., Fox (2007)) and, in the context of "long" panel data, mixed logit models with non-parametric distributions of random coefficients (e.g., Dubois et al. (2019)).[1]

When choice sets are *unobserved*, however, it is not clear how to construct proper subsets of consumers' true choice sets and therefore how to implement these estimators (e.g., Frejinger et al. (2009)). Inspired by Chamberlain (1980), we propose the use of consumers' observed choices paired with assumptions about the evolution of their unobserved choice sets over time as a practical tool to construct proper subsets in panel data environments. We call these subsets "sufficient sets." For example, in the case of the MNL with fixed effects studied by Chamberlain (1980), the set of permutations of a consumer's observed choices over time represents a sufficient set, indeed one which Chamberlain (1980) specifically chose to difference out fixed effects but that incidentally also differences out true and potentially unobserved choice sets.

The most popular approach to address unobserved choice set heterogeneity models the joint probability that a consumer is matched to a certain choice set and that she makes a specific choice from that choice set. As first proposed by Manski (1977), one can then obtain the marginal choice probability by "integrating over" unobserved choice set heterogeneity, as is routinely done with any other form of unobserved heterogeneity (e.g., standard mixed logit models that integrate over unobserved preference heterogeneity). The integrating over approach has a long tradition in marketing and transportation studies analysing consumers' consideration sets (e.g., Roberts and Lattin (1991), Ben-Akiva and Boccara (1995), Bronnenberg and Vanhonacker (1996), Chiang et al. (1998), Başar and Bhat (2004), Erdem and Swait (2004), Bruno and Vilcassim (2008), Van Nierop et al. (2010), Draganska and Klapper (2011), Ching et al. (2014), and Huang and Bronnenberg (2017)) and has also recently been applied in economics (e.g., Goeree (2008), De los Santos et al. (2012), Conlon and Mortimer (2013), Honka (2014), Abaluck and Adams (2017)).[2] Because the number of choice sets to integrate over grows exponentially in the number of available products, the practical implementation of this

[1] To the best of our knowledge, in the context of cross-sectional or "short" panel data, results of this kind are still not available for mixed logit models with continuous distributions of random coefficients, even though some interesting approximations have been proposed by Keane and Wasi (2012) and Guevara and Ben-Akiva (2013a).

[2] For a recent survey of these applied literatures in the context of consumer search, see Honka et al. (2018).

method is subject to a curse of dimensionality. Sufficient sets can be used to greatly reduce the number of choice sets to integrate over (i.e., only *supersets* of the sufficient sets can have positive probability mass), making this approach more viable in applications with large choice sets.[3]

We highlight the costs and benefits of differencing out unobserved choice sets relatively to integrating over them, a fact often overlooked in the applied literature, which typically only integrates over them. While the differencing out approach treats the choice set formation process as a nuisance parameter to be dropped from the likelihood function, the integrating over approach instead requires the specification of a choice set formation model which is estimated along with preferences given choice sets. On the one hand, while integrating over the distribution of unobserved choice sets requires additional functional form assumptions, data on the choice set formation process, and it is computationally more intensive, it enables researchers to learn about both consumer preferences and choice set formation. This may be essential in applications in which the key counterfactuals involve re-matching of choice sets to consumers (e.g., Gaynor et al. (2016)). On the other hand, differencing out unobserved choice sets requires less prior knowledge and data on the choice set formation process, and it is simpler to implement, but it does not allow inference about choice set formation. Whenever information on the choice set formation process is available, integrating over is likely to be the preferred approach. However, there are cases where such information is not available or is unreliable, or where the curse of dimensionality prevents implementation of the integrating over approach. In these cases, one can still learn about consumer preferences by differencing out unobserved choice sets.

Importantly, the economic characteristics of the choice environment can inform the specification of sufficient sets in both approaches. In other words, sufficient sets can help map concrete economic information about choice environments into suitable econometric assumptions. For example, settings characterized by non-sequential or fixed-sample search (e.g., Morgan and Manning (1985)) imply a choice environment that is stable for any individual consumer over time, but (possibly) varying across consumers. These models imply sufficient sets based on the collection of products chosen by a consumer over time, what we call a Full Purchase History sufficient set. Settings characterized by sequential search (e.g., Caplin et al. (2011)), limited attention (e.g., Eliaz and Spiegler (2011)), or consumer focus (e.g., Kőszegi and Szeidl (2013)) imply choice environments that evolve over time, and

[3]Goeree (2008) proposes a convenient importance sampling procedure (which we detail in Appendix B) that also greatly simplifies the practical implementation of this approach. Goeree (2008)'s importance sampling and sufficient sets are not mutually exclusive and can be used together.

suggest sufficient sets based on the accumulation of products chosen by a consumer in the past, what we call a Past Purchase History sufficient set. Cross-sectional settings where a group of individuals face a common choice environment, as for example in the analyses of Currie et al. (2010) of whether greater availability of fast food increases obesity or Gaynor et al. (2016) of hospital choice by doctors, suggest sufficient sets made of the collection of products observed to be chosen by individuals in the relevant group, what we call an Inter-Personal sufficient set. These are only a few examples of sufficient sets; they are neither mutually exclusive nor exhaustive. Sufficient sets can be combined or devised to reflect a large range of choice environments. We discuss the use of specification tests to aid comparisons between alternative sufficient sets within a particular application.

We implement some of the surveyed approaches in both Monte Carlo simulations and an empirical illustration. The illustration estimates price and advertising sensitivity for demand on-the-go for chocolate among adult women in the UK. Advertising can affect both a consumer's valuation of a specific product and the likelihood that the product enters the consumer's choice set. We first estimate a mixed logit model along the lines of Dubois et al. (2019) from an enlarged choice set including all chocolate products available in the type of outlet in which the consumer is shopping (what we call the Complete sufficient set). We then compare the estimates from this model to those from more robust mixed logit models that condition on various forms of consumer-specific past purchase histories at the type of outlet from which they are currently purchasing. The estimates of the impact of advertising on purchase probabilities differ significantly between the model based on the Complete sufficient set and the model based on the past purchase history sufficient sets; the Complete sufficient set may be incorrectly including alternatives in consumers' choice sets that were not in fact present, and this may be biasing upwards the estimated advertising sensitivities. These results are consistent with models of imperfect consumer attention as in Eliaz and Spiegler (2011). We also find that the estimated price sensitivities associated to the Complete sufficient set are on the whole upwardly biased, in line with the empirical findings by Goeree (2008).

The methods surveyed in this paper rely on an exogeneity assumption between consumer preferences and the matching of consumers to choice sets. This assumption accommodates models in which firms select the products to sell or in which consumers search for alternatives on the basis of both observable characteristics and expectations over unobservable characteristics, but rules out the

matching of consumers to choice sets on the basis of the realizations of unobserved preferences.[4] The problems of "endogenous" matching and of "unobserved" choice sets are of a fundamentally different nature and can be better understood in isolation. On the one hand, endogenous matching of choice sets gives rise to econometric problems akin to sample selection even when choice sets are perfectly observable.[5] On the other, unobserved choice set heterogeneity represents a non-trivial concern even when choice sets are exogenously matched to consumers. In this paper we limit our discussion to exogenous unobserved choice set heterogeneity and refer the reader to Hickman and Mortimer (2016) and Honka et al. (2018) for recent surveys on the endogenous matching of choice sets.[6]

In addition to the applied literatures mentioned earlier, there is a fast-growing theoretical literature in which limited attention is used to rationalize apparently incongruent consumer and firm behaviours that strongly motivates our interest in contributing to the debate on empirical approaches to accommodate such theories. These include, for example, consumer attention as in Eliaz and Spiegler (2011), Masatlioglu et al. (2012), and Manzini and Mariotti (2014), rational inattention as in Gabaix (2014) and Matejka and McKay (2015), search as in Janssen and Moraga-González (2004) and Rhodes (2014), screening rules as in Gilbride and Allenby (2004), models of salience as in Bordalo et al. (2014), focus as in Kőszegi and Szeidl (2013). In our empirical illustration, we develop sufficient sets and estimate preferences inspired by the work of Eliaz and Spiegler (2011).

The rest of the paper is structured as follows. In section 2, we introduce the model and illustrate how heterogeneity in unobserved choice sets may lead to inconsistent estimators of MNL models. This helps to build intuition for the possible solutions discussed in section 3. In section 4, we discuss the economic foundations of possible sufficient sets. We discuss specification tests in section 5 and provide an empirical illustration in section 6. A final section concludes. Several appendices provide Monte Carlo simulations, additional derivations, and additional results.

[4] Most applied papers dealing with choice set formation processes also rely on assumptions that guarantee exogenous matching, see for example: Goeree (2008), Draganska et al. (2009), Conlon and Mortimer (2013), and Eizenberg (2014).

[5] Some papers dealing with endogenous choice set heterogeneity include Iaria (2014), Musalem (2015), Ciliberto et al. (2016), Li et al. (2018), and Lu (2018).

[6] In addition, Hickman and Mortimer (2016) and Honka et al. (2018) also discuss the problem of unobserved choice set heterogeneity in the context of aggregate-level data on market shares (e.g., Goeree (2008) and Bruno and Vilcassim (2008)), while in the current paper we limit our discussion to individual-level purchase data. In our view, the idea of "sufficient set" does not suit well the case of aggregate data: similar to Chamberlain (1980), we propose to construct subsets of each individual's unobserved choice set on the basis of observed individual-level past purchases. In the case of aggregate data, it is not clear how one could use panel data on the evolution of market shares to obtain similar results to those available in the literature for individual-level discrete choice models.

2 The Problem of Unobserved Choice Set Heterogeneity

To build intuition, we start by illustrating that unobserved choice set heterogeneity leads to violations of the Independence of Irrelevant Alternatives (IIA) property in models with Gumbel errors and therefore to inconsistent estimators. Understanding the nature of the problem introduced by unobserved choice set heterogeneity helps to better appreciate the solutions proposed in the literature, which we will describe in the next section.

2.1 Basic Model and Notation

Let there be a panel of $i = 1, \ldots, I$ individuals each observed making a sequence of T choices, one per choice situation $t = 1, \ldots, T$. Denote i's sequence of choices by $Y_i = (Y_{i1}, \ldots, Y_{iT})$. For simplicity, we assume to observe exactly T choice situations for each i, but this is inessential.

We consider a situation in which i is matched to her choice set $CS_{it}^{\star}$ in choice situation t, but this choice set is unobserved to the researcher. Denote by $\times$ the cartesian product and let i's set of possible choice sequences be given by $\mathcal{CS}_i^{\star} = \times_{t=1}^{T} CS_{it}^{\star}$. By construction, any observed choice sequence Y_i must belong to $\mathcal{CS}_i^{\star}$.

Let preferences be defined by a parameter vector θ and let the probability with which i is matched to a given set of possible choice sequences, $\mathcal{CS}_i^{\star} = c$, be given by $\Pr\left[\mathcal{CS}_i^{\star} = c | \, \gamma\right]$, with γ also a parameter vector. In principle, γ could include some or all of the parameters that are in θ and could be the result of, for example, limited consumer attention, consumer search behavior, or strategic decision-making by firms.

Given θ and a specific match with a set of possible choice sequences, $\mathcal{CS}_i^{\star} = c$, each individual i is observed to make a sequence of choices $Y_i = j = (j_1, ..., j_T)$. We assume that the conditional indirect utility of alternative j_t in choice situation t for individual i is

$$U_{ij_tt} = V\left(X_{ij_tt}, \theta\right) + \epsilon_{ij_tt}, \tag{2.1}$$

where X_{ij_tt} is a vector of observable characteristics, and ϵ_{ij_tt} is the portion of i's utility that is unobserved to the econometrician. The probability that i is matched to the set of possible choice

sequences $\mathcal{CS}_i^\star = c$ *and* makes a sequence of choices $Y_i = j$ is:

$$\Pr\left[Y_i = j, \mathcal{CS}_i^\star = c \middle| \theta, \gamma\right] = \Pr\left[Y_i = j \middle| \mathcal{CS}_i^\star = c, \theta\right] \Pr\left[\mathcal{CS}_i^\star = c \middle| \gamma\right]. \tag{2.2}$$

The first term in (2.2) is the probability of choosing j solely due to preferences given the sequence of choice sets i is matched to, and the second is the probability that the individual is matched to that sequence of choice sets. The following assumption implies that $\Pr\left[Y_i = j \middle| \mathcal{CS}_i^\star = c, \theta\right]$ is multinomial logit (MNL) for any c.

Assumption 1. Conditional on all $V(X_{ij_t t}, \theta)$'s and on $\mathcal{CS}_i^\star = c$, $\epsilon_{ij_t t}$ from (2.1) is distributed i.i.d. Gumbel.

Assumption 1 allows for general matching processes: $\Pr[\mathcal{CS}_i^\star = c|\gamma]$ can take any form and be a function of any element of the MNL model $\Pr[Y_i = j|\mathcal{CS}_i^\star = c, \theta]$. For example, this accommodates models in which firms select the products to sell or in which individuals search for alternatives on the basis of both observable characteristics and expectations over unobservable characteristics, but rules out the matching of individuals to choice sets on the basis of the *realizations* of the unobservables, $\epsilon_{ij_t t}$'s. In other words, Assumption 1 rules out the possibility that individuals and choice sets are *endogenously* matched.[7] The problems of "endogenous" matching and of "unobserved" choice sets are of a fundamentally different nature and can be better understood in isolation. On the one hand, endogenous matching of choice sets gives rise to econometric problems akin to sample selection even when choice sets are perfectly observable. On the other, as we will detail below, unobserved choice set heterogeneity represents a non-trivial concern even when choice sets are exogenously matched to individuals. In the interest of space, in this paper we limit our discussion to *exogenous* unobserved choice set heterogeneity and refer the reader to Hickman and Mortimer (2016) and to Honka et al. (2018) for recent surveys on endogenous matching of choice sets.

We use Assumption 1 to provide some intuition about the econometric problem introduced by unobserved choice set heterogeneity and to illustrate the basic ideas behind the main solutions proposed in the literature. Afterwards, we relax Assumption 1 and extend the discussion to more general models than the MNL.

[7] See footnote 5 in the introduction for references on the endogenous matching of choice sets in demand estimation.

An implication of Assumption 1 is that conditional Maximum Likelihood estimators of θ can be constructed from $\Pr[Y_i = j|\mathcal{CS}_i^\star = c, \theta]$, since this conditional probability is multinomial logit.[8] Using Assumption 1, i's conditional probability of selecting choice sequence $Y_i = j$, given their set of possible choice sequences, $\mathcal{CS}_i^\star = c = \times_{t=1}^{T} c_t$, is the familiar product of T MNL's, each one specific to a choice situation along the sequence:

$$\Pr\left[Y_i = j \middle| \mathcal{CS}_i^\star = c, \theta\right] = \prod_{t=1}^{T} \frac{\exp\left(V\left(X_{ij_t t}, \theta\right)\right)}{\sum_{k \in CS_{it}^\star = c_t} \exp\left(V\left(X_{ik_t t}, \theta\right)\right)}. \tag{2.3}$$

2.2 The Problem of Adding Unavailable Alternatives

Because $\mathcal{CS}_i^\star = c$ is unobserved, equation (2.3) cannot be directly used for estimation. In order to proceed, usually the researcher will specify "a" set of choice sequences, $\mathcal{S}_i = s$, possibly different from $\mathcal{CS}_i^\star = c$, on the basis of which to construct a likelihood function. Researchers often specify $\mathcal{S}_i$ to be common across i and given by, for example, all those products above a certain market share threshold or a given number of products with the largest market shares. Suppose that the researcher specifies the likelihood function to be used in estimation as the conditional probability of i choosing $Y_i = j$ from the set of choice sequences $\mathcal{S}_i = s = \times_{t=1}^{T} s_t$. Then, the potentially misspecified model is:

$$\Pr\left[Y_i = j \middle| \mathcal{S}_i = s, \theta\right] = \prod_{t=1}^{T} \frac{\exp\left(V\left(X_{ij_t t}, \theta\right)\right)}{\sum_{k \in S_{it} = s_t} \exp\left(V\left(X_{ik_t t}, \theta\right)\right)}, \tag{2.4}$$

where the difference between (2.3) and (2.4) lies in the terms included in the summations in the denominator of each. As it is well known, using model (2.4) to estimate θ in the presence of unobserved choice set heterogeneity will *not* be a problem whenever the researcher manages to specify $\mathcal{S}_i = s \subseteq \mathcal{CS}_i^\star = c$. Differently, whenever $\mathcal{S}_i = s$ includes alternatives not originally available in $\mathcal{CS}_i^\star = c$, then the use of model (2.4) will lead to inconsistent estimators of θ. Not surprisingly, this is a consequence of model (2.4) satisfying (when $\mathcal{S}_i = s \subseteq \mathcal{CS}_i^\star = c$) or violating (when $\mathcal{S}_i = s \nsubseteq \mathcal{CS}_i^\star = c$) the IIA property from Assumption 1. To see this, note that individual i's probability of choosing sequence j among the potentially misspecified $\mathcal{S}_i$, given that the true set of sequences is $\mathcal{CS}_i^\star = c$ and conditional

[8] Given Assumption 1, if γ shares some common element with θ (as is likely), failing to control for the choice set matching process $\Pr\left[\mathcal{CS}_i^\star = c \middle| \gamma\right]$ only causes a loss of efficiency in the resulting conditional Maximum Likelihood estimator relative to a joint Maximum Likelihood estimator derived from Equation (2.2).

on the vector of parameters θ is:

$$\begin{aligned}\Pr\left[Y_i = j \middle| \mathcal{S}_i = s, \mathcal{CS}_i^\star = c, \theta\right] &= \prod_{t=1}^{T} \frac{\Pr[Y_{it} = j_t | CS_{it}^\star = c_t, \theta]}{\sum_{r_t \in s_t \cap c_t} \Pr[Y_{it} = r_t | CS_{it}^\star = c_t, \theta] + \sum_{k_t \in s_t \setminus c_t} \Pr[Y_{it} = k_t | CS_{it}^\star = c_t, \theta]} \\ &= \prod_{t=1}^{T} \frac{\exp\left(V\left(X_{ij_tt}, \theta\right)\right)}{\sum_{r_t \in S_{it} = s_t \cap CS_{it}^\star = c_t} \exp\left(V\left(X_{ir_tt}, \theta\right)\right)} \\ &= \Pr\left[Y_i = j \middle| \mathcal{S}_i = s \cap \mathcal{CS}_i^\star = c, \theta\right],\end{aligned} \tag{2.5}$$

where the denominator in the first line decomposes s_t into those alternatives that are in c_t ($r_t \in s_t \cap c_t$) and those that are not ($k_t \in s_t \setminus c_t$). The second equality obtains as the probability i selects an alternative not in her true choice set is zero, $\Pr[Y_{it} = k_t | CS_{it}^\star = c_t, \theta] = 0$ for all $k_t \notin CS_{it}^\star = c_t$. In other words, because Assumption 1 implies the IIA property only when the choice set assumed by the researcher is a *subset* of i's true choice set, $\mathcal{S}_i = s \subseteq \mathcal{CS}_i^\star = c$, equation (2.5) is not guaranteed to equal (2.4). By expressing (2.5) in terms of (2.4), we obtain:

$$\begin{aligned}\Pr\left[Y_i = j \middle| \mathcal{S}_i = s \cap \mathcal{CS}_i^\star = c, \theta\right] &= \prod_{t=1}^{T} \frac{\exp\left(V\left(X_{ij_tt}, \theta\right)\right)}{\sum_{r_t \in s_t \cap c_t} \exp\left(V\left(X_{ir_tt}, \theta\right)\right)} \frac{\sum_{m_t \in s_t} \exp\left(V\left(X_{im_tt}, \theta\right)\right)}{\sum_{m_t \in s_t} \exp\left(V\left(X_{im_tt}, \theta\right)\right)} \\ &= \prod_{t=1}^{T} \frac{\exp\left(V\left(X_{ij_tt}, \theta\right) - \ln\left(\frac{\sum_{r_t \in s_t \cap c_t} \exp\left(V\left(X_{ir_tt}, \theta\right)\right)}{\sum_{m_t \in s_t} \exp\left(V\left(X_{im_tt}, \theta\right)\right)}\right)\right)}{\sum_{m_t \in S_{it} = s_t} \exp\left(V\left(X_{im_tt}, \theta\right)\right)} \\ &= \prod_{t=1}^{T} \frac{\exp(V\left(X_{ij_tt}, \theta\right) - \ln(\pi_{it}(\theta)))}{\sum_{m_t \in S_{it} = s_t} \exp\left(V\left(X_{im_tt}, \theta\right)\right)}.\end{aligned} \tag{2.6}$$

Suppose $s_t \cap c_t \subset s_t$ for some t's (i.e., s_t includes alternatives not in c_t), then $\ln(\pi_{it}) < 0$ for those t's and models (2.4) and (2.5) will differ. In this case, if estimation proceeds on the basis of model (2.4), the likelihood function will be mistakenly ignoring a sequence of up to T fixed effects for each i, $\ln(\pi_{it})$'s, which are functions of the rest of the model. Differently, suppose instead $s_t \subseteq c_t$ for all t's, then $\ln(\pi_{it}) = 0$ for all t's and (2.5) equals (2.4). Consequently, the model used in estimation will correspond to the true conditional choice model.

More succinctly, given true model (2.3), the likelihood function obtained from model (2.4) will mistakenly ignore a sequence of (i,t)-specific fixed effects if and only if at least one choice set $S_{it} = s_t$ of the sequence $\mathcal{S}_i = s$ includes at least one alternative not originally available in $CS^{\star}_{it} = c_t$. The $\ln(\pi_{it}(\theta))$'s are (i,t)-specific fixed effects that cause inconsistency in estimation. $\pi_{it}(\theta)$ measures the probability that individual i would choose one of the alternatives in her true choice set when faced with the set of products assumed by the researcher. If $s_t \cap c_t \subset s_t$, i.e. if s_t includes alternatives not in c_t, this probability will be strictly less than one, and smaller (and thus the likely inconsistency greater) the more likely it is that i would have preferred one of the products mistakenly included in s_t.

Discussion. The previous subsection illustrates that the econometric issue introduced by unobserved choice set heterogeneity can be characterized, even in MNL models, as a violation of the IIA property. This violation introduces individual-time-specific fixed effects that are functions of the rest of the model and that lead to the inconsistency of the estimator of θ. In Appendix A, we explore the severity of this econometric bias in some Monte Carlo experiments and demonstrate that it can be substantial.

The inconsistency induced by incorrectly including in individuals' choice sets alternatives among which they did not choose cannot be resolved by including alternative-specific constants or random coefficients; they will not control for the individual-time-specific fixed effects, because they are *individual- and time*-specific and their distribution is a function of all of the observables. In Appendix A, we illustrate this point with simulations.

The methods proposed in the literature to address unobserved choice set heterogeneity can be grouped into two families: those that "integrate over" and those that "difference out" unobserved choice sets. The first is the approach of Manski (1977), which models the unconditional probability i chooses j by integrating over *all* possible unobserved choice sets that include j. This is akin to treating unobserved choice set heterogeneity in a manner analogous to unobserved preference heterogeneity, and has been used in the applied literature (e.g., Goeree (2008) and Van Nierop et al. (2010)).

We propose a second approach that aims at "differencing out" unobserved choice sets by conditioning choice probabilities on *subsets* of the true choice sets. This approach builds on well known estimators originally developed for other purposes, such as McFadden (1978), Fox (2007), Fox et al.

(2011), and Dubois et al. (2019) to "difference out" unobserved choice sets in various models (beyond the MNL). In the next section, we discuss both of these approaches and their relative advantages and disadvantages.

3 Two Solutions: Integrating Over and Differencing Out Heterogeneous Unobserved Choice Sets

We start the section by discussing the "integrating over" approach proposed by Manski (1977). We then illustrate the "differencing out" approach in the context of three popular models: the MNL, the mixed logit with (a) non-parametric distribution of random coefficients in panel data with "large" T and (b) discrete distribution of random coefficients in panel data with "small" T, and the class of semi-parametric models studied by Fox (2007). Finally, we discuss how the two approaches can be combined to reduce the computational burden of the "integrating over" approach.

3.1 "Integrating Over" Unobserved Choice Sets

The most popular approach used in the literature to address unobserved choice set heterogeneity is to jointly model choice set formation and the purchase decision given a choice set. As originally discussed by Manski (1977), the unconditional probability of i selecting choice sequence j can be written as:

$$\Pr[Y_i = j|\theta, \gamma] = \sum_{c \in C_i^\star} \Pr[Y_i = j|\mathcal{CS}_i^\star = c, \theta] \Pr[\mathcal{CS}_i^\star = c|\gamma], \tag{3.1}$$

where $C_i^\star$ is the collection of possible sets of choice sequences to which individual i can be matched. By having information on the matching process between individuals and choice sets, one can integrate over unobserved choice set heterogeneity, as it is routinely done with unobserved preference heterogeneity. Until recently, it was believed that the identification of model (3.1) relied on the availability of auxiliary data about $\Pr[\mathcal{CS}_i^\star = c|\gamma]$ (e.g., Roberts and Lattin (1991)) and/or the availability of instruments that exclusively affected the matching between individuals and choice sets (e.g., Goeree (2008)). In a recent

paper, however, Abaluck and Adams (2017) present identification results for this model that do not require the availability of such auxiliary data.[9]

Whenever model (3.1) is correctly specified and $C_i^\star$ is not too large, so that estimation is actually possible (e.g., by a Simulated Maximum Likelihood Estimator of (3.1)), then one can use this approach to learn both about preferences θ and about the matching process between individuals and choice sets γ. Knowledge of both θ and γ is essential in several contexts, especially when the researcher is interested in simulating counterfactuals that may involve re-matching of choice sets to individuals (e.g., Gaynor et al. (2016)). In Appendix B, we describe the details of a convenient importance sampling procedure proposed by Goeree (2008) for the joint estimation of (θ, γ).

Even though the Manski (1977)'s approach represents the best option in many instances, there are also cases in which it may not be appropriate. We will spend the rest of the subsection discussing three of the main possible drawbacks of the integrating over approach. One difficulty relates to the requirement of further functional form information: in addition to $\Pr[Y_i = j|\mathcal{CS}_i^\star = c, \theta]$, the practical implementation of model (3.1) requires knowledge also of the functional form of $\Pr[\mathcal{CS}_i^\star = c|\gamma]$. Another potential drawback, perhaps less obvious, is that the collection of choice sets over which expectations are taken is hardly observable and $C_i^\star$ may easily be misspecified. Finally, even when $\Pr[Y_i = j|\theta, \gamma]$ is correctly specified, in practice the estimation of model (3.1) may prove difficult. Indeed, this model suffers from a curse of dimensionality related to the number of elements in $C_i^\star$, which grows exponentially in the number of alternatives available to individual i, J_i (e.g., Abaluck and Adams (2017) limit their estimations to the case of J_i=10).[10]

We illustrate the potential drawbacks just discussed in the context of Goeree (2008), a popular paper that relies on the Manski (1977)'s approach. In our notation, Goeree (2008)'s choice model can be written as:

$$\Pr[Y_{it} = j_t|\theta, \gamma] = \sum_{c_t \in C_t^j} \frac{\exp\left(V(X_{ij_tt}, \theta)\right)}{\sum_{r_t \in c_t} \exp\left(V(X_{ir_tt}, \theta)\right)} \left[\prod_{l_t \in c_t} \phi_{il_tt}(\gamma) \prod_{k_t \notin c_t} \left(1 - \phi_{ik_tt}(\gamma)\right)\right], \tag{3.2}$$

[9]The intuition of their argument is that whenever some alternatives are not in the choice sets of some individuals, the discrete-choice analogue of Slutsky symmetry will be violated. They show that one can therefore use deviations from Slutsky symmetry to separately identify γ and θ.

[10]Two possible ways of alleviating the practical consequences of this curse of dimensionality are the importance sampling procedure proposed by Goeree (2008) (detailed in Appendix B) and the idea proposed in subsection 3.3 below. Note that both simplifications are not "free" and require additional assumptions on the choice set formation process $\Pr[\mathcal{CS}_i^\star = c|\gamma]$.

where C_t^j is the collection of *all* period t choice sets that include product j_t. This specification relies on:

- Our Assumption 1, with $\Pr[Y_{it} = j_t | CS_{it}^\star = c_t, \theta] = \dfrac{\exp\left(V(X_{ij_tt}, \theta)\right)}{\sum_{r_t \in c_t} \exp\left(V(X_{ir_tt}, \theta)\right)}$ and
- The additional assumption that consideration of each product is *independent* of the consideration of the other products: $\Pr\left[CS_{it}^\star = c_t | \gamma\right] = \prod_{l_t \in c_t} \phi_{il_tt}(\gamma) \prod_{k_t \notin c} \left(1 - \phi_{ik_tt}(\gamma)\right)$.

Even given this second assumption, in Goeree (2008) the total number of products is still very large ($> 2,100$), so the non-parametric estimation of all the ϕ's is not feasible. Consequently, she further assumes that:

$$\phi_{il_tt}(\gamma) = \frac{\exp\left(W_{il_tt}(\gamma)\right)}{1 + \exp\left(W_{il_tt}(\gamma)\right)}. \tag{3.3}$$

This implies that *every* $c_t \in C_t^j$ will have a strictly positive probability in the distribution of choice sets for each (i,t), so that $\Pr\left[CS_{it}^\star = c_t | \gamma\right] > 0$ for every (i,t). However, it may be that for some (i,t) combination $\Pr\left[CS_{it}^\star = c_t | \gamma\right] = 0$ for some c_t, i.e. c_t is not in the support of the choice set distribution to which individual i can be matched to in period t:

$$\Pr\left[CS_{it}^\star = c_t | \gamma\right] = \begin{cases} \displaystyle\prod_{l_t \in c_t} \phi_{il_tt}(\gamma) \prod_{k_t \notin c_t} \left(1 - \phi_{ik_tt}(\gamma)\right) & \text{if } c_t \in C_{it}^{j\star} \\ 0 & \text{if } c_t \in C_t^j \setminus C_{it}^{j\star}, \end{cases} \tag{3.4}$$

where $C_{it}^{j\star}$ is the collection of choice sets to which individual i can possibly be matched to in period t. Since $C_{it}^{j\star}$ is typically unobserved and heterogeneous across (i,t) combinations, model (3.2) and (3.3) will suffer from support misspecification whenever there exist observations where the true collection of unobserved choice sets to which an individual can be matched to is restricted, i.e. $\exists\,(i,j,t)$ combination such that $C_{it}^{j\star} \subset C_t^j$. With support misspecification, standard estimators of model (3.2) and (3.3) will be inconsistent.

To be clear, computing expectations over the power set of the universal set, as with C_t^j in (3.2), would *not* be a problem if one could afford to estimate a truly flexible specification for ϕ that was able to accommodate $\Pr\left[CS_{it}^\star = c_t | \gamma\right] = 0$ whenever necessary. The problem arises because one is not usually able to estimate a truly flexible model for ϕ, and needs to make additional assumptions

along the lines of (3.3). Taken together, C_t^j in (3.2) and (3.3) may introduce bias due to the potential inclusion of infeasible choice sets.[11]

3.2 "Differencing Out" Unobserved Choice Sets

When Manski (1977)'s approach is not appropriate, possibly because of the drawbacks discussed in the previous section, one can still hope to estimate the preference parameters θ by *differencing out* unobserved choice sets. We start by introducing the differencing out approach in the context of the MNL model, then in the next two subsections we extend the discussion to models that relax the IIA property.

3.2.1 Multinomial Logit Model

The differencing out approach relies on i's observed choice sequence, $Y_i = j$, to construct *subsets* of i's true but unobserved choice set, $\mathcal{CS}_i^\star = c$, without requiring additional data. We call these subsets *sufficient sets*. Specifically, consider any correspondence $f(Y_i)$ that satisfies the following property.

Condition 1. Given any choice sequence $Y_i \in \mathcal{CS}_i^\star$, the correspondence f is such that $Y_i \in f(Y_i)$ and $f(Y_i) \subseteq \mathcal{CS}_i^\star$.

It is easy to see that if Assumption 1 (i.e., MNL model) and Condition 1 hold, then $f(Y_i)$ will be a sufficient statistic for $\mathcal{CS}_i^\star$ or, equivalently, a MNL model conditional on $f(Y_i)$ will be guaranteed to satisfy the IIA property even if choice sets are unobserved. It is for this reason that we call any f that satisfies Condition 1 a sufficient set. More precisely, if Assumption 1 and Condition 1 hold, then for every individual i and choice sequence $Y_i = j$ such that $f(j) = r$:

$$\Pr[Y_i = j | f(Y_i) = r, \theta] = \frac{\prod_{t=1}^T \exp(V(X_{ij_t t}, \theta))}{\sum_{k \in f(Y_i)=r} \prod_{t=1}^T \exp(V(X_{ik_t t}, \theta))} \tag{3.5}$$

[11] Two other papers that take a similar approach to Goeree (2008) are Van Nierop et al. (2010) and Draganska and Klapper (2011). Van Nierop et al. (2010) assume that the model for $\Pr\left[CS_{it}^\star = c_t|\gamma\right]$ is a J-dimensional multivariate normal distribution, which again cannot be 0 for any c_t. In the corresponding equation to (3.2), they compute expectations over C_t^j, associating positive mass to every $c_t \in C_t^j$ rather than only to each $c_t \in C_{it}^{j\star} \subset C_t^j$. Draganska and Klapper (2011) assume $\Pr\left[CS_{it}^\star = c_t|\gamma\right]$ to be a MNL model with choice set C_t^j, and compute expectations over C_t^j, where again the multinomial logit model cannot be exactly 0 for any $c_t \in C_t^j$.

and θ can be consistently estimated by the conditional Maximum Likelihood Estimator derived from $\Pr[Y_i = j | f(Y_i) = r, \theta]$. Equation (3.5) is a direct consequence of the IIA property.[12] We define the MNL in (3.5) as the **Sufficient Set Logit (SSL)** model.

In essence, one can estimate preferences θ based only on the variation in characteristics of those products in i's sufficient set, rather than on her full (but unobserved) sequence of choice sets. This is evident as equation (3.5) does not depend on i's unobserved sequence of choice sets, $\mathcal{CS}_i^* = c$. Whenever $\mathcal{CS}_i^\star = c$ is observed, the econometrician can easily detect appropriate subsets of $CS_{it}^\star = c_t$ for any i in t, and rely on McFadden (1978) to consistently estimate θ. However, as we saw in subsection 2.2, whenever $\mathcal{CS}_i^\star$ is unobserved the econometrician needs to be careful in constructing sets of choice sequences for each i and t that are actual *subsets* of $CS_{it}^\star$. The main idea of the differencing out approach is to exploit individual i's observed choice sequence Y_i, paired with assumptions about the evolution of $CS_{it}^\star$ over t, to construct a proper subset of $\mathcal{CS}_i^\star$, the sufficient set $f(Y_i)$, and then to rely on McFadden (1978) to consistently estimate θ.

The differencing out approach is quite general and works for any f that generates subsets of $\mathcal{CS}_i^\star$. In section 4 we discuss how different economic theories naturally lead to sufficient sets that satisfy Condition 1. While these examples of sufficient sets are suggestive, we note here that they represent *a* set of sufficient conditions that imply the SSL in (3.5), but they are neither necessary nor the minimal sufficient conditions for the result to hold. In subsection 5.1 we discuss statistical tests to help researchers choose among different sufficient sets.

The SSL expression in (3.5) makes clear that individual i will have a non-zero log-likelihood contribution whenever her observed choice sequence $Y_i = j$ gives rise to a non-singleton $f(j) = r$. In the practical examples of sufficient sets we discuss in the paper, similar to Chamberlain (1980), this happens whenever the observed choice sequence $Y_i = (j_1, ..., j_t, ..., j_T)$ entails "some" switching, so that there exist at least two elements j_t and $j_{t'}$ in the sequence for which $j_t \neq j_{t'}$. All those observations for which only one alternative is chosen repeatedly throughout the sequence will be dropped from the log-likelihood function.

Note that, in practice, the SSL is a MNL in which the choice set is given by a (potentially huge) set of choice sequences over T choice situations, $f(Y_i) = r$. While this can be numerically inconvenient for some peculiar $f(Y_i)$, it will usually be possible to re-express (3.5) in a way that greatly simplifies

[12] For completeness, we report a derivation of (3.5) in Appendix C.

its practical implementation. In particular, the SSL in (3.5) over choice sequences can be equivalently expressed as the product of T separate t-specific MNL's over *alternatives* if and only if the sufficient set (over *choice sequences*) $f(Y_i)$ can be expressed as the cartesian product of T separate t-specific sufficient sets (over alternatives). In other words:

$$\begin{aligned} \Pr[Y_i = j | f(Y_i) = r, \theta] &= \prod_{t=1}^{T} \Pr[Y_{it} = j_t | f_t(Y_i) = r_t, \theta] \\ &= \prod_{t=1}^{T} \frac{\exp(V(X_{ij_t t}, \theta))}{\sum_{k_t \in f_t(k) = r_t} \exp(V(X_{ik_t t}, \theta))} \end{aligned} \tag{3.6}$$

if and only if $f(Y_i) = \times_{t=1}^{T} f_t(Y_i)$.[13] To see why expression (3.6) results in more convenient estimators than expression (3.5), suppose that the econometrician specifies a $f(Y_i)$ such that in each $t = 1, ..., 10$ an individual can choose one out of $J_t = J = 5$ different alternatives. It follows that $f(Y_i)$ contains 5^{10} possible choice sequences of length $T = 10$. By using the SSL in (3.5), the econometrician would have to estimate a huge MNL model with a summation over 5^{10} addends in the denominator. However, since $f(Y_i)$ can be obtained as the cartesian product of $T = 10$ separate t-specific sets each containing $J = 5$ alternatives, expression (3.6) guarantees that this SSL model can equivalently be expressed as the product of $T = 10$ MNL models each with a summation over $J = 5$ addends in the denominator. The examples of sufficient sets that we propose in section 4 all satisfy this condition, giving rise to computationally simple estimators. As we will illustrate later, a famous exception to the equivalence between (3.5) and (3.6) is the Choice Permutations sufficient set first proposed by Chamberlain (1980), which defines—in our terminology—the sufficient set as the collection of all permutations of the observed $Y_i = j$ and which results in Chamberlain (1980)'s fixed effect logit model.[14]

3.2.2 Mixed Logit Models

In this subsection, we illustrate how sufficient sets can be used in the context of mixed logit models. We start by discussing the simpler case in which the econometrician observes a large T for each

[13] For a derivation of this result, see Appendix D.

[14] In those cases in which $f(Y_i)$ cannot be expressed as $\times_{t=1}^{T} f_t(Y_i)$, the econometrician can directly apply McFadden (1978) and estimate θ from subsets of the *observed* but potentially huge sufficient sets $f(Y_i)$. As an example, D'Haultfœuille and Iaria (2016) implement this idea in the context of Chamberlain (1980)'s fixed effect logit model. Matlab codes are available on the authors' personal webpages.

individual. We then move on to the more complex scenario in which the econometrician only observes a small T for each individual: after a brief introduction of the model and basic challenges, we focus on a discrete mixture version of it for which both identification and estimation can be discussed in simple and intuitive terms.

Non-Parametric Distribution of Random Coefficients: Large T

When the econometrician observes a large number of choice situations per individual, $T \to \infty$ for fixed I, then she can rely on the results from the previous subsection and on the estimator proposed by Dubois et al. (2019) to estimate a mixed logit model with unobserved choice sets and non-parametric distribution of random coefficients.

Assumption 2(a). Suppose that each i has systematic utilities of the form $V(X_{ij_tt}, \theta_i)$ with individual-specific preferences θ_i distributed according to $p(\theta_i = \theta|\psi)$, where ψ is a vector of parameters. Conditional on all $V(X_{ij_tt}, \theta_i)$'s, on $\theta_i = \theta$, and on $\mathcal{CS}_i^\star = c$, ϵ_{ij_tt} from (2.1) is distributed i.i.d. Gumbel.

Irrespective of the distribution of θ_i, when $f(Y_i) = \times_{t=1}^T f_t(Y_i)$, it follows from the previous subsection that i's probability of choosing $Y_i = j$ conditional on $f(Y_i) = r$ and on θ_i is given by:

$$\Pr[Y_i = j | f(Y_i) = r, \theta_i] = \prod_{t=1}^{T} \frac{\exp(V(X_{ij_tt}, \theta_i))}{\sum_{k_t \in f_t(k)=r_t} \exp(V(X_{ik_tt}, \theta_i))}, \tag{3.7}$$

which we call **Individual Sufficient Set Logit (ISSL)** to distinguish it from the standard SSL from (3.6). Even though the realization of the random coefficients θ_i is unobserved and potentially hetereogeneous across individuals, when $T \to \infty$ the ISSL model (3.7) can be directly used as an individual-specific likelihood function on the basis of which to construct a consistent individual-specific MLE of θ_i. In other words, for any finite I, when T is large the econometrician can treat the unobserved θ_i as a "fixed effect" and estimate it from the individual-specific MLE derived from (3.7), for each $i = 1, ..., I$. Dubois et al. (2019) propose this idea for situations with observed choice sets, but the results from the previous subsection imply that their methods—when combined with sufficient sets from Condition 1—can be readily applied also to situations with unobserved choice sets.

At an intuitive level, by following this procedure, the econometrician would estimate by MLE a separate $\widehat{\theta_i}$ for each individual $i = 1, ..., I$ and then recover non-parametrically the distribution of random coefficients $p(\theta_i = \theta|\psi)$ by simply computing the frequency of each realization $\theta_i = \theta$ among the estimates. For more details about this estimation procedure, see Dubois et al. (2019).

Discrete Distribution of Random Coefficients: Small T

Here we discuss the more complex scenario in which the econometrician only observes a small number of choice situations T per individual, with $I \to \infty$.[15] In this case, unfortunately, the econometrician will not be able to treat θ_i as a fixed effect and directly rely on the ISSL model (3.7) to consistently estimate it. She will rather need to treat θ_i as a random effect and, relying on its distribution $p(\theta_i = \theta|\psi)$, to integrate it over when deriving i's choice probability.

Given Assumption 2(a) and Condition 1, by conditioning the probability of choice sequence $Y_i = j$ on the sufficient set $f(Y_i) = r$, with $f(Y_i) = r \subseteq \mathcal{CS}_i^\star = c$, we obtain the Sufficient Set Mixed Logit (SSML):

$$\begin{aligned}
&\Pr[Y_i = j|f(Y_i) = r, \psi] \\
&= \int_\theta \Pr[Y_i = j, \theta_i = \theta|f(Y_i) = r, \psi]d\theta \\
&= \int_\theta \Pr[Y_i = j|f(Y_i) = r, \theta_i = \theta]p(\theta_i = \theta|f(Y_i) = r, \psi)d\theta \\
&= \int_\theta \frac{\prod_{t=1}^T \exp\left(V(X_{ij_t t}, \theta)\right)}{\sum_{k \in f(Y_i)=r} \prod_{t=1}^T \exp\left(V(X_{ik_t t}, \theta)\right)} p(\theta|f(Y_i) = r, \psi)d\theta.
\end{aligned} \tag{3.8}$$

Note from (3.8) that the distribution of random coefficients used to integrate over unobserved preference heterogeneity is *conditional* on the realization of the sufficient set $f(Y_i) = r$, i.e. $p(\theta_i = \theta|f(Y_i) = r, \psi)$. As a consequence, and differently from the *unconditional* mixed logit model commonly estimated by applied researchers, the SSML in (3.8) is a conditional mixed logit model which requires the specification of the conditional distribution of random coefficients $p(\theta_i = \theta|f(Y_i) = r, \psi)$, as opposed to the more standard unconditional distribution of random coefficients $p(\theta_i = \theta|\psi)$. The two are related by:

[15] This is the typical situation considered in the current paper, with the only exception being the ISSL model.

$$p(\theta_i = \theta|\psi) = \sum_r p(\theta_i = \theta|f(Y_i) = r, \psi) \Pr[f(Y_i) = r], \tag{3.9}$$

where the probability of each realization r of the sufficient set, $\Pr[f(Y_i) = r]$, is observed in the data. Given (3.9), it is apparent how parametric assumptions on the unconditional distribution of random coefficients, $p(\theta_i = \theta|\psi)$, will not typically translate into convenient restrictions on the conditional distributions, $p(\theta_i = \theta|f(Y_i) = r, \psi)$. For example, if $p(\theta_i = \theta|\psi)$ is a normal density, that certainly does not imply that $p(\theta_i = \theta|f(Y_i) = r, \psi)$ will also be a normal density. This complexity is a consequence of the sample selection on the realized random coefficients introduced by the conditioning $f(Y_i) = r$. This selection problem greatly complicates general treatments about identification and estimation of the SSML model (3.8), for a related discussion see Keane and Wasi (2012). However, there are simple versions of (3.8) whose identification can be readily shown and estimation easily performed in practice.

We now turn to one of these cases, where both the random coefficients θ_i and the regressors X_i have a discrete, finite, and known support. In this scenario, the identification of the model can be described in very transparent terms, while the estimation can be performed by Ordinary Least Squares (OLS) or, to improve efficiency, by the easy-to-implement inequality-constrained least square estimator proposed by Bajari et al. (2007) and Fox et al. (2011).

Assumption 2(b). Suppose the matrix $X_i = [X_{i1}, \ldots, X_{it}, \ldots, X_{iT}]$ gathers all of i's regressors over the T choice situations, $\Theta = \{\theta_1, \ldots, \theta_q, \ldots, \theta_Q\}$ is the discrete and finite support of the random coefficients θ_i, and $\Psi^r = \left[\psi_1^r, \ldots, \psi_q^r, \ldots, \psi_Q^r\right]'$ the associated conditional weights or conditional probability masses. There is a finite number P of different values taken by the regressors X_i, so that any $X_i \in \{X_1, \ldots, X_p, \ldots, X_P\}$. Both Θ and $\{X_1, \ldots, X_p, \ldots, X_P\}$ are known to the econometrician.

Given the additional Assumption 2(b), the SSML model (3.8) simplifies to:

$$\Pr\left[Y_i = j \middle| X_i, f\left(Y_i\right) = r, \Theta, \Psi^r\right] = \sum_{q=1}^{Q} \Pr\left[Y_i = j \middle| X_i, f\left(Y_i\right) = r, \theta_i = \theta_q\right] \times \psi_q^r. \tag{3.10}$$

For the purpose of identification, the left hand side of (3.10), $\Pr\left[Y_i = j \middle| X_p, f\left(Y_i\right) = r, \Psi^r\right]$, is known for any combination (j, p, r). Because $\Theta = \{\theta_1, \ldots, \theta_q, \ldots, \theta_Q\}$ is known, each $\Pr\left[Y_i = j \middle| X_p, f\left(Y_i\right) = r, \theta_i = \theta_q\right]$

from (3.10) is also known for any (j, X_p, r, θ_q) combination (i.e., a simple SSL with parameters θ_q). For given r, as a consequence, identification boils down to guaranteeing the existence of a unique solution Ψ^r to the system of linear equations in (3.10).

There are Q sufficient set logit probabilities on the right hand side of (3.10). For brevity, we call each of them $\Pr[j|\, X_p, r, \theta_q]$ and re-write (3.10) as:

$$\begin{aligned} \Pr[Y_i = j|\, X_p, f(Y_i) = r, \Theta, \Psi^r] &= \sum_{q=1}^{Q} \Pr[j|\, X_p, r, \theta_q] \times \psi_q^r \\ &= \mathbf{Pr}[j|\, X_p, r, \Theta]\, \Psi^r, \end{aligned} \tag{3.11}$$

where $\mathbf{Pr}[j|\, X_p, r, \Theta] = [\Pr[j|\, X_p, r, \theta_1], \ldots, \Pr[j|\, X_p, r, \theta_Q]]$ and $\Psi^r = \left[\psi_1^r, \ldots, \psi_q^r, \ldots, \psi_Q^r\right]'$. Denote the number of choice sequences in $f(Y_i) = r$ by S_r (i.e., $|r| = S_r$). For each (X_p, r) combination, with $p = 1, \ldots, P$, we have a system of S_r equations like (3.11) and, in turn, we have P such systems of S_r equations (one for each X_p). By stacking all of these equations together, we obtain a (potentially huge) system of $P \cdot S_r$ equations for any r:

$$\begin{aligned} \mathbf{Pr}[\boldsymbol{X}, r, \Theta, \Psi^r] &= \begin{bmatrix} \mathbf{Pr}[X_1, r, \Theta] \\ \vdots \\ \mathbf{Pr}[X_p, r, \Theta] \\ \vdots \\ \mathbf{Pr}[X_P, r, \Theta] \end{bmatrix} \Psi^r \\ &= \mathbf{Pr}[\boldsymbol{X}, r, \Theta]\, \Psi^r, \end{aligned} \tag{3.12}$$

where $\boldsymbol{X} = [X_1, \ldots, X_p, \ldots, X_P]$, $\mathbf{Pr}[\boldsymbol{X}, r, \Theta, \Psi^r]$ is the $P \cdot S_r \times 1$ vector that stacks together all the observed $\Pr[Y_i = j|\, X_p, f(Y_i) = r, \Theta, \Psi^r]$ conditional choice probabilities, for all X_p's and all $j \in f(Y_i) = r$. For any $p = 1, \ldots, P$ $\mathbf{Pr}[X_p, r, \Theta]$ is a $S_r \times Q$ matrix with rows given by $\mathbf{Pr}[j|\, X_p, r, \Theta]$, for $j = 1, \ldots, S_r$. $\mathbf{Pr}[\boldsymbol{X}, r, \Theta]$ is the $P \cdot S_r \times Q$ matrix that stacks together all the P matrices $\mathbf{Pr}[X_p, r, \Theta]$ with $p = 1, \ldots, P$.

It is then immediate to see that given Assumptions 2(a), 2(b), and Condition 1, the discrete conditional distribution of random coefficients Ψ^r, conditional on sufficient set $f(Y_i) = r$, is identified whenever $\mathbf{Pr}[\boldsymbol{X}, r, \Theta]$ is of full column rank:

$$\Psi^r = \left(\mathbf{Pr}[\boldsymbol{X}, r, \Theta]' \mathbf{Pr}[\boldsymbol{X}, r, \Theta]\right)^{-1} \mathbf{Pr}[\boldsymbol{X}, r, \Theta]' \mathbf{Pr}[\boldsymbol{X}, r, \Theta, \Psi^r]. \tag{3.13}$$

This full column rank condition is sufficient but not necessary for the identification of Ψ^r.[16] An obvious necessary condition for $\mathbf{Pr}[\boldsymbol{X}, r, \Theta]$ to be of full column rank is $Q \leq P \cdot S_r$. In other words, the rank condition embedded in (3.13) requires one to have "enough" measurements in the sense of many potential choice sequences (a high S_r) and large variation in the regressors (a high P). Intuitively, when $P \cdot S_r$ is high, then it is "easier" to sustain a finer grid of points Θ (a high Q). In addition, at the potential cost of some loss of information (or efficiency in terms of estimation), equation (3.13) can be used in isolation to recover Ψ^r for each or only for some of the realizations of the sufficient set $r = 1, ..., R$.

In practice, these features suggest to focus on the sub-sample of observations corresponding to realizations of the sufficient set r's with high $P \cdot S_r$, so to be able to identify and estimate a fine grid Θ with a high Q. This is important because, given any Q, for all those r's with $Q > P \cdot S_r$, the vector of weights Ψ^r may not be identified (i.e., $\mathbf{Pr}[\boldsymbol{X}, r, \Theta]$ will not be of full column rank). This highlights a trade-off in the choice of the number of grid points Q. On the one hand, the larger the Q the more credible the model, to the extreme of being able to approximate even continuous mixing distributions (see Fox et al. (2016)), but at the cost of having to use potentially only a small part of the full sample, i.e. those individuals with realizations of $f(Y_i)$ with large $P \cdot S_r$. On the other hand, with a small Q one may be able to use a larger portion of the sample, but the risk of misspecification will be higher. This trade-off is salient because only by having an estimate of Ψ^r for all $r = 1, ..., R$ one can recover the *unconditional* distribution of random coefficients, the object typically estimated in standard mixed logit models. This stresses how, by allowing for this further dimension of unobserved heterogeneity without any additional data, one will typically be able to learn less about the distribution of random coefficients.

[16] There are at least two reasons why the full column rank condition behind (3.13) may be stronger than necessary for identification (i.e., it is possible to achieve identification with $\mathbf{Pr}[\boldsymbol{X}, r, \Theta]$ of rank less than Q): the possibility of sparsity in Ψ^r (i.e., $\varphi_q^r = 0$ for some q) and the fact that the correspondence f is the *same* across different realizations $r = 1, ..., R$ (i.e., the *marginal* distribution of the random coefficients (3.9) imposes restrictions across the R realizations of the sufficient set). We leave the investigation of the necessary conditions for the identification of Ψ^r to future work.

Equation (3.13) readily leads to a simple estimator: a separate (and potentially huge) OLS estimator for each $r = 1, ..., R$. Each individual OLS will provide an estimate of the vector of weights Ψ^r (the distribution of random coefficients conditional on $f(Y_i) = r$). In the context of the unconditional mixed logit (with known choice sets), such a simple-to-implement estimator was first proposed by Bajari et al. (2007) and further extended by Fox et al. (2011). To improve efficiency, these papers propose an inequality-constrained least square estimator that complements the OLS with the natural constraints implied by Ψ^r being a vector of probabilities.[17] More recently, a refined version of this estimator that does not require perfect ex-ante knowledge of the grid Θ was proposed by Fox et al. (2016).

3.2.3 Beyond Gumbel Errors: Pairwise Maximum Score Estimator

McFadden (1978) showed that MNL models can be consistently estimated by a conditional Maximum Likelihood Estimator (MLE) using subsets of individuals' true choice sets. More recently, Bierlaire et al. (2008) extended the result to discrete-choice models with block-diagonal Generalized Extreme Value errors. To the best of our knowledge, in the context of cross-sectional or "short" panel data, results of this kind are still not available for the nested logit and for the mixed logit models, even though some interesting approximations have been proposed by Keane and Wasi (2012), Guevara and Ben-Akiva (2013a), and Guevara and Ben-Akiva (2013b).[18] Building on Manski (1975), Fox (2007) extended McFadden (1978) by showing that semi-parametric discrete-choice models can be consistently estimated with a Pairwise Maximum Score Estimator (PMSE) using subsets of individuals' true choice sets. In this subsection we discuss the use of sufficient sets in the context of the PMSE proposed by Fox (2007).

Assumption 3. Suppose that $V(X_{ikt}, \theta) = X_{ikt}\theta$ and that X_{it} is the matrix stacking all the X_{ikt}'s for $k \in CS^{\star}_{it} = c_{it}$. For any given (i, t) and $k, k' \in CS^{\star}_{it} = c_{it}$, $X_{ikt}\theta > X_{ik't}\theta$ if and only if $\Pr[Y_{it} = k | X_{it}, CS^{\star}_{it} = c_{it}, \theta] > \Pr[Y_{it} = k' | X_{it}, CS^{\star}_{it} = c_{it}, \theta]$.

[17] So that for each r: $0 \leq \psi^r_q \leq 1$, $q = 1, ..., Q$ and $\sum_{q=1}^{Q} \psi^r_q = 1$.

[18] The lack of general extensions of McFadden (1978) to mixed logit models motivates our focus on mixed logit models in the context of "long" panel data and with *discrete* distributions of random coefficients when only "short" panel data are available.

Assumption 3 states that the alternatives belonging to (i,t)'s true but unobserved choice set with higher systematic utilities are more likely to be chosen. Note that, differently from parametric models such as the multinomial logit or probit, Assumption 3 does not impose that the distribution of ϵ_{ikt} is the same across individuals or even that the distribution is the same across the choice situations of the same individual (e.g., ϵ_{ikt} could be distributed Laplace while $\epsilon_{ikt'}$ normal). However, Assumption 3 does constrain the joint distribution of ϵ_{ikt} across alternatives for any given (i,t). Goeree et al. (2005) show that a sufficient condition for Assumption 3 is that, for any (i,t), the joint density of the errors across alternatives is *exchangeable*.[19] See Fox (2007) for more details about Assumption 3. The implementation of the PMSE with unobserved and heterogeneous choice sets requires an additional condition on sufficient sets.

Condition 2. Suppose that $f(Y_i) = \times_{t=1}^{T} f_t(Y_i) = \times_{t=1}^{T} f_{it}$ and that there is a non-empty set N of (i,t)'s with $|N| = n \leq I \cdot T$ for which $K = \cap_{(i,t)\in N} f_{it}$ contains at least two alternatives, $|K| \geq 2$.

Condition 2 imposes two restrictions. First, it requires the sufficient set over choice sequences $f(Y_i)$ to be the cartesian product of t-specific sufficient sets f_{it}'s over alternatives. Second, it requires that there is a set of (i,t)'s whose sufficient sets f_{it}'s contain the same two or more alternatives. The PMSE makes pairwise comparisons of alternatives belonging to some subset K for all those (i,t)'s that are known to have originally made choices from some choice set $CS_{it}^{\star}$ such that $K \subseteq CS_{it}^{\star}$. Condition 2 uses sufficient sets to construct a K guaranteed to be strictly included in the true but unobserved choice set $CS_{it}^{\star}$ of every (i,t) belonging to N.[20]

With a slight abuse of notation, we re-label the alternatives in subset K so that $K = \{1, ..., k, ..., K\}$. The PMSE using choice-based data on the subset K of alternatives is the parameter vector $\hat{\theta}_n^K$ that maximizes:

$$Q_n^K(\theta) = \sum_{k=1}^{K-1} \sum_{k'=k+1}^{K} \frac{1}{n} \sum_{(i,t)\in N} (1[Y_{it} = k] \cdot 1[X_{ikt}\theta > X_{ik't}\theta] + 1[Y_{it} = k'] \cdot 1[X_{ik't}\theta > X_{ikt}\theta]). \quad (3.14)$$

[19] Despite the flexibility, there are popular models among applied researchers that violate Assumption 3, such as the mixed logit model.

[20] As we will see below, the sizes of N and K directly affect the number of pairwise comparisons, i.e. observations, used to construct the PMSE.

Given Assumption 3, Condition 1, Condition 2, and some additional technical assumptions (i.e. Assumptions 3 and 4 from Fox (2007) at p.1009 and p.1011), Fox (2007)'s Theorem 1 guarantees that the Pairwise Maximum Score Estimator $\hat{\theta}_n^K$ is consistent for θ.

3.3 Combining Sufficient Sets with the "Integrating Over" Approach

So far we have discussed the use of sufficient sets to specify conditional discrete-choice models that "difference out" unobserved choice sets. Here we discuss how sufficient sets can be used to simplify the practical estimation of *unconditional* discrete-choice models that "integrate over" unobserved choice sets, alleviating the curse of dimensionality embedded in the Manski (1977)'s approach. An early application of this idea can be found in Chiang et al. (1998).

As discussed in subsection 3.1, even when model (3.1) is correctly specified and identified, its estimation can prove difficult. In particular, estimation of model (3.1) is likely to suffer from a curse of dimensionality because the number of elements in $C_i^\star$ grows exponentially in the number of alternatives J_i available to individual i (i.e., $|C_i^\star| = 2^{J_i} - 1$). A direct consequence of this curse of dimensionality is that, unless the researcher makes strong functional form assumptions on $\Pr[\mathcal{CS}_i^\star = c|\gamma]$, the model can be estimated only when J_i is small (e.g., Abaluck and Adams (2017) limit their estimations to the case of J_i=10). Sufficient sets can provide additional restrictions on model (3.1) and make its estimation more tractable for any given J_i (or even possible if J_i is large). If Condition 1 is satisfied, $f(Y_i) = r \subseteq \mathcal{CS}_i^\star = c$, where c is the true set of choice sequences to which i is matched. It therefore follows that any set of choice sequences $c' \in C_i^\star$ such that $f(Y_i) = r \not\subseteq \mathcal{CS}_i^\star = c'$ cannot be the set of choice sequences to which i is matched, so that $\Pr[\mathcal{CS}_i^\star = c'|\gamma]$ must be zero. In other words, the researcher knows that i's true but unobserved choice set must contain all the choice sequences in the sufficient set, and consequently any candidate choice set that does not include even just one of these choice sequences can be removed from the collection of possible sets of choice sequences $C_i^\star$. Let $C_{f(Y_i)} = \{c|f(Y_i) = r \subseteq \mathcal{CS}_i^\star = c\}$ be the collection of choice sets consistent with $f(Y_i) = r$ (i.e., that include r). Then model (3.1) simplifies to:

$$\Pr[Y_i = j|\theta, \gamma] = \sum_{c \in C_{f(Y_i)}} \Pr[Y_i = j|\mathcal{CS}_i^\star = c, \theta] \times \Pr[\mathcal{CS}_i^\star = c|\gamma]. \tag{3.15}$$

where the only difference with (3.1) is in the terms included in the summation. Note that $C_{f(Y_i)}$ will typically be substantially smaller than the unrestricted $C_i^\star$. For example, suppose that there are four possible choice sequences: a, b, c, and d. Depending on their observed choice sequence Y_i, individual i will have a sufficient set of one of four possible sizes: $|f(Y_i)| = 1$ (e.g., $f(Y_i) = \{a\}$, $f(Y_i) = \{b\}$, etc.), $|f(Y_i)| = 2$ (e.g., $f(Y_i) = \{a, b\}$, $f(Y_i) = \{b, c\}$, etc.), $|f(Y_i)| = 3$ (e.g., $f(Y_i) = \{a, b, c\}$, $f(Y_i) = \{b, c, d\}$, etc.), or $|f(Y_i)| = 4$ (i.e., $f(Y_i) = \{a, b, c, d\}$). The collection $C_i^\star$, usually specified as the the power set of $\{a, b, c, d\}$, will then contain $2^4 - 1 = 15$ possible (non-empty) choice sets. However, $C_{f(Y_i)}$ will only contain: 8 choice sets if $|f(Y_i)| = 1$, 4 choice sets if $|f(Y_i)| = 2$, 2 choice sets if $|f(Y_i)| = 3$, and 1 choice set if $f(Y_i) = \{a, b, c, d\}$. Importantly, note that this use of the sufficient sets is quite general and does not rely on the specific functional form assumptions made by the researcher in specifying model (3.1).

3.4 Pros and Cons of Each Approach

Each of the two main approaches to the problem of unobserved choice set heterogeneity discussed above presents advantages and disadvantages. While "integrating over" requires additional functional form assumptions and data on the choice set formation process and it is computationally more intensive, it enables researchers to learn about both the preference parameters θ and the choice set formation parameters γ. Learning about both θ and γ may be essential in applications in which the key counterfactuals involve re-matching of choice sets to individuals. Differently, "differencing out" requires less prior knowledge and data on the choice set formation process and is simpler to implement, but it does not allow the estimation of the parameters γ.

Within the differencing out approach, we discussed four models: the Sufficient Set Logit (SSL) model, the Individual Sufficient Set Logit (ISSL) model (which can be used to estimate mixed logit models with non-parametric distributions of random coefficients when $T \to \infty$), the Sufficient Set Mixed Logit (SSML) model, and the semi-parametric model studied by Fox (2007). Despite its simplicity and pedagogical value, the SSL model may be unattractive in many applications because of the IIA property. When data on large T are available, the ISSL may be the most attractive option given its practical simplicity (basically, the estimation of I separate MNL models) while allowing for very flexible distributions of unobserved preferences. However, the necessary requirement of a large T may prevent its use in some applications. When only data on small T are available, the

choice is between the SSML and Fox (2007)'s semi-parametric model. The primary advantage of Fox (2007)'s semi-parametric model is to allow for flexible distributions of unobserved preferences (within the boundaries of Assumption 3, which are however not compatible with the mixed logit model). By contrast, the SSML model requires the distribution of random coefficients to be discrete and its support Θ to be known in advance (see Assumption 2(b)). Despite (or because of) its greater flexibility and robustness in the estimation of the preference parameters θ, however, Fox (2007)'s PMSE does not allow for the calculation of some objects typically of interest to applied researchers. Researchers using any of the models discussed here can evaluate the willingness to pay for alternatives' attributes (see, for example, Bajari et al. (2008) or subsection 6.3 below), but knowledge of $\hat{\theta}_n^K$ does not allow for the evaluation of predicted choice probabilities, marginal effects, price elasticities, and consumer surplus. As we detail in Appendix E, the SSL, ISSL, and SSML models instead lead themselves to natural lower and upper bounds on such objects of interest. Finally, while the SSML model can be easily estimated by a simple OLS or by the inequality-constrained least square estimator proposed by Bajari et al. (2007) and Fox et al. (2011), the implementation of Fox (2007)'s PMSE is not as straightforward.[21]

4 Economic Foundations of Sufficient Sets

In this section, we describe how choice environments that have been analyzed in a wide variety of literatures in economics map into particular sufficient sets $f(Y_i)$ which can be used to implement the estimators discussed in the previous section. The sufficient sets introduced here are examples and the list is not meant to be exhaustive: any $f(Y_i)$ satisfying Conditions 1 (and 2 possibly) gives rise to valid estimators, however different from the sufficient sets described in what follows. In Appendix F, we illustrate the practical performance of some of the sufficient sets discussed in this section when choice sets are generated by models of screening on product characteristics (such as price) and of costly search over alternatives.

[21] To facilitate its use, Jeremy Fox and David Santiago share on their personal webpages Mathematica codes for its implementation.

4.1 Stable Choice Sets

4.1.1 Examples from the Literature

Fixed-Sample Search. In an influential paper in the search literature, Morgan and Manning (1985) present general results on the existence and properties of expected-utility maximizing search rules for dynamic search problems in which individuals may choose both the number of periods in which samples of alternatives are searched and the size of the sample searched in each period. Individuals conduct searches over T periods. In a non-sequential or fixed-sample search strategy, individuals do all their search in the first period, construct a choice set, and then make sequences of T choices from this choice set. The authors show that if individuals have "full recall," i.e. once alternatives are searched, individuals do not forget they exist until period T, and "no lost alternatives," i.e. alternatives are not removed from the market until period T, then a fixed-sample search strategy is optimal if either the marginal cost of searching or individuals' discount factors are sufficiently high. Intuitively, fixed-sample search strategies are appealing when individuals find it more advantageous to gather information quickly, because the search results are observed with some delay and the cost of waiting is high.

Fixed-sample search strategies have been studied in both product and labor markets. For example, Janssen and Moraga-González (2004) study oligopolistic markets characterized by individuals who engage in costly fixed-sample searches for the best prices, while De los Santos et al. (2012) find evidence in support of fixed-sample search strategies in the online market for books from data on the web browsing and purchasing behavior of a large panel of individuals.

Whenever a fixed-sample search strategy is optimal, unobserved choice sets will be stable across the T choice situations for each i, but potentially different across i's: i.e., $CS^{\star}_{it} = CS^{\star}_{it'}$, for all $t \neq t'$. For brevity, in this case we will refer to $CS^{\star}_{it}$, for any t, simply as to $CS^{\star}_{i}$. Any alternative purchased by i in any t, Y_{it}, is guaranteed to belong to $CS^{\star}_{i}$.

Referral networks. Another context in which the assumption of stable choice sets plausibly applies is for specialists' referral networks. For example, Gaynor et al. (2016) study the impact of a regulatory policy that expanded the set of hospitals to which a physician could refer a patient in the English National Health Service (NHS). They consider the post-reform period and assume that in this period,

a physician's choice of referral hospitals is unconstrained, showing that this allows them to identify preference parameters. They then use these and apply them to the pre-reform period to learn about the impact of the pre-reform constraint on physicians' referral choices.

The authors assume that physicians' referral networks are stable over time. As such, and similar to the example of fixed-sample search above, $CS^{\star}_{it} = CS^{\star}_{i}$, $\forall\ t$ and any hospital chosen by i in any t, Y_{it}, is guaranteed to belong to $CS^{\star}_{i}$.

School Choice. In many countries, the set of public schools that parents can consider to send their children for T years is constrained by the neighbourhood where they live and by various allocation mechanisms such as the Gale-Shapley deferred acceptance mechanism (e.g., Abdulkadiroglu and Sönmez (2003)). Usually, the set of schools in any neighbourhood does not evolve rapidly, and those households that do not change neighbourhood over the T years will face a stable set of schools for their children. In this context, Walters (2014) investigates the demand for charter middle schools in Boston, while Fack et al. (2015) study the demand for high schools in the southern district of Paris. Similar to the two previous examples above, the set of schools faced by household i in academic year t can be assumed to be stable $\forall\ t$, $CS^{\star}_{it} = CS^{\star}_{i}$, and any school chosen by i for their children in any t, Y_{it}, is guaranteed to belong to $CS^{\star}_{i}$.

4.1.2 Sufficient Sets Consistent with Stable Choice Sets

Full Purchase History Sufficient Set. Consistent with the three examples above, suppose that individuals' choice sets are potentially heterogeneous across i's but stable over the T choice situations, $CS^{\star}_{it} = CS^{\star}_{i}$. Let $H_i = \bigcup_{t=1}^{T} \{Y_{it}\} \subseteq CS^{\star}_{i}$ be the collection of all the alternatives that individual i is observed to choose in any of the T choice situations. We define the Full Purchase History (FPH) sufficient set as $f_{FPH}(Y_i) = (H_i)^T$, the set of choice sequences given by the cartesian product of H_i in each of the T choice situations. Note that $f_{FPH}(Y_i)$ implies SSL, ISSL, and SSML models along the lines of (3.6) and (3.7) which are very simple to implement.

Practically, the FPH sufficient set assumes that an individual that chooses a collection of alternatives H_i over a given sequence of T choice situations is familiar with all of those alternatives in all of those choice situations. The intuition of this approach, evident in the denominator of (3.6) for the case of SSL models, is that the researcher exploits variation in the characteristics of these alternatives

over choice situations, and the differences in choices made by each i over choice situations, to estimate preference parameters θ in the absence of information about i's true choice set, $CS_i^\star$.

Further pursuing this point, note that each individual may have considered other alternatives beyond those included in her FPH sufficient set, $f_{FPH}(Y_i)$, that were ultimately not chosen. All is required is that a sufficient set is a *subset* of an individual's true choice set (as stated by Condition 1 for SSL, ISSL, and SSML models and by Conditions 1 and 2 for semi-parametric models). Note also that, to simplify exposition, we have assumed that choice sets are stable across all T choice situations of each individual, but this is not necessary. FPH sufficient sets will be effective in differencing out choice sets whenever one observes at least two (different) purchase decisions by the same individual from the same choice set. More generally, i's full choice sequence of length T can be divided into sub-sequences of length of at least two. Then, the assumption of stable choice sets implies fixed choice sets within each sub-sequence, but potentially different choice sets between the different sub-sequences of individual i. In the next section, we describe how to form specification tests to check the length of the sequence over which choice sets are plausibly stable.

Choice Permutations Sufficient Set. In addition to the FPH sufficient set, the assumption of stable choice sets also underpins another sufficient set: the sufficient set proposed by Chamberlain (1980) for the classic Fixed Effect logit model (FE logit). In a model with systematic utilities given by $V_i(X_{ij_tt}, \theta) = \delta_{ij_t} + X_{ij_tt}\beta$, Chamberlain (1980) shows that β can be consistently estimated by the ML estimator of a SSL model with sufficient set $f_{CP}(Y_i) = \mathcal{P}(Y_i)$: the set of all possible permutations of observed choice sequence Y_i.[22] As such, we call this the Choice Permutations (CP) sufficient set and the corresponding model the CP SSL.

Matejka and McKay (2015) propose a choice model with rational inattention in which the reduced form choice probabilities take the form of a CP SSL model (see Theorem 1, p.282). In this model, individuals have priors about the indirect utilities associated with consuming any alternative in $CS_i^\star$ and, before choosing an alternative, can decide to gather more precise payoff-relevant information at a cost. In the reduced form, the individual-alternative specific fixed effects δ_{ij_t}'s are not functions of i's preferences, but rather of the cost of gathering information and of i's priors.[23] As a special

[22] For example, if i is observed to choose alternatives 1, 3, and 5 in choice situations 1, 2, and 3, the $f_{CP}(Y_i)$ sufficient set will be the collection of all possible permutations of sequence (1, 3, 5), i.e. (1, 3, 5), (1, 5, 3), (3, 1, 5), (3, 5, 1), etc.

[23] Recall j indexes the full sequence of i's choices and j_t indexes the choice made in the t^{th} choice situation.

case, the reduced form choice probabilities simplify to a MNL model when i's priors are completely uninformative (i.e., each alternative in $CS_i^\star$ is perceived to generate the same level of indirect utility).

Chamberlain (1980)'s expressed motivation for the sufficient set $f_{CP}(Y_i) = \mathcal{P}(Y_i)$ was to difference out the fixed effects (δ_{ij_t}) from each individual's systematic utility. But his assumption of choice set stability also implies that $f_{CP}(Y_i) \subseteq \mathcal{CS}_i^\star$. As such, sufficient set $f_{CP}(Y_i) = \mathcal{P}(Y_i)$ will not only accommodate unobserved preference heterogeneity in the form of individual-alternative specific fixed effects, but also unobserved choice set heterogeneity.[24]

While this is a significant benefit, the CP SSL also comes with meaningful costs. To obtain predicted choice probabilities and their functions, such as elasticities, researchers typically need to be able to identify the whole vector of preference parameters θ. The CP SSL does not, however, usually allow the identification of i's fixed effects δ_{ij_t}'s, but only those elements of β associated with time-varying observables. This can limit its usefulness to applied researchers. By contrast, the SSL and SSML models obtained from $f_{FPH}(Y_i)$, while relying on the same assumption of choice set stability, typically allow the identification of all parameters.

Note that the CP sufficient set, $f_{CP}(Y_i)$, *cannot* be expressed as the cartesian product of t-specific sufficient sets, giving rise to models that are harder to implement (e.g., the CP SSL model can be expressed as in (3.5) but not as in (3.6)). For those cases where T is large and/or there is substantial heterogeneity in the alternatives chosen across the T choice situations, the computational burden implied by the CP SSL can be considerable. D'Haultfœuille and Iaria (2016) show how to ease this computational burden by applying the insights of McFadden (1978) to the estimation of β from (uniform) random subsets of $f_{CP}(Y_i)$.

4.2 Growing Choice Sets

4.2.1 Examples from the Literature

Sequential Search. Morgan and Manning (1985) show that, in the general context described in subsection 4.1, if the assumptions ensuring "full recall' and "no lost alternatives" hold, then any sequential search strategy over T periods will imply choice sets that are weakly growing over time, so that $CS_{it}^\star \subseteq CS_{it+1}^\star$.

[24] The argument is essentially identical to that leading to (3.5), except for the different systematic utilities that in Chamberlain (1980) have individual-alternative specific coefficients. By replacing $V(X_{ij_tt}, \theta)$ with $\delta_{ij_t} + X_{ij_tt}\beta$ and $f(Y_i)$ with $\mathcal{P}(Y_i)$ in equation (3.5), Chamberlain (1980)'s result follows.

Caplin and Dean (2011) propose two models of sequential search: the alternative-based search (ABS) model and the reservation-based search (RBS) model. The ABS model "captures the process of sequential search with [full] recall, in which the [individual] evaluates an ever-expanding set of objects, choosing at all times the best object thus far identified" (Caplin and Dean (2011), p.23). This model provides the micro foundations underlying some of the functional form restrictions used by Goeree (2008), Manzini and Mariotti (2014), and Abaluck and Adams (2017) to aid the identification of Manski (1977)'s model. By contrast, the RBS model is a formalization of Simon (1955)'s satisficing model. In related work, Caplin et al. (2011) find experimental evidence in support of this model.

In these settings, a sequential search strategy implies that alternatives observed to be chosen by the individual in the (recent) past are in their choice set and can be used to form a sufficient set. We describe such a sufficient set after introducing other examples.

Choice by Iterative Search. Masatlioglu and Nakajima (2013) propose another dynamic search framework that they call Choice by Iterative Search (CIS). In each period t, the history of search until $t-1$, also called the "status quo" (i.e., the set of alternatives searched so far), and the feasible set of alternatives both can affect the evolution of the choice set $CS^{\star}_{it}$ in arbitrary ways. However, if assumptions analogous to "full recall" and "no lost alternatives" hold, then $CS^{\star}_{it}$ will coincide with next period's status quo and the sequence of status quos will weakly grow over time, so that $CS^{\star}_{it} \subseteq CS^{\star}_{it+1}$.

Several models in the fast-growing literature on limited attention build on the CIS framework. An example is Eliaz and Spiegler (2011), who study a setting in which individuals have a singleton status quo, i.e. a choice set only including one product (possibly different across individuals), and firms seek to use marketing devices, e.g. advertising, to include their products in individuals' choice sets.[25] Preferences are themselves unaffected by such advertising; it only influences the alternatives individuals include in their choice sets. While Eliaz and Spiegler (2011) only consider a static environment, one could imagine a dynamic extension within the CIS framework, in which multiple firms compete in each period with advertising to encourage individuals to consider their products and status quos evolve over time as in Masatlioglu and Nakajima (2013). In such a setting, again the alternatives an

[25]Other relevant examples of the application of the CIS framework are Ho et al. (2015) and Heiss et al. (2016). Both papers study the Medicare Part D program and document that individuals switch health plans infrequently and search imperfectly, possibly because of high search costs.

individual has purchased in the (recent) past will be in their choice set and can be used to form a sufficient set.[26]

Focus. Another example related to the CIS framework is the work of Kőszegi and Szeidl (2013), who analyze the impact of "focus" on individuals' choices. They provide numerous examples of individuals focusing on one of an alternative's (possibly many) attributes, leading them to select an alternative that exceeds others in this attribute, even if a comparison of the alternatives across all attributes would lead to a different choice.[27]

Formally they model this as individuals making choices from subsets of their full choice sets. They motivate these subsets using conjunctive and disjunctive screening rules like those from Gilbride and Allenby (2004) (Kőszegi and Szeidl, 2013, p.61). They then apply their model of focus to inter-temporal choices (e.g. consumption-savings decisions).

As they describe themselves, "Formally, there are T periods and in period t, a consumer makes a choice x_t from the deterministic ... consideration set $X_t(h_{t-1})$, where $h_{t-1} = (x_1, \ldots, x_{t-1})$ is the history of choices up to period $t-1$." While in a consumption-savings environment, past decisions limit current choices by means of a summary statistic (e.g., how much income an individual has in period t), in a repeat-purchase environment (e.g., retail purchases of household goods), h_{t-1} would consist of the history of that individual's previous purchase decisions, a fact that can be used to form a sufficient set as we describe next.

4.2.2 The Past Purchase History Sufficient Set

The three examples above suggest the following sufficient set. Let $H_{it} = \bigcup_{b=1}^{t} \{Y_{ib}\} \subseteq CS_{it}^{\star}$ be the collection of all the alternatives that individual i is observed to choose between choice situation 1 and t. We define the Past Purchase History (PPH) sufficient set as $f_{PPH}(Y_i) = \times_{t=1}^{T} H_{it}$, the cartesian product of H_{it} between choice situation 1 and T.[28] Note that, similarly to the Full Purchase History

[26] This is the framework which motivates our empirical application in section 6.

[27] For example, a person comparing the quality of life in California with that in the Midwest may focus more on climate than other aspects of life satisfaction in which the two areas are more similar (e.g. crime rate, availability of public goods, etc.), and therefore be too likely to believe that California is a better place to live.

[28] An analytically related sufficient set to f_{PPH} is the sufficient set compatible with choice sets that are weakly *shrinking*, rather than growing, over choice situations. This can be obtained by just "turning around" the choice situations of each choice sequence, so to have them re-ordered from T to 1, and then by applying the same definition of f_{PPH} to the re-ordered choice sequences.

sufficient set, $f_{PPH}(Y_i)$ also implies SSL, ISSL, and SSML models along the lines of (3.6) and (3.7), which are very simple to implement.

Practically, the PPH sufficient set assumes that an individual has in their choice set the set of alternatives chosen between some beginning period and the current one. Importantly, alternatives are only assumed to be in i's choice set after they are observed to have been chosen. As with the FPH sufficient set, the intuition is to exploit the variation in the characteristics of only these alternatives over time. As for the FPH sufficient set, any individual i's full choice sequence $(Y_{i1}, ..., Y_{iT})$ can be divided into sub-sequences of length of at least two. The assumption here allows for the possibility of choice sets that are growing within each sub-sequence, but with potentially different choice sets between the different sub-sequences for the same individual. In the next section, we describe how to form specification tests to check the length of the sequence over which choice sets are allowed to weakly grow.

4.3 Other Sufficient Sets

Cross-Sectional Environments. Each of the examples above assumed a panel of individuals making decisions in multiple choice situations. Here we illustrate how sufficient sets can also be constructed in cross-sectional environments to a group of individuals, each making a separate purchase decision at a single point in time from the same choice set.

As a motivating example, consider the question of whether greater availability of fast food outlets causes obesity as in Currie et al. (2010). One would like to be able to identify whether it is the availability of fast food outlets that leads to increased consumption or whether preferences are the driving factor. The authors collected precise geographic data on the location of fast food outlets and where children live and attend school and examined the effect of the presence of a fast food restaurant within 0.1, 0.25, and 0.5 miles of the school attended by the student. While this is reasonable, this definition did not exploit the location of each child's home and, even if it did, one cannot be sure of exactly which outlets lie within individual children's choice sets. By contrast, if the authors were willing to assume that all children living on the same street *and* attending the same school faced the same choice set, they could conclude that all such outlets were in the choice set for all such children and this could form the basis for a sufficient set in our approach.

More generally, in a cross-sectional environment one can call each i a "consumer type" and each t one of the T individuals of that type.[29] Then, when the same choice set is faced by the T individuals of the same consumer type i, the econometrician can use the Inter-Personal sufficient set: $f_{IP}(Y_i) = (H_i)^T$, where $H_i = \bigcup_{t=1}^{T} \{Y_{it}\} \subseteq CS^{\star}_{it}$.[30] The sufficient set $f_{IP}(Y_i)$ imputes to each individual t the collection of all the alternatives observed to be chosen by any of the T individuals of consumer type i. Note the validity of this sufficient set further relies on the assumption that the observable characteristics of any product j are the same for each of the T individuals of consumer type i, or that the econometrician knows how they change across t's.[31] As for $f_{FPH}(Y_i)$ and $f_{PPH}(Y_i)$, note that $f_{IP}(Y_i)$ also implies SSL, ISSL, and SSML models along the lines of (3.6) and (3.7), which are very simple to implement.

Combining sufficient sets. The sufficient sets described above are neither mutually exclusive nor exhaustive. Regarding the first point, if one had panel data and a stable choice environment (suggesting the application of the FPH sufficient set) and some subset of the individuals faced the same choice set in each period (suggesting the application of the IP sufficient set), one could combine the sufficient sets into a Inter-Personal Full Purchase History sufficient set. This is also more generally true: if multiple sufficient set definitions apply, then one can use the intersection of each sufficient set to form a new, composite, sufficient set. In the empirical application in section 6, we rely on composite sufficient sets that combine an individual's past purchases in different types of retail stores, i.e. in our store-type-specific Past Purchase History sufficient sets, each i has a separate PPH sufficient set that is specific to each of the store-types in our data.

[29] This is without loss of generality. We could allow each type to have a different number of individuals, T_i, but this would only complicate notation and provide no deeper insights into the underlying mechanism at work.

[30] Note that this definition of $f_{IP}(Y_i)$ is identical to that for the Full Purchase History sufficient set, but because the underlying economic environments are so different (e.g. i is an individual and t is a time period in $f_{FPH}(Y_i)$, while i is a consumer type and t is an individual in $f_{FPH}(Y_i)$), we prefer to define the two separately.

[31] For example, if individual t and t' of consumer type i are observed purchasing, respectively, product j at price p_{ijt} and product k at price $p_{ikt'}$, then we assume that each individual could have purchased the product bought by the other at the same price, i.e. $p_{ijt} = p_{ijt'} = p_{ij}$ and $p_{ikt} = p_{ikt'} = p_{ik}$. If instead different products have observable characteristics that take different values for different t's of type i, say driving distance d_{ijt} between j and t in a model of supermarket choice, then the econometrician needs to be able to compute $d_{ijt'}$ for any other individual t' of type i.

4.3.1 Discussion

It is important to note that the examples provided in this section are meant to illustrate possible ways of linking sufficient sets to popular economic environments and datasets. In general, *any set* that somehow combines i's choices across t's can in principle serve as a sufficient set.

Typically, there may be some choice sets that vary in predictable ways due to idiosyncracies of someone's choice environment. For example, an individual might choose from a different set of alternatives for lunch on Monday, when they are working at the office, than on Tuesdays when they are working from home. This could naturally be accommodated in the definition of sufficient set by conditioning on past purchases made in that state (e.g., lunch choices made on past Mondays instead of lunch choices made any day of the week).

More generally, information available to the researcher can and should be incorporated into the definition of sufficient sets. So, generalizing Eliaz and Spiegler (2011) and considering how to measure the impact of an in-store advertising campaign for a new alternative, if one had information about both each individual's past purchase history as well as their exposure to and awareness of the advertising campaign for the new alternative *and* one was willing to make the assumption that exposure and awareness to the campaign meant the new alternative was in an individual's choice set, one could use each individual's past purchase history to form an (initial) sufficient set for that individual and then augment it with the new alternative for those individuals known to have been exposed to the advertisement.

We conclude this section summarizing the results of some Monte Carlo simulations. In Appendix F, we report the results of simulations evaluating the practical performance of MNL and SSL models in the presence of various forms of unobserved choice set heterogeneity. In a first set of experiments, we directly vary the extent of choice set heterogeneity by randomly removing alternatives from choice sets, independently of the indirect utilities or the product characteristics of the removed alternatives. In a second set of experiments, we implement two more economically relevant choice set formation processes along the lines of the above discussion: a model of screening on product characteristics (such as price) and a model of costly search. There, our aim is to illustrate that even when choice set heterogeneity is the outcome of selection processes involving the alternatives' systematic utilities and/or product characteristics, the proposed SSL models work well without requiring the econometrician to know much about such possibly complex processes.

5 Specification Tests: Choice Set Stability and IIA

The correct implementation of the differencing out approach relies on two kinds of assumptions: assumptions about the evolution of choice sets across choice situations and assumptions about unobserved preference heterogeneity. In the previous two sections, we discussed examples of choice set formation processes giving rise to sufficient sets compatible with Conditions 1 and 2, and different models that rely on the IIA property to different extents. In what follows, we illustrate how existing testing procedures can be used to discriminate among alternative choice set formation processes and various departures from the IIA property.

5.1 Testing Among Competing Sufficient Sets

In the context of the ML estimator of the SSL and the ISSL models, alternative sufficient sets lead to more or less robust and/or efficient estimators along the lines of Hausman and McFadden (1984) and can be used to form specification tests. We discuss here how to test for some of the assumptions implicit in several sufficient sets, such as the length of the sequence of choice situations for which choice sets are stable or grow.

The basis for these specification tests is the *Factorization Theorem* proposed by Ruud (1984) and further explored by Hausman and Ruud (1987). Ruud (1984)'s result enables one to "rank" Maximum Likelihood Estimators (MLEs) of SSL or ISSL models with different sufficient sets in terms of their efficiency: the MLE of a SSL or ISSL model with sufficient set f_L is more efficient than the MLE of a SSL or ISSL model with sufficient set $f_Z \subset f_L$. This result can be applied recursively, so that if two subsets of f_L are available, say f_Z and f_{XZ} with $f_{XZ} \subset f_Z$, then the efficiency rank of the three MLEs will be $f_L \succ f_Z \succ f_{XZ}$. As we detail in Appendix G, building on the Factorization Theorem one can construct Hausman tests between SSL or ISSL models based on different sufficient sets and implicitly test for underlying economic assumptions such as choice set stability or the IIA property. For example, in the context of the SSL model, the sufficient sets discussed earlier rely on the following economic assumptions:

- f_{CP}: Choice set stability across T choice situations *and* the possibility of IIA violations in the form of individual-alternative specific fixed effects, δ_{ij_t}.
- f_{FPH}: Choice set stability across T choice situations *and* the IIA property.

- f_{PPH}: Choice set evolution in the form of weakly growing choice sets (or, symmetrically, weakly shrinking choice sets) across T choice situations *and* the IIA property.[32]

The first possibility is to compare f_{CP}, f_{FPH}, and f_{PPH} for choice sequences of constant length T. In this case, both the CP and PPH sufficient sets are subsets of the FPH sufficient set: $f_{CP}(Y_i) \subset f_{FPH}(Y_i)$ and $f_{PPH}(Y_i) \subset f_{FPH}(Y_i)$ for any $Y_i \in \mathcal{CS}_i^\star = c$. As we discuss in Appendix G.1, these relationships can be used to test for the assumption of choice set stability and for violations of the IIA property.

The second possibility is to fix a specific f, say f_{CP}, and to compare choice sequences with some of their *sub*-sequences: for example, the sequence $1, 2, \ldots, T^L$ can be split into two mutually exclusive sub-sequences $1, 2, \ldots, T^Z$ and $T^Z + 1, \ldots, T^L$, and this gives rise to different f_{CP}'s, f_{CP}^Z (separately from 1 to T^Z and from $T^Z + 1$ to T^L) and f_{CP}^L (from 1 to T^L) such that $f_{CP}^Z(Y_i) \subset f_{CP}^L(Y_i)$ for any $Y_i \in \mathcal{CS}_i^\star = c$. The same holds both for f_{FPH} and f_{PPH}. As illustrated in Appendix G.1, these comparisons allow one to test for general forms of choice set stability or evolution.

5.2 Testing for Departures from the IIA

A classic simple test for the IIA property proposed by McFadden et al. (1977) involves a comparison between a MNL with its *true* choice set against another MNL with a *restricted* choice set. The specification test described in the previous subsection is based on the same logic but in a more complex environment where true choice sets are not observed. The additional layer of complexity leads to some ambiguity in the classic testing procedure, because rejection of the null can now be motivated by either a failure of the IIA property (as in Hausman and McFadden (1984)) or by the sufficient sets being too large (a violation of Condition 1), or by both simultaneously. As illustrated in subsection 2.2, the imputation in estimation of a choice set that is too "large" (so that Condition 1 does not hold) is mechanically equivalent to a violation of the IIA property. In general, such ambiguity cannot be fully resolved: any testing procedure of this kind will be valid and informative only under some maintained assumptions. At a deeper level, this is a fundamental identification problem: as discussed by McFadden (1987), *any* discrete choice model can be formally re-written as a model satisfying the IIA property, but with a complex dependence on the explanatory variables. We

[32] Importantly, the IIA requirement follows from the SSL model and it is not intrinsic in the f_{FPH} and f_{PPH} sufficient sets. Neither sufficient set relies on the IIA property *across* individuals when employed in more general models such as the ISSL and the SSML discussed in subsection 3.2.2.

illustrate some examples of maintained assumptions necessary for the test to be valid in Appendix G.1. For instance, under the maintained assumption of choice set stability and a specific alternative model of unobserved preferences (i.e., Gumbel errors plus individual-alternative specific fixed effects), one can test for departures from the IIA property by comparing the estimates of a CP SSL versus those of a FPH SSL. Both the CP sufficient set and the FPH sufficient set require choice sets to be stable, but—differently from the FPH SSL—the CP SSL controls for individual-alternative specific fixed effects which may induce violations of the IIA.

A second way to test for departures of the IIA can be based on the mixed logit models discussed in subsection 3.2.2. For any given correctly specified sufficient set, in the ISSL model (3.7) one can check whether the I estimates $\widehat{\theta}_i$'s are statistically indistinguishable across individuals. Similarly, in the SSML model (3.11), for any given correctly specified $f(Y_i) = r$, one can check whether Ψ^r is degenerate, i.e. all but one of the Q probability mass functions ψ_q^r's are equal to zero. A third tool that can be used to further investigate failures of the IIA property is the nested logit version of the testing procedure proposed by Hausman and McFadden (1984), which consists of comparing a nested logit against a MNL, both from the *true* choice set. In Appendix H, we illustrate how under similar assumptions to those required by the MNL, sufficient sets can also be used for the consistent estimation of nested logit models with unobserved choice set heterogeneity. In particular, for any given correctly specified $f(Y_i) = r$, it is possible to consistently estimate the *within-nest* part of a nested logit model at very little additional cost with respect to a MNL model, and this is enough to implement a test for departures of the IIA along the lines of Hausman and McFadden (1984) in the context of unobserved choice set heterogeneity.

6 Empirical Illustration

To show how our ideas can be applied in practice, we present an empirical illustration. In Section 4.2, we discussed models of limited attention and the role that marketing expenditure can play at influencing consumers' choice sets (as in the models of Eliaz and Spiegler (2011) and Goeree (2008)). We use data and methods similar to those in Dubois et al. (2019), who estimate demand for soft drinks, to estimate demand for chocolate bars by a sample of adult women making decisions on-the-

go, i.e. chocolate purchased outside of the home in small corner stores, vending machines, concession stands, and other outlets for immediate consumption.[33]

We are interested in estimating consumers' responsiveness to price and how advertising might affect consumers' choices. Advertising is important in the chocolate market, and there is intuitive appeal to the idea that ads might play an important role both in bringing products to consumers' attention (as in Eliaz and Spiegler (2011) and Goeree (2008)) as well as potentially entering their utility directly (as in Becker and Murphy (1993)).

At any point in time there are more than 100 products available to choose from. In such a choice environment it is unlikely that an individual will spend the time to consider each one, and collecting information on which products the individual considered (for example, using eye-tracking technologies) is expensive. We compare results from estimation based on alternative assumptions on sufficient sets. First, we assume that each individual considers all of the products that are available in the type of store in which they are currently shopping; we call this the **Complete** sufficient set. Second, we assume that each individual considers (at minimum) the products that the individual has purchased in the past and are available in the type of store in which they are currently shopping; we call this the individual's **Past Purchase History (PPH)** sufficient set. Third, we allow for the possibility that individuals have finite memory of products that they have purchased, and consider sufficient sets based on purchase histories of shorter duration (described below). For brevity we omit the "store-type specific" modifier from each of these descriptions.

6.1 Model

We adapt the general mixed logit model presented in Section 3.2.2 to the specifics of on-the-go chocolate demand and our data. We assume that each individual makes a purchase from their own choice set $CS_{it}^{\star}$. This set is not observed. It could include many or only a few of the products currently available in the market; it always includes the option not to purchase. We observe what product was purchased, the price paid, the type of store the product was purchased in, and what products are available in that type of store. See Appendix I for further details.

We rely on the large number of choice situations observed per individual to specify choice probabilities as in the ISSL model (3.7). The probability individual i buys the sequence of products

[33] We focus on adult women because the advertising exposure information in our data is for the main shopper in the household and these are most often adult women.

$j = (j_1, ..., j_t, ..., j_T)$ given her sufficient set $f(Y_i) = r_i$ is given by:

$$\Pr[Y_i = j | f(Y_i) = r_i, \theta_i] = \prod_{t=1}^{T} \frac{\exp(V(X_{ij_t t}, \theta_i))}{\sum_{k \in r_{it}} \exp(V(X_{ikt}, \theta_i))}. \tag{6.1}$$

where $f(Y_i) = \times_{t=1}^{T} f_t(Y_i) = \times_{t=1}^{T} r_{it}$ and each r_{it} is the set of chocolate bars belonging to individual i's sufficient set in week t. Utility for any chocolate bar j_t in week t is given by

$$U_{ij_t t} = V(X_{ij_t t}, \theta_i) + \epsilon_{ij_t t}, \tag{6.2}$$

with

$$V(X_{ij_t t}, \theta_i) = \delta_{gb} + \alpha_i p_{oj_t t} + \beta_g \ln a_{ibt}, \tag{6.3}$$

where δ_{gb} is a brand (to which product j_t belongs) fixed effect for demographic group g, $p_{oj_t t}$ is the price of product j_t in store-type o in week t, and $\ln a_{ibt}$ is log advertising exposure to brand b in week t.[34] The price variable and our measure of advertising exposure are defined in the next subsection. Following Dubois et al. (2019), we allow each individual to have her own price sensitivity, α_i, with a common brand and advertising sensitivity according to their membership in one of nine demographic groups defined by age and equivalised income and indexed by g.

The utility of the outside option of not purchasing a chocolate bar is given by

$$U_{i0t} = \delta_{g0} + \sum_{m} \tau_{gm} + \epsilon_{i0t}, \tag{6.4}$$

where the τ_{gm}'s are demographic-group-specific month effects meant to capture seasonality and/or cyclicality in on-the-go chocolate demand. We estimate versions of ISSL model (6.1) corresponding to three different specifications of $f(Y_i)$.

6.2 Data

We use data on 532 women (ages 25 and older) who are the main shoppers in their household. The data are from the Kantar Worldpanel on-the-go survey, collected from individuals who record purchases

[34] We specific brand dummies for eight large chocolate brands.

that they make on-the-go for immediate consumption.[35] We use information on 130,304 purchase occasions over the period 2009-2014. A purchase occasion is when the woman is observed purchasing a snack of any form on-the-go.

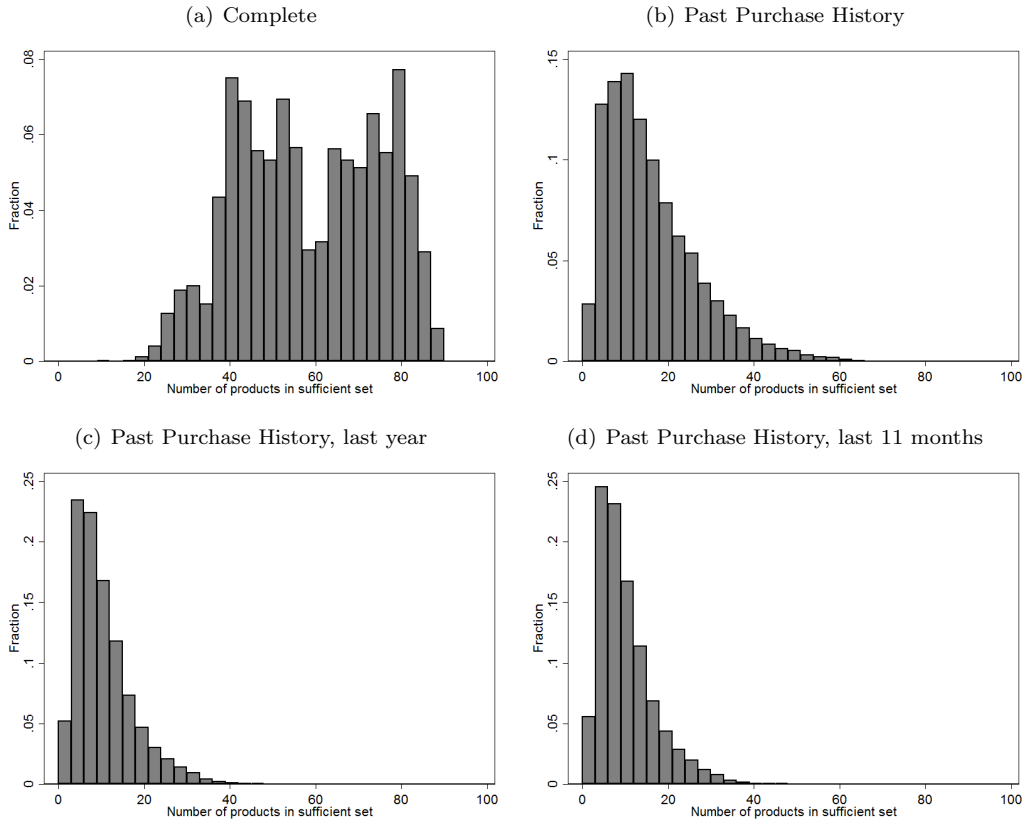
At any one point in time there are more than 100 different types of chocolate products available in the market. The outside option, when a chocolate bar is not purchased, has a 39.6% market share. The three largest market share products are KitKat, with a market share of 3.7%, Cadbury's Twirl, 2.7%, and Cadbury's Dairy Milk, 2.5%.

Individuals purchase products in different outlets. We consider four types of outlets—large national chains (30.1% of sales), news agents (25.2% of sales), vending machines (5.3% of sales), and other types of small stores and outlets (38.6% of sales). We assume that the outlet that we observe the individual shopping in is chosen independently from demand shocks for any specific product. Prices are constructed at the level of the store-type o and week t. We observe prices on each individual transaction and aggregate them to the level of the outlet and week (using the median); most national chains in the UK price nationally, we allow prices in news agents and other outlets to vary across broad regions. 95% of prices range from 20 pence to £1.00, with a few exceptional items available at very low price (for example, Cadburys Dairy Milk Buttons for 19 pence) and a few large items (for example, a 360g Toblerone Milk Chocolate bar for £4.99).

Figure 6.1 shows the size of the sufficient sets used in estimation; panel (a) shows the distribution of the number of chocolate bars in the Complete sufficient set across all purchase occasions. The distribution is bi-modal, with sufficient sets when purchasing from national outlets populating the right mode (up to a maximum of 90 chocolate bars) and sufficient sets when purchasing from a vending machine populating the bulk of the left tail. Panel (b) shows the distribution of the number of chocolate bars in Past Purchase History (PPH) sufficient sets; these range from 2 to 64. Panel (c) shows the distribution for the Past Purchase History using only purchases made in the year before the current choice occasion, and panel (d) using only those made in the 11 months prior to the current choice occasion—this reduces the sufficient sets to a maximum of 48 products.

[35] These data were used to analyze the effects of banning advertising in the market for junk foods in Dubois et al. (2016) and in Dubois et al. (2019) to study the impact of soda taxes; we follow their lead in many aspects of our data construction.

Figure 6.1: *Number of Products in Sufficient Sets*



Note: The histograms shows the distribution of the number of products in the sufficient sets across all purchase occasions.

To measure advertising exposure we convert weekly advertising ("flows") into an advertising "stock;" advertising stocks are the depreciated accumulation of the flows. We use minutes of TV advertising to define advertising flows. Following Goeree (2008), we measure advertising exposure at the *individual* level. We use detailed information about when individual ads were aired on television matched with self-reported viewing information. We denote the stock of advertising $stock_{ibt}$. $stock_{ibt}$ ranges from 0 for individuals that do not watch TV, or only watch advertising-free public TV (the BBC), to over 100 minutes of accumulated exposure to advertisements for a particular brand. The mean is 10 minutes of accumulated exposure. Finally, we follow Dubé et al. (2005) and allow for diminishing returns to advertising by transforming the stock of advertising, $stock_{ibt}$, using the log inverse hyperbolic sine function, $\ln a_{ibt} = ln\left(stock_{ibt} + \sqrt{1 + stock_{ibt}^2}\right)$. Further details are available in Appendix I.

6.3 Coefficient estimates

Table 6.1 presents the mean and standard deviation of the estimated price and advertising coefficients using each of the four sufficient sets.[36] The mean of the coefficient on price reduces substantially from the Complete sufficient set to the Past Purchase History, and reduces again when we use only information on purchases made in the year prior to the current choice occasion; restricting to using only the past 11 months does not substantially change the mean. The standard deviation of the individual estimates is smaller for the estimates using the Past Purchase History. Similarly, for the advertising coefficients, the mean of the estimates is higher when using the Complete sufficient set than when using the Past Purchase Histories.

Table 6.1: *Coefficient Estimates*

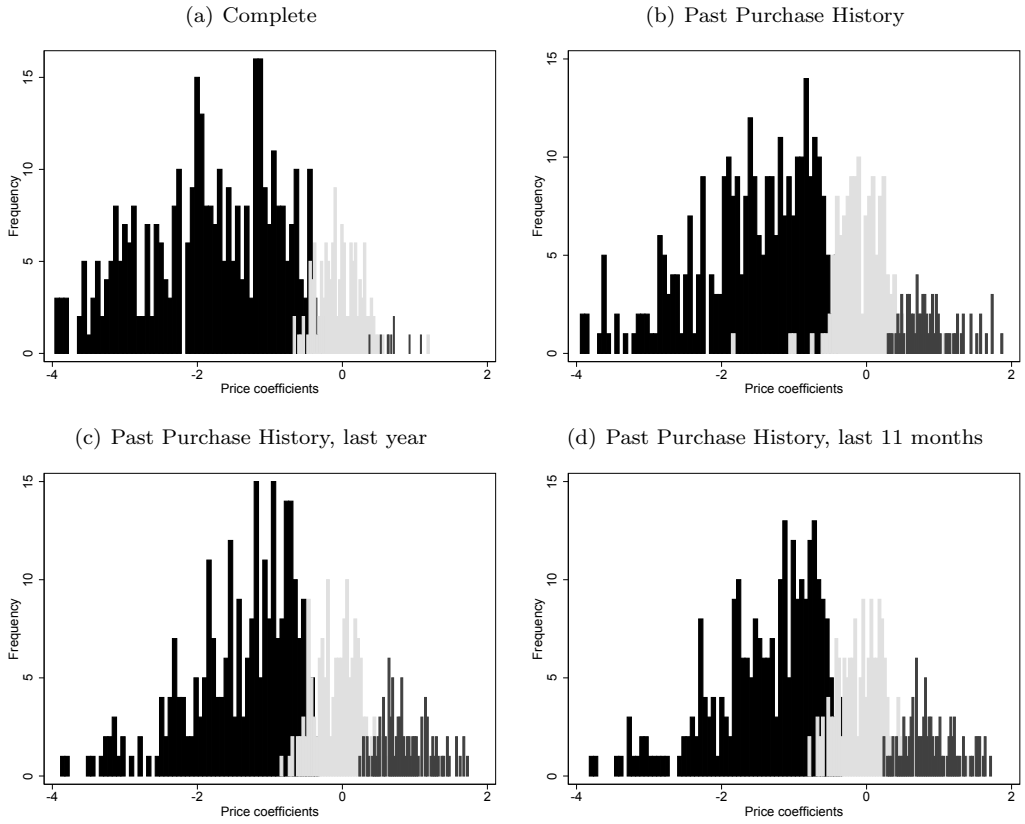
		Complete	PPH	PPH 1 year	PPH 11 months
Price	Mean	-1.777	-0.921	-0.643	-0.623
	Std Dev	5.154	1.822	1.759	1.845
Advertising	Mean	0.168	0.054	0.033	0.031
	Std Dev	0.047	0.036	0.024	0.022
product effects		yes	yes	yes	yes
time effects		yes	yes	yes	yes

Notes: The Table describes the mean and standard deviation of the 532 individual price and 532 advertising parameters. Each column shows the estimates using the indicated sufficient set.

Figures 6.2 shows the distribution of 532 estimated price coefficients across the four sufficient sets. For any individual i, the Complete sufficient set is a superset of the full PPH and the "full" PPH is a superset of the "1 year" PPH, which in turn is a superset of the "11 months" PPH sufficient set. It is evident that assumptions on individuals' choice sets have a material impact on these estimated distributions. This chain of inclusions enables us to perform Hausman tests as discussed in Section 5.1

[36] We excluded the results for a small number of women in our sample for which there is not sufficient price variation in their sufficient sets to identify all of the price coefficients, and a small number for whom the estimated price sensitivity in all specifications was positive. Including them in the analysis would change none of the qualitative conclusions drawn from this illustration.

Figure 6.2: *Distributions of Estimated Price Coefficients*



Note: The histograms show the distribution of estimated price coefficients for our sample of adult women for specifications using different sufficient sets. Bars in black are statistically significant and negative, in light gray are statistically not different from zero, and dark gray statistically significant and positive at conventional significance levels.

In Table 6.2, we report some examples of the possible Hausman tests one can devise to learn about the appropriateness of the sufficient sets used in estimation. In particular, in Table 6.2 we report the distribution of p-values of a Hausman test on the price coefficient and separately on the advertising coefficient for each individual. The validity of these Hausman tests relies on the maintained assumptions that unobserved preference heterogeneity is correctly specified by ISSL model (6.1) and that the "smallest" of the sufficient sets used as a reference is small enough to satisfy Condition 1. Overall, these results suggest that both the Complete and the PPH may be too large and systematically include products not considered or unavailable to individuals when making choices on-the-go. Among the proposed sufficient sets, the most robust—in the sense of Condition 1—is the PPH using

11 months of previous purchases. Consequently, we regard the comparison between the PPH 1 year versus the PPH 11 months as the most informative: for a substantial share of the sample (37% in the top panel and 63% in the bottom panel), the Hausman tests provide some evidence that, during any purchase occasion, individuals consider at least the chocolate bars they bought in the previous year. In general, if a researcher is not satisfied by the frequency with which the Hausman test is rejected, she can then specify smaller sufficient sets for those individuals with p-values below 10%, re-estimate the model, take these estimates as the reference points for another round of Hausman tests, and so on until the rate of non-rejection is deemed satisfactory. In the interest of space, we end the specification testing here and consider a PPH of 1 year sufficient for our illustrative example.

Table 6.2: *Hausman Tests*

	% of sample with p-value on individual Hausman test			
	>0.1	0.05-0.1	0.01-0.05	<0.01
Price coefficients				
Complete v PPH	12.0	1.9	3.4	82.7
PPH v PPH 1 year	16.2	2.8	5.6	75.4
PPH 1 year v PPH 11 months	37.0	8.5	13.0	41.5
Advertising coefficients				
Complete v PPH	0.0	0.0	0.0	100.0
PPH v PPH 1 year	0.0	8.5	0.0	91.5
PPH 1 year v PPH 11 months	63.0	10.2	0.0	26.9

Notes: The Table summarizes p-values of a Hausman test for each individual.

Our empirical results are in line with Goeree (2008)'s. With respect to price sensitivity, in a simplified model with three products, (Goeree, 2008, Appendix B, pages 2-7) shows analytically that the more likely are individuals to select among less than the full choice set (what she calls "limited information"), the more attenuated will price elasticities be (i.e., closer to zero). This is also what she finds in her empirical results (Goeree, 2008, Table VII), with price elasticities smaller in absolute value than their full-information counterparts (estimated on what we would call the Complete sufficient set).

Across specifications, we find that our estimates of advertising sensitivity are smaller when using the Past Purchase History sufficient set. As described above, the literature analyzing the economics of advertising has argued that advertising can both inform individuals about products' existence and so increase the likelihood that they are in individuals' choice sets, as well as directly influence

individual utility, shifting their preferences. The estimates using the Complete sufficient set can, at some intuitive level, be considered as a "reduced form" that captures both of these effects, while the estimates using the Past Purchase History sufficient sets, by focusing on those products for which individual attention is presumed to be already high, identifies the effects of advertising mainly through its influence on preferences. If this story is an accurate characterization of behavior in the on-the-go chocolate market, then we would expect to find, as we do, *smaller* estimated advertising sensitivity with the Past Purchase History than with the Complete sufficient set.

We can use the estimated preference parameters to compute the willingness-to-pay for advertising, something in which firms and advertising executives are likely to be interested, and the willingness-to-pay for individual products, something in which firms and retailers are likely to be interested. In our simple illustrative application, willingness-to-pay for (log) advertising is computed as:

$$\widehat{WTP}_{ia} = -\frac{\partial V_{ij_t t}/\partial \ln a_{ibt}}{\partial V_{ij_t t}/\partial p_{oj_t t}} = -\frac{\widehat{\beta_g}}{\widehat{\alpha_i}}, \tag{6.5}$$

whose mean and standard deviation we report in Table 6.3.

Table 6.3: *Willingness-to-Pay for Advertising*

		Complete	PPH	PPH 1 year	PPH 11 months
WTP	Mean	0.756	0.656	0.403	0.404
	Std Dev	0.168	0.236	0.249	0.261

Notes: The Table reports the mean and standard deviation of the estimated willingness-to-pay for advertising. Each column shows the estimates for the indicated sufficient set.

These WTP estimates suggest that there is significant bias in willingness-to-pay for advertising when using estimates from the Complete sufficient set. At the mean WTP for advertising, a one-standard deviation increase in the log advertising stock, $\ln a_{ibt}$, equal to 0.69 (or 69%), implies an increase in valuation of a product of 52.2 pence when using the Complete sufficient set.[37] As the average price of a chocolate product is 58 pence, this is a 90% increase. By contrast, the estimates obtained using the PPH 1 year sufficient set suggest a one-standard deviation increase in the log advertising stock increases the value of a product by 27.8 pence, or a 48% increase.[38]

[37] $(0.69 \times 0.756) = 0.522$, where 0.756 is the mean WTP for advertising from Table 6.3 using the Complete sufficient set.

[38] $(0.69 \times 0.403) = 0.278$, where 0.403 is the mean WTP for advertising from Table 6.3 using the Past Purchase History for the past year.

This illustration in the on-the-go chocolate market in the UK shows how failing to account for unobserved choice set heterogeneity can significantly bias preference estimates and the economic inferences that might arise from them in discrete-choice demand estimation. From a business strategy perspective, failing to account for unobserved choice set heterogeneity would lead a researcher to conclude that individuals are more sensitive to price, that advertising has a greater impact on preferences, and that most products are more desired than individual preferences truly indicate.

7 Conclusion

In this paper, we survey the two main empirical approaches to tackling the problem of unobserved choice set heterogeneity: "integrating over" and "differencing out" unobserved choice sets. The two approaches originate from different econometric literatures, started respectively by Manski (1977) and McFadden (1978). While integrating over heterogeneous unobserved choice sets is commonly done in empirical applications, differencing them out appears to be less popular in this context, possibly because the McFadden (1978)'s original motivation was to facilitate estimation with large but *observed* choice sets. We provide a unifying notation for understanding the two approaches and, inspired by Chamberlain (1980), we propose the use of consumers' observed choices paired with assumptions about the evolution of their unobserved choice sets over time as a practical tool to construct proper choice *subsets* in panel data environments. We call these subsets "sufficient sets."

Sufficient sets serve several purposes. First, sufficient sets help clarify that differencing out can also address the problem of unobserved choice sets, and that it is complementary to integrating over them. Second, sufficient sets prove useful to implement both approaches in practice. Third, they help translate economic assumptions derived from the characteristics of a given choice environment into econometric assumptions appropriate for estimation.

We illustrate some of the relevant issues and methods both in Monte Carlo simulations and in an empirical illustration of on-the-go demand for chocolate bars in the UK. Both exercises highlight how different assumptions on individuals' choice sets will have a material impact on estimation, but also that existing methods combined with the use of sufficient sets may substantially help in reducing the detrimental consequences of unobserved choice set heterogeneity.

References

Abaluck, Jason and Abi Adams, "What Do Consumers Consider Before They Choose: Identification from Asymmetric Demand Responses," *Working Paper*, 2017.

Abdulkadiroglu, Atila and Tayfun Sönmez, "School choice: A mechanism design approach," *The American Economic Review*, 2003, *93* (3), 729–747.

Arellano, Manuel, *Panel data econometrics*, Oxford University Press, 2003.

Bajari, Patrick, Jeremy T Fox, and Stephen P Ryan, "Linear regression estimation of discrete choice models with nonparametric distributions of random coefficients," *American Economic Review*, 2007, *97* (2), 459–463.

_ , _ , and _ , "Evaluating wireless carrier consolidation using semiparametric demand estimation," *Quantitative Marketing and Economics*, 2008, *6* (4), 299.

Başar, Gözen and Chandra Bhat, "A parameterized consideration set model for airport choice: an application to the San Francisco Bay area," *Transportation Research Part B: Methodological*, 2004, *38* (10), 889–904.

Ben-Akiva, M. and Brian Boccara, "Discrete choice models with latent choice sets," *International Journal of Research in Marketing*, 1995, pp. 9–24.

Berry, Steven, James Levinsohn, and Ariel Pakes, "Differentiated Products Demand Systems from a Combination of Micro and Macro Data: The New Car Market," *Journal of Political Economy*, 2004, pp. 68–105.

Bierlaire, Michel, Denis Bolduc, and Daniel McFadden, "The estimation of generalized extreme value models from choice-based samples," *Transportation Research Part B: Methodological*, 2008, *42* (4), 381–394.

Bordalo, P., N. Gennaioli, and A. Shleifer, "Salience and Consumer Choice," *Journal of Political Economy*, 2014, *121* (5), 803–843.

Bronnenberg, B. and V. Vanhonacker, "Limited choice sets, local price response, and implied measures of price competition," *Journal of Marketing Research*, 1996, pp. 163–173.

Bruno, Hernán A and Naufel J Vilcassim, "Research NoteStructural Demand Estimation with Varying Product Availability," *Marketing Science*, 2008, *27* (6), 1126–1131.

Caplin, Andrew and Mark Dean, "Search, choice, and revealed preference," *Theoretical Economics*, 2011, *6* (1), 19–48.

_ , _ , and Daniel Martin, "Search and satisficing," *The American Economic Review*, 2011, *101* (7), 2899–2922.

Chamberlain, Gary, "Analysis of Covariance with Qualitative Data," *The Review of Economic Studies*, 1980, *47*, 225–238.

Chiang, Jeongwen, Siddhartha Chib, and Chakravarthi Narasimhan, "Markov chain Monte Carlo and models of consideration set and parameter heterogeneity," *Journal of Econometrics*, 1998, *89* (1), 223–248.

Ching, Andrew T, Tülin Erdem, and Michael P Keane, "A simple method to estimate the roles of learning, inventories and category consideration in consumer choice," *Journal of choice modelling*, 2014, *13*, 60–72.

Ciliberto, Federico, Charles Murry, and Elie T Tamer, "Market structure and competition in airline markets," *Working paper*, 2016.

Conlon, Christopher T and Julie Holland Mortimer, "Demand estimation under incomplete product availability," *American Economic Journal-Microeconomics*, 2013, *5* (4), 1–30.

Currie, Janet, Stefano DellaVigna, Enrico Moretti, and Vikram Pathania, "The Effect of Fast Food Restaurants on Obesity and Weight Gain," *American Economic Journal: Economic Policy*, 2010, *2* (3), 32–63.

de Palma, Andr, Nathalie Picard, and Paul Waddell, "Discrete Choice Models with Capacity Constraints: An Empirical Analysis of the Housing Market of the Greater Paris Region," *Journal of Urban Economics*, 2007, pp. 204–230.

D'Haultfœuille, Xavier and Alessandro Iaria, "A Convenient Method for the Estimation of the Multinomial Logit Model with Fixed Effects," *Economics Letters*, 2016, *141*, 77–79.

Draganska, M., M. Mazzeo, and K. Seim, "Beyond Plain Vanilla: Modeling Joint Product Assortment and Pricing Decisions," *Quantitative Marketing and Economics*, 2009, *7*, 105–146.

Draganska, Michaela and Daniel Klapper, "Choice set heterogeneity and the role of advertising: An analysis with micro and macro data," *Journal of Marketing Research*, 2011, *48* (4), 653–669.

Dubé, JP, G. Hitsch, and P Manchanda, "An empirical model of advertising dynamics," *Quantitative Marketing and Economics*, 2005, *3*, 107–144.

Dubois, Pierre, Rachel Griffith, and Martin O'Connell, "The Effects of Banning Advertising in Junk Food Markets," *CEPR Discussion Papers*, 2016.

_ , _ , and _ , "How well targeted are soda taxes?," *CEPR Discussion Papers*, 2019.

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

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

Erdem, Tülin and Joffre Swait, "Brand credibility, brand consideration, and choice," *Journal of consumer research*, 2004, *31* (1), 191–198.

Fack, Gabrielle, Julien Grenet, and Yinghua He, "Beyond truth-telling: Preference estimation with centralized school choice," *Working Paper*, 2015.

Fox, Jeremy, Kyoo il Kim, Stephen Ryan, and Patrick Bajari, "A SImple Estimator for the Distribution of Random Coefficients," *Quantitative Economics*, 2011, *2* (3), 381–418.

Fox, Jeremy T, "Semiparametric estimation of multinomial discrete-choice models using a subset of choices," *The RAND Journal of Economics*, 2007, *38* (4), 1002–1019.

_ , Kyoo il Kim, and Chenyu Yang, "A simple nonparametric approach to estimating the distribution of random coefficients in structural models," *Journal of Econometrics*, 2016, *195* (2), 236–254.

Frejinger, Emma, Michel Bierlaire, and Moshe Ben-Akiva, "Sampling of alternatives for route choice modeling," *Transportation Research Part B: Methodological*, 2009, *43* (10), 984–994.

G., S. Becker and K. M. Murphy, "A Simple Theory of Advertising as a Good or Ban," *The Quarterly Journal of Economics*, 1993, *108* (4), 941–964.

Gabaix, Xavier, "A Sparsity-Based Model of Bounded Rationality," *Quarterly Journal of Economics*, 2014, *129* (4), 1661–1710.

Gaynor, Martin, Carol Propper, and Stephan Seiler, "Free to Choose? Reform, Choice and Consideration Sets in the English National Health Service," *American Economic Review*, 2016, *forthcoming.*

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

Goeree, J. K., C. A. Holt, and T. R. Palfrey, "Regular quantal response equilibrium," *Experimental Economics*, 2005, *8* (4), 347–367.

Goeree, Michelle Sovinsky, "Limited Information and Advertising in the U.S. Personal Computer Industry," *Econometrica*, 2008, *76* (5), 1017–1074.

Guevara, C Angelo and Moshe E Ben-Akiva, "Sampling of alternatives in logit mixture models," *Transportation Research Part B: Methodological*, 2013, *58*, 185–198.

_ and _ , "Sampling of alternatives in multivariate extreme value (MEV) models," *Transportation Research Part B: Methodological*, 2013, *48*, 31–52.

Hausman, Jerry A and Paul A Ruud, "Specifying and testing econometric models for rank-ordered data," *Journal of econometrics*, 1987, *34* (1), 83–104.

Hausman, Jerry and Daniel McFadden, "Specification tests for the multinomial logit model," *Econometrica*, 1984, pp. 1219–1240.

Heiss, Florian, Daniel McFadden, Joachim Winter, Amelie Wuppermann, and Bo Zhou, "Inattention and switching costs as sources of inertia in medicare part d," 2016. Working paper.

Hickman, William and Julie Holland Mortimer, "Demand Estimation with Availability Variation," *Handbook on the Economics of Retailing and Distribution*, 2016, pp. 306–342.

Ho, Kate, Joseph Hogan, and Fiona Scott Morton, "The impact of consumer inattention on insurer pricing in the Medicare Part D program," 2015. Working paper.

Honka, E., A. Hortaçsu, and M. Wildenbeest, "Empirical Search and Consideration Sets," *Working Paper*, 2018.

Honka, Elisabeth, "Quantifying search and switching costs in the US auto insurance industry," *The RAND Journal of Economics*, 2014, *45* (4), 847–884.

Huang, Yufeng and Bart J Bronnenberg, "Pennies for your thoughts: Costly product consideration and purchase quantity thresholds," 2017.

Iaria, Alessandro, "Consumer benefit of big-box supermarkets: The importance of controlling for endogenous entry," *Working paper*, 2014.

Imbens, Guido W and Charles F Manski, "Confidence intervals for partially identified parameters," *Econometrica*, 2004, pp. 1845–1857.

Janssen, Maarten CW and José Luis Moraga-González, "Strategic pricing, consumer search and the number of firms," *The Review of Economic Studies*, 2004, *71* (4), 1089–1118.

Keane, Michael P and Nada Wasi, "Estimation of discrete choice models with many alternatives using random subsets of the full choice set: With an application to demand for frozen pizza," 2012.

Kőszegi, Botond and Adam Szeidl, "A model of focusing in economic choice," *The Quarterly Journal of Economics*, 2013, *128* (1), 53–104.

Leicester, Andrew and Zoe Oldfield, "An analysis of consumer panel data," *IFS Working Papers W09/09*, 2009.

Li, Sophia Ying, Joe Mazur, Yongjoon Park, James W Roberts, Andrew Sweeting, and Jun Zhang, "Endogenous and Selective Service Choices After Airline Mergers," *National Bureau of Economic Research*, 2018.

los Santos, Babur De, Ali Hortaçsu, and Matthijs R Wildenbeest, "Testing models of consumer search using data on web browsing and purchasing behavior," *The American Economic Review*, 2012, *102* (6), 2955–2980.

Lu, Zhentong, "Estimating Multinomial Choice Models with Unobserved Choice Sets," 2018.

Manski, Charles F, "Maximum score estimation of the stochastic utility model of choice," *Journal of econometrics*, 1975, *3* (3), 205–228.

_ , "The structure of random utility models," *Theory and decision*, 1977, *8* (3), 229–254.

Manzini, Paola and Marco Mariotti, "Stochastic choice and consideration sets," *Econometrica*, 2014, *82* (3), 1153–1176.

Masatlioglu, Yusufcan and Daisuke Nakajima, "Choice by iterative search," *Theoretical Economics*, 2013, *8* (3), 701–728.

_ , _ , and Erkut Y Ozbay, "Revealed attention," *American Economic Review*, 2012, pp. 2183–2205.

Matejka, Filip and Alisdair McKay, "Rational inattention to discrete choices: A new foundation for the multinomial logit model," *American Economic Review*, 2015, *105* (1), 272–98.

McFadden, Daniel, "Modeling the Choice of Residential Location," in A. Karlqvist, L. Lundqvist, F. Snickars, and J. Weibull, eds., *Spatial Interaction Theory and Planning Models*, Vol. 1, North-Holland, 1978, pp. 75–96.

_ , "Regression-based specification tests for the multinomial logit model," *Journal of econometrics*, 1987, *34* (1-2), 63–82.

_ , William B Tye, and Kenneth Train, *An application of diagnostic tests for the independence from irrelevant alternatives property of the multinomial logit model*, Institute of Transportation Studies, University of California, 1977.

Morgan, Peter and Richard Manning, "Optimal search," *Econometrica*, 1985, pp. 923–944.

Musalem, Andrés, "When demand projections are too optimistic: A structural model of product line and pricing decisions," *Working paper*, 2015.

Nierop, Erjen Van, Bart Bronnenberg, Richard Paap, Michel Wedel, and Philip Hans Franses, "Retrieving unobserved consideration sets from household panel data," *Journal of Marketing Research*, 2010, *47* (1), 63–74.

Rhodes, Andrew, "Multiproduct retailing," *The Review of Economic Studies*, 2014, *82* (1), 360–390.

Roberts, John H and James M Lattin, "Development and testing of a model of consideration set composition," *Journal of Marketing Research*, 1991, pp. 429–440.

Ruud, Paul A, "Tests of specification in econometrics," *Econometric Reviews*, 1984, *3* (2), 211–242.

Simon, Herbert A, "A behavioral model of rational choice," *The quarterly journal of economics*, 1955, *69* (1), 99–118.

Train, K and C Winston, "Vehicle choice behavior and the decling market share of US automakers," *International economic review*, 2007.

Train, Kenneth E, Daniel L McFadden, and Moshe Ben-Akiva, "The demand for local telephone service: A fully discrete model of residential calling patterns and service choices," *The RAND Journal of Economics*, 1987, pp. 109–123.

Walters, Christopher R, "The demand for effective charter schools," Technical Report, National Bureau of Economic Research 2014.

Appendices

A Quantifying the Size of the Bias in MNL model

Table A.1 provides Monte Carlo evidence in which we quantify the size of the bias in MNL and mixed MNL models from mistakenly attributing to individuals alternatives that were not available to them. The three panels describe the relative importance of different features of the choice set generating process on the extent of the bias arising from unobserved choice set heterogeneity. We report the average bias and the standard deviation of the estimates (across 20 replications) that arise if the researcher imputes the full choice set instead of the true (heterogeneous and unobserved) choice set to all individuals in all choice situations.

In each scenario, the data generating process is a MNL model with systematic utility $V(X_{ij_tt}, \theta) = X_{ij_tt}\beta$ and heterogeneous choice sets across individuals. Given these data, we report in the first column estimates from a MNL model with systematic utility $V(X_{ij_tt}, \theta) = \delta_{j_t} + X_{ij_tt}\beta$ and with full choice sets, while we report in the second column estimates from a mixed MNL model with $V(X_{ij_tt}, \theta_i) = \delta_{j_t} + X_{ij_tt}\beta_i$, $\beta_i = \beta + \sigma \times \nu_i$, ν_i distributed standard normal, and with full choice sets.[39] Note that, differently from the data generating process, both estimated models include alternative-specific constants δ_{j_t}'s. In the top panel, we report results for the baseline model, where all individuals make choices from the full choice set. In this case, both the MNL and the mixed MNL models with full choice sets are correctly specified and virtually unbiased. In the second panel, an increasing share of individuals make choices from a choice set of four randomly selected alternatives. In both the standard and mixed MNL models, the bias increases with the share of individuals facing constrained choice sets. In the mixed MNL model both the estimated mean $\widehat{\beta}$ and the estimated standard deviation $\widehat{\sigma}$ of the random coefficient get increasingly biased. In the third panel, 30% of individuals make choices from a choice set of two, three, or four randomly selected alternatives. For a given share of individuals with constrained choice sets, in both estimated models the bias increases with the severity of the constraint in choice sets. In the bottom panel, 10% of individuals have their first-best alternative removed from the choice set, with an increasing "distance" between the systematic utilities of the (removed) first best

[39] The mixed MNL is estimated by simulated maximum likelihood using 150 shifted and shuffled Halton standard normal draws of ν_i per individual. By substantially increasing the number of draws per individual, results remain qualitatively unaffected.

and the (chosen) second best. Again, the bias increases the more individuals prefer the alternatives that are not included in their true choice sets (but that are mistakenly included in the choice sets of the models used in estimation). Interestingly, and differently from the previous two panels, the bias in the mixed MNL estimates appears to be concentrated in the estimated standard deviation $\widehat{\sigma}$ of the random coefficient, while the mean $\widehat{\beta}$ is always basically unbiased.

The results are intuitive and show the consequences of failing to account for unobserved choice set heterogeneity. The size of the bias can be substantial. Collectively, these results confirm the theoretical insights from section 2.2: the estimation bias is proportional to the extent of the (incorrect) choice set enlargement and to the size of the systematic utilities of the alternatives mistakenly added to the choice sets. Furthermore, the results show that simply adding alternative-specific constants and random coefficients to a MNL model does not address the econometric problems introduced by unobserved choice set heterogeneity.

Table A.1: *Size of the bias in the MNL model with full choice sets*

	MNL	Mixed MNL	
	Bias (StdDev) $\widehat{\beta}$	Bias (StdDev) $\widehat{\beta}$	Bias (StdDev) $\widehat{\sigma}$
Baseline			
100% of consumers have full choice set	0.005	0.013	0.071
	(0.032)	(0.034)	(0.080)
Increasing the share of individuals with constrained choice sets			
90% have full choice set, 10% choose from 4 out of 5	-0.223	-0.019	0.498
	(0.021)	(0.034)	(0.023)
70% have full choice set, 30% choose from 4 out of 5	-0.525	-0.272	0.531
	(0.013)	(0.023)	(0.019)
50% have full choice set, 50% choose from 4 out of 5	-0.719	-0.529	0.446
	(0.007)	(0.015)	(0.015)
Increasing the num. of alt. randomly removed from choice sets			
30% have 4 out of 5 available	-0.525	-0.272	0.531
	(0.013)	(0.023)	(0.019)
30% have 3 out of 5 available	-0.839	-0.425	0.683
	(0.007)	(0.021)	(0.019)
30% have 2 out of 5 available	-1.139	-0.474	0.916
	(0.003)	(0.022)	(0.020)
Increasing the differentiation of alt., 10% have first-best removed			
First-best alternative is slightly better	-0.346	-0.064	0.682
$(V_1 - V_2)/V_1 \simeq 10\%$, $\sigma_X^2 = 1.5$	(0.017)	(0.031)	(0.024)
First-best alternative is better	-0.471	-0.026	0.775
$(V_1 - V_2)/V_1 \simeq 35\%$, $\sigma_X^2 = 2.5$	(0.012)	(0.029)	(0.020)
First-best alternative is much better	-0.740	0.010	0.891
$(V_1 - V_2)/V_1 \simeq 110\%$, $\sigma_X^2 = 5.5$	(0.009)	(0.036)	(0.025)

We consider a population of 1,000 individuals making a sequence of choices over 10 choice situations. On each choice situation they choose between a maximum of five alternatives. The indirect utility of each alternative is specified as in equation (2.1). The true systematic utility is $V(X_{ij_tt}, \theta) = X_{ij_tt}\beta$, and the unobserved portion of utility, ϵ_{ij_tt}, is distributed i.i.d. Gumbel. In the baseline specification, X_{ij_tt} is drawn from a normal distribution with mean 0 and variance 5, and $\beta = 2$. In the MNL column, we report estimates of a MNL model with $V(X_{ij_tt}, \theta) = \delta_{j_t} + X_{ij_tt}\beta$ and where all individuals are incorrectly assumed to always have full choice sets. In the mixed MNL column, we report estimates of a mixed MNL model with $V(X_{ij_tt}, \theta_i) = \delta_{j_t} + X_{ij_tt}\beta_i$, $\beta_i = \beta + \sigma \times \nu_i$, where we draw 150 shifted and shuffled Halton ν_i's per individual from a standard normal, and all individuals are incorrectly assumed to always have full choice sets. The mixed MNL is exclusively used in estimation, the data are always generated from a standard MNL model. The table reports averages of the biases and standard deviations of the estimates across 20 replications per scenario. In the top panel, all individuals make choices from the full choice set. In the second panel, an increasing share of individuals make choices from a choice set of four randomly selected alternatives. In the third panel, 30% of individuals make choices from a choice set of two, three, or four randomly selected alternatives. In the bottom panel, 10% of individuals have their first-best alternative removed from the choice set, with an increasing "distance" between the systematic utilities of the (removed) first best and the (chosen) second best. In the second to bottom panels, choice sets differ across individuals but are constant across choice situations within individual.

B Importance Sampling Procedure from Goeree (2008)

In this Appendix, we detail how to implement the importance sampling procedure proposed by Goeree (2008) for the estimation of model (3.2) when choice sets are potentially large. To obtain further computational simplifications, this procedure can be combined with the use of sufficient sets, as described in subsection 3.3.

The unconditional probability of individual i choosing alternative j_t in choice situation t is:

$$\Pr\left[Y_{it}=j_t|\,\theta,\gamma\right]=\sum_{c_t\in C_t^j}\underbrace{\frac{\exp\left(V\left(X_{ij_tt},\theta\right)\right)}{\sum_{d_t\in c_t}\exp\left(V\left(X_{id_tt},\theta\right)\right)}}_{\Pr\left[Y_{it}=j_t|CS_{it}^{\star}=c_t,\theta\right]}\times\overbrace{\prod_{l_t\in c_t}\phi_{il_tt}\left(\gamma\right)\prod_{k_t\notin c_t}\left(1-\phi_{ik_tt}\left(\gamma\right)\right)}^{\Pr[CS_{it}^{\star}=c_t|\gamma]},\tag{B.1}$$

where C_t^j is the collection of all possible choice sets that include alternative j_t in period t, $\Pr\left[Y_{it}=j_t|\,CS_{it}^{\star}=c_t,\theta\right]$ is the choice probability conditional on choice set $CS_{it}^{\star}=c_t$, and $\Pr\left[CS_{it}^{\star}=c_t|\,\gamma\right]$ is the probability of i being matched to choice set c_t at choice situation t. Individual i's probability of alternative l_t to be in their choice set in t is:

$$\phi_{il_tt}(\gamma)=\frac{\exp(W_{il_tt}(\gamma))}{1+\exp(W_{il_tt}(\gamma))},\tag{B.2}$$

where γ are the choice set generating process parameters. The aim is to estimate both θ and γ by maximum likelihood on the basis of (B.1). Doing this directly is often numerically infeasible because the set of possible choice sets in each t is usually too large to handle. Goeree (2008) proposed a simulation method to ease the computation of (B.1). In the basic version of it, for each i and t one would approximate (B.1) by drawing R choice sets $\{c_{it}^r|\,r=1,\ldots,R\}$ according to probability $\Pr\left[CS_{it}^{\star}=c_{it}^r|\,\gamma\right]$ and then by averaging out across the resulting conditional choice probabilities:[40]

$$\widehat{\Pr}\left[Y_{it}=j_t|\,\theta,\gamma\right]=R^{-1}\sum_{r=1}^{R}\Pr\left[Y_{it}=j_t|\,CS_{it}^{\star}=c_{it}^r,\theta\right].\tag{B.3}$$

An estimator based on (B.3) would still be numerically demanding since the probability with which each c_{it}^r is drawn, $\Pr\left[CS_{it}^{\star}=c_{it}^r|\,\gamma\right]$, is a function of γ. This means that at each iteration of the

[40] As a useful complement, Honka (2014) proposed a kernel smoothing procedure to prevent lumpiness in these simulated choice probabilities even for "manageable" numbers of draws, R.

maximization routine, one would have to re-draw the R simulated choice sets $\{c_{it}^r | r = 1, \ldots, R\}$ for each i and t. To overcome also this problem, Goeree (2008) proposed an importance sampling version of simulator (B.3) that allows her to draw all the choice sets once and for all at the beginning of estimation. The idea of the importance sampling is that we wish to draw random variable c from probability $p(c)$ but we are not able to directly. However, we know how to draw from probability $g(c)$ and we have a closed-form solution for expression $\frac{p(c)}{g(c)}$. Consequently, one can draw several c^r's from $g(c^r)$, multiply each draw c^r by $\frac{p(c^r)}{g(c^r)}$, and the resulting distribution of the drawn c^r's will be the desired $p(c^r)$. In our context, the desired probability is $p(c_{it}^r) = \prod_{l_t \in c_{it}^r} \phi_{il_tt}(\gamma) \prod_{k_t \notin c_{it}^r} (1 - \phi_{ik_tt}(\gamma))$, while $g(c_{it}^r) = \prod_{l_t \in c_{it}^r} \phi_{il_tt}^0 \prod_{k_t \notin c_{it}^r} \left(1 - \phi_{ik_tt}^0\right)$ where $\phi_{il_tt}^0$ is (B.2) evaluated at some initial guess γ_0 that will be picked at the beginning of estimation and will not change until the end. Then, the importance sampling simulator of (B.1) is:

$$\begin{aligned}
\widehat{\widehat{\Pr}}[Y_{it} = j_t | \theta, \gamma] &= R^{-1} \sum_{r=1}^{R} \frac{p(c_{it}^r)}{g(c_{it}^r)} \times \Pr[Y_{it} = j_r | CS_{it}^\star = c_{it}^r, \theta] \\
&= R^{-1} \sum_{r=1}^{R} \left[\frac{\prod_{l_t \in c_{it}^r} \phi_{il_tt}(\gamma) \prod_{k_t \notin c_{it}^r} (1 - \phi_{ik_tt}(\gamma))}{\prod_{l_t \in c_{it}^r} \phi_{il_tt}^0 \prod_{k_t \notin c_{it}^r} \left(1 - \phi_{ik_tt}^0\right)} \times \Pr[Y_{it} = j_t | CS_{it}^\star = c_{it}^r, \theta] \right].
\end{aligned} \tag{B.4}$$

Simulator (B.4) can be implemented with the following algorithm. Before starting estimation, one should draw R choice sets $\{c_{it}^r | r = 1, \ldots, R\}$ and compute their "drawing" probabilities $g(c_{it}^r)$ for each i and t. This can be done as follows.

1. Set some initial value for γ and call it γ_0.
2. Given γ_0, compute $\phi_{il_tt}^0 = \phi_{il_tt}(\gamma_0) = \frac{\exp(W_{il_tt}(\gamma_0))}{1+\exp(W_{il_tt}(\gamma_0))}$ for each i, alternative $l_t \in J_t$, and t.[41]
3. For each i and t, draw R choice sets $\{c_{it}^r | r = 1, \ldots, R\}$ from ϕ_{ilt}^0, $l \in J_t$. Each choice set c_{it}^r can be drawn as follows:
   (a) Draw an independent uniform $u_{ilt}^r \in [0, 1]$ for each $l_t \in J_t$.
   (b) Alternative $l_t \in J_t$ belongs to c_{it}^r if and only if $u_{ilt}^r \leq \phi_{ilt}^0$.

[41] Remember that J_t is the set of alternatives available in the market in t.

4. For each i and t, compute the probability of having drawn each of the R choice sets $\{c_{it}^r | r = 1, \ldots, R\}$ as $g(c_{it}^r) = \prod_{l_t \in c_{it}^r} \phi_{il_t t}^0 \prod_{k_t \notin c_{it}^r} (1 - \phi_{ik_t t}^0)$.

Then, given $\{c_{it}^r | r = 1, \ldots, R\}$ and their "drawing" probabilities $g(c_{it}^r)$ for each i and t, one can proceed to the estimation of θ and γ by simulated maximum likelihood.

1. For each guessed value of (θ, γ), compute the predicted probability of the observed choice sequence of each i as:

$$\widehat{\widehat{\Pr}}[Y_i = (Y_{i1} = j_1, \ldots, Y_{it} = j_t, \ldots, Y_{iT} = j_T) | \theta, \gamma] = \prod_{t=1}^{T} \widehat{\widehat{\Pr}}[Y_{it} = j_t | \theta, \gamma], \tag{B.5}$$

   where each $\widehat{\widehat{\Pr}}[Y_{it} = j_t | \theta, \gamma]$ in the product is computed as in (B.4) given $\{c_{it}^r | r = 1, \ldots, R\}$ and their probabilities $g(c_{it}^r)$.

2. Take the log of the individual likelihood contribution from (B.5) and sum across all individuals.

3. Keep iterating with new guesses of (θ, γ) until the log-likelihood function computed at the previous step is maximized.

C Derivation of Sufficient Set Logit (SSL) Model (3.5)

$$
\begin{aligned}
&\Pr[Y_i = j | f(Y_i) = r, \theta] \\
&= \Pr[Y_i = j | f(Y_i) = r, \mathcal{CS}_i^\star = c, \theta] \\
&= \frac{\Pr[Y_i = j, Y_i \in r, \mathcal{CS}_i^\star = c | \theta, \gamma]}{\Pr[Y_i \in r, \mathcal{CS}_i^\star = c | \theta, \gamma]} \\
&= \frac{\Pr[Y_i = j, Y_i \in r | \mathcal{CS}_i^\star = c, \theta] \Pr[\mathcal{CS}_i^\star = c | \gamma]}{\Pr[Y_i \in r | \mathcal{CS}_i^\star = c, \theta] \Pr[\mathcal{CS}_i^\star = c | \gamma]} \\
&= \frac{\Pr[Y_i = j, Y_i \in r | \mathcal{CS}_i^\star = c, \theta]}{\sum_{k \in \mathcal{U}} \Pr[Y_i = k, Y_i \in r | \mathcal{CS}_i^\star = c, \theta]} \qquad \text{(C.1)} \\
&= \frac{\prod_{t=1}^{T} \frac{\exp(V(X_{ij_t t}, \theta))}{\sum_{v_t \in CS_{it}^\star = c_t} \exp(V(X_{iv_t t}, \theta))}}{\sum_{k \in f(Y_i) = r} \prod_{t=1}^{T} \frac{\exp(V(X_{ik_t t}, \theta))}{\sum_{v_t \in CS_{it}^\star = c_t} \exp(V(X_{iv_t t}, \theta))}} \\
&= \frac{\prod_{t=1}^{T} \exp(V(X_{ij_t t}, \theta))}{\sum_{k \in f(Y_i) = r} \prod_{t=1}^{T} \exp(V(X_{ik_t t}, \theta))}
\end{aligned}
$$

Assumption 1 and Condition 1 imply the IIA property, and the first equality follows from its definition. Note that conditioning the choice probability on $f(Y_i) = r$ is equivalent to conditioning the choice Y_i to be from the set r, or $Y_i \in r$. The second and third equalities follow from the definition of conditional probability, while the fourth follows from the law of total probability. In the fourth equality, $\mathcal{U}$ is the universal set of *all* choice sequences. The fifth equality follows from $\Pr[Y_i = k, Y_i \in r | \mathcal{CS}_i^\star = c, \theta]$ being equal to $\Pr[Y_i = k | \mathcal{CS}_i^\star = c, \theta]$ for any $k \in r$ or, alternatively, being equal to 0 for any $k \notin r$. In the last equality, $\sum_{v_t \in CS_{it}^\star = c_t} \exp(V(X_{iv_t t}, \theta))$ cancels out. Finally, consistency of the conditional Maximum Likelihood Estimator derived from $\Pr[Y_i = j | f(Y_i) = r, \theta]$ follows from McFadden (1978).

D Derivation of Sufficient Set Logit (SSL) Model (3.6)

In this Appendix we demonstrate that equation (3.6) holds if and only if $f(Y_i) = \times_{t=1}^{T} f_t(Y_i)$.

IF part. Suppose that $f(Y_i) = \times_{t=1}^{T} f_t(Y_i)$. Then we can re-write the denominator of conditional logit model (3.5), $\Pr[Y_i = j | f(Y_i) = r, \theta]$, as (omitting the f in the conditioning for simplicity):

$$\begin{aligned}
\sum_{(k_1,\ldots,k_T)\in r} \prod_{t=1}^{T} \exp(V_{ik_tt}) &= \sum_{(k_1,\ldots,k_T)\in r_1\times\cdots\times r_T} \prod_{t=1}^{T} \exp(V_{ik_tt}) \\
&= \left(\sum_{k_1\in r_1} \exp(V_{ik_11})\right) \sum_{(k_2,\ldots,k_T)\in r_2\times\cdots\times r_T} \prod_{t=2}^{T} \exp(V_{ik_tt}) \\
&= \left(\sum_{k_1\in r_1} \exp(V_{ik_11})\right)\left(\sum_{k_2\in r_2} \exp(V_{ik_22})\right) \sum_{(k_3,\ldots,k_T)\in r_3\times\cdots\times r_T} \prod_{t=3}^{T} \exp(V_{ik_tt}) \\
&\vdots \\
&= \prod_{t=1}^{T}\left(\sum_{k_t\in r_t} \exp(V_{ik_tt})\right),
\end{aligned} \tag{D.1}$$

which implies that:

$$
\begin{aligned}
\Pr\left[Y_i = j \middle| f\left(Y_i\right) = r, \theta\right] &= \frac{\prod_{t=1}^{T} \exp\left(V_{ij_t t}\right)}{\sum_{(k_1,\ldots,k_T)\in r} \prod_{t=1}^{T} \exp\left(V_{ik_t t}\right)} \\
&= \frac{\prod_{t=1}^{T} \exp\left(V_{ij_t t}\right)}{\prod_{t=1}^{T}\left(\sum_{k_t \in r_t} \exp\left(V_{ik_t t}\right)\right)} \\
&= \prod_{t=1}^{T} \frac{\exp\left(V_{ij_t t}\right)}{\sum_{k_t \in r_t} \exp\left(V_{ik_t t}\right)} \\
&= \prod_{t=1}^{T} \Pr\left[Y_{it} = j_t \middle| f_t\left(Y_i\right) = r_t, \theta\right].
\end{aligned}
\tag{D.2}
$$

To complete the proof, we are now going to show that $\Pr[Y_{it} = j_t | f(Y_i) = r, \theta] = \Pr[Y_{it} = j_t | f_t(Y_i) = r_t, \theta]$. Define the set $M\left(\widetilde{j}_s\right) = \left\{ (z_1, \ldots, z_s, \ldots, z_T) \middle| z \in f\left(Y_i\right) = r, z_s = \widetilde{j}_s \right\}$ as the collection of choice sequences that have alternative $\widetilde{j}_s$ in position s. It then follows that:

$$
\begin{aligned}
&\Pr\left[Y_{is} = \widetilde{j}_s \middle| f\left(Y_i\right) = r, \theta\right] \\
&= \sum_{j \in M\left(\widetilde{j}_s\right)} \Pr\left[Y_i = j \middle| f\left(Y_i\right) = r, \theta\right] \\
&= \sum_{j \in M\left(\widetilde{j}_s\right)} \frac{\prod_{t=1}^{T} \exp\left(V_{ij_t t}\right)}{\sum_{(k_1,\ldots,k_T)\in r} \prod_{t=1}^{T} \exp\left(V_{ik_t t}\right)} \\
&= \left(\sum_{(k_1,\ldots,k_T)\in r} \prod_{t=1}^{T} \exp\left(V_{ik_t t}\right)\right)^{-1} \left(\sum_{j \in M\left(\widetilde{j}_s\right)} \prod_{t=1}^{T} \exp\left(V_{ij_t t}\right)\right).
\end{aligned}
\tag{D.3}
$$

Similarly to (D.1), $f\left(Y_i\right) = \times_{t=1}^{T} f_t\left(Y_i\right)$ implies that $M\left(\widetilde{j}_s\right) = r_1 \times \cdots \times \left\{\widetilde{j}_s\right\} \times \cdots \times r_T$, and consequently that the numerator of (D.3) can be re-written as:

$$\begin{aligned}\sum_{j\in M\left(\widetilde{j}_s\right)}\prod_{t=1}^{T}\exp\left(V_{ij_tt}\right) &= \sum_{(j_1,\ldots,j_s,\ldots,j_T)\in r_1\times\cdots\times\left\{\widetilde{j}_s\right\}\times\cdots\times r_T}\prod_{t=1}^{T}\exp\left(V_{ij_tt}\right)\\ &= \exp\left(V_{i\widetilde{j}_ss}\right)\prod_{t\neq s}\left(\sum_{j_t\in r_t}\exp\left(V_{ij_tt}\right)\right).\end{aligned} \tag{D.4}$$

Plugging (D.1) and (D.4) into (D.3), we obtain:

$$\begin{aligned}&\Pr\left[Y_{is}=\widetilde{j}_s\middle|\, f\left(Y_i\right)=r,\theta\right]\\ &=\sum_{j\in M\left(\widetilde{j}_s\right)}\Pr\left[Y_i=j\middle|\, f\left(Y_i\right)=r,\theta\right]\\ &=\left(\prod_{t=1}^{T}\left(\sum_{k_t\in r_t}\exp\left(V_{ik_tt}\right)\right)\right)^{-1}\left(\exp\left(V_{i\widetilde{j}_ss}\right)\prod_{t\neq s}\left(\sum_{j_t\in r_t}\exp\left(V_{ij_tt}\right)\right)\right)\\ &=\left(\left(\sum_{k_s\in r_s}\exp\left(V_{ik_ss}\right)\right)\prod_{t\neq s}\left(\sum_{k_t\in r_t}\exp\left(V_{ik_tt}\right)\right)\right)^{-1}\left(\exp\left(V_{i\widetilde{j}_ss}\right)\prod_{t\neq s}\left(\sum_{j_t\in r_t}\exp\left(V_{ij_tt}\right)\right)\right)\\ &=\frac{\exp\left(V_{i\widetilde{j}_ss}\right)}{\sum_{k_s\in rfanocaria_s}\exp\left(V_{ik_ss}\right)}\\ &=\Pr\left[Y_{is}=\widetilde{j}_s\middle|\, f_s\left(Y_i\right)=r_s,\theta\right].\end{aligned} \tag{D.5}$$

ONLY IF part. Consider two choice situations t and s. For these, define $f_t=\{j_t|\,(j_1,\ldots,j_t,\ldots,j_T)\in f\left(Y_i\right)=r\}$ and $f_s=\{j_s|\,(j_1,\ldots,j_s,\ldots,j_T)\in f\left(Y_i\right)=r\}$ as the collections of alternatives that appear in at least one sequence belonging to $f\left(Y_i\right)=r$ at positions t and s, respectively. Suppose $f\left(Y_i\right)\neq\times_{t=1}^{T}f_t\left(Y_i\right)$, then $\exists\ t$ and s such that $\left(\widetilde{j}_t\in f_t,\widetilde{j}_s\in f_s\right)$ and $\left(\widetilde{j}_1,\ldots,\widetilde{j}_t,\ldots,\widetilde{j}_s,\ldots,\widetilde{j}_T\right)\notin f\left(Y_i\right)=r$. It then follows that:

$$\Pr\left[Y_{it}=\widetilde{j}_t\middle|\, Y_{is}=\widetilde{j}_s,f\left(Y_i\right)=r,\theta\right]=0,$$

while, since $\widetilde{j}_t \in f_t$:

$$\Pr\left[Y_{it} = \widetilde{j}_t \middle| f(Y_i) = r, \theta\right] > 0.$$

This implies that Y_{it} and Y_{is} are not conditionally independent.

E Sufficient Sets and Bounds on Choice Probabilities, Elasticities, and Consumer Surplus

The various approaches presented in section 3 differ in the extent to which they allow us to evaluate functions of θ (or θ_i for the ISSL model), for example: willingness to pay, elasticities, consumer surplus, or the analysis of counterfactuals, such as evaluating the effects of a change in tax policy or a merger between manufacturers. At one extreme is Fox (2007)'s PMSE discussed in subsection 3.2.3. The PMSE can point-estimate the preference parameters θ, as well as simple functions of them (e.g., willingness-to-pay), but cannot reveal functions of θ that involve knowledge of the distribution of the unobserved portion of utility.[42] At the other extreme are models based on Manski (1977) that integrate over unobserved choice sets discussed in subsections 3.1 and 3.3. These approaches involve specifying a model of choice set formation that reveals the distribution of choice sets in the population of individuals, so they also allow point-identification of all functions of θ that depend on this distribution.

The SSL, the ISSL, and the SSML represent an intermediate case. In this Appendix, we describe parameters and functions of parameters that we can point-identify, and how we can use sufficient sets to derive bounds on several useful functions of these parameters. For ease of notation, we limit the discussion to the SSL with the understanding that similar ideas readily apply also to the ISSL and to the SSML. To simplify exposition, suppose that the systematic utilities take the form:

$$V(X_{ij_tt}, \theta) = \delta_{j_t} + X_{ij_tt}\beta + \alpha p_{j_tt},$$

[42] See Fox (2007) and Bajari et al. (2008) for more details on this point.

where $\theta = [\delta_1, ..., \delta_J, \beta, \alpha]$ and $p_{j_t t}$ is the price of alternative j_t. We can point-identify the vector of preference parameters θ from the SSL model $\Pr[Y_i = j | f(Y_i) = r_i, \theta]$.[43] We can similarly point-identify simple functions of θ. For example, we are often interested in willingness-to-pay (WTP) for product characteristic k, $X^k_{ij_t t}$. By Roy's Identity, this can be computed as:

$$WTP_k = -\frac{\partial V_{ij_t t}/\partial X^k_{ij_t t}}{\partial V_{ij_t t}/\partial p_{j_t t}} = -\frac{\beta_k}{\alpha}. \tag{E.1}$$

Other outputs of economic interest, however, require information about the distribution of choice sets in the population for point-identification. We cannot point-identify these functions, but we can place bounds on them. This makes clear that these functions are only point-identified when one relies on strong assumptions about the choice set formation process. The probability with which i chooses alternative j_t given choice set $CS^\star_{it} = c_{it}$ is

$$Pr^{CS^\star}_{ij_t t}(\theta) \equiv \Pr\left[Y_{it} = j_t | CS^\star_{it} = c_{it}, \theta\right] = \frac{\exp\left(\delta_{j_t} + X_{ij_t t}\beta + \alpha p_{j_t t}\right)}{\sum_{m \in c_{it}} \exp\left(\delta_m + X_{imt}\beta + \alpha p_{mt}\right)} \tag{E.2}$$

if $j_t \in CS^\star_{it} = c_{it}$ and zero otherwise. This choice probability depends on i's (unobserved) choice set, $CS^\star_{it}$. Suppose that we observe a superset Q_{it} of the true but unobserved choice set, so that $CS^\star_{it} \subseteq Q_{it}$. This could be, for example, the collection of all alternatives observed to be chosen by any i in choice situation t. It follows that, even if we do not directly observe $CS^\star_{it} = c_{it}$, $f_t(Y_i) \subseteq CS^\star_{it}$ and $CS^\star_{it} \subseteq Q_{it}$. We can therefore use these conditions to bound the true but unobserved denominator of the SSL choice probabilities for any $X_{it} = [X_{i1t}, p_{1t}, ..., X_{iJt}, p_{Jt}]$ and θ:

$$\sum_{m \in f_t(Y_i) = r_{it}} \exp\left(\delta_m + X_{imt}\beta + \alpha p_{mt}\right) \leq \sum_{m \in CS^\star_{it} = c_{it}} \exp\left(\delta_m + X_{imt}\beta + \alpha p_{mt}\right) \leq \sum_{m \in Q_{it} = q_{it}} \exp\left(\delta_m + X_{imt}\beta + \alpha p_{mt}\right). \tag{E.3}$$

Similar to $Pr^{CS^\star}_{ij_t t}(\theta)$, denote for brevity also $Pr^{Q}_{ij_t t}(\theta) \equiv \Pr[Y_{it} = j_t | Q_{it} = q_{it}, \theta]$ and $Pr^{f}_{ij_t t}(\theta) \equiv \Pr[Y_{it} = j_t | f_t(Y_i) = r_{it}, \theta]$. It then follows from (E.3) that for any $j_t \in f_t(Y_i) = r_{it}$:

$$Pr^{Q}_{ij_t t}(\theta) \leq Pr^{CS^\star}_{ij_t t}(\theta) \leq Pr^{f}_{ij_t t}(\theta). \tag{E.4}$$

[43] Note that here, differently from most other parts in the paper, we will keep track of the "i" subscript in the realizations of the sufficient sets, $f(Y_i) = r_i$, and of the choice sets, $CS^\star_{it} = c_{it}$. This is essential to avoid confusion when computing averages across individuals, as detailed below.

That is to say, the true choice probability with which i chooses j_t in t is bounded from below by the same probability assuming i chooses from some superset of the unobserved choice set, $Q_{it} = q_{it}$, and from above by the same probability assuming i chooses from just their sufficient set, $f_t(Y_i) = r_{it}$. Observe that $Pr^f_{ij_tt}(\theta)$ takes the usual logit form whenever $j_t \in r_{it}$, but that it equals zero whenever $j_t \notin r_{it}$. Hence, for those $j_t \in q_{it}$ but $j_t \notin r_{it}$, $Pr^f_{ij_tt}(\theta)$ will not be a valid upper bound for $Pr^{CS^\star}_{ij_tt}(\theta)$: even if $j_t \notin r_{it}$, it can still be the case that $j_t \in CS^\star_{it} = c_{it}$ and so that $Pr^{CS^\star}_{ij_tt}(\theta) > 0$. Similarly, among the $j_t \in q_{it}$ that $j_t \notin r_{it}$, there can be some $j_t \notin c_{it}$. But for those $j_t \in q_{it}$ that $j_t \notin c_{it}$, $Pr^Q_{ij_tt}(\theta) > Pr^{CS^\star}_{ij_tt}(\theta) = 0$: $Pr^Q_{ij_tt}(\theta)$ will not be a valid lower bound for $Pr^{CS^\star}_{ij_tt}(\theta)$. It is then unclear how to bound $Pr^{CS^\star}_{ij_tt}(\theta)$ for those $j_t \in q_{it}$ but $j_t \notin r_{it}$. However, it is always possible to construct bounds for the probability with which i would choose j_t if indeed j_t were to be *added* to their true but unobserved choice set, $CS^\star_{it} \cup \{j_t\} = c_{it} \cup \{j_t\}$:

$$Pr^{CS^\star \cup j}_{ij_tt}(\theta) = \Pr\left[Y_{it} = j_t | CS^\star_{it} \cup \{j_t\} = c_{it} \cup \{j_t\}, \theta\right] = \frac{\exp\left(\delta_{j_t} + X_{ij_tt}\beta\right)}{\sum_{m \in c_{it} \cup \{j_t\}} \exp\left(\delta_m + X_{imt}\beta\right)}. \tag{E.5}$$

By defining $Pr^{Q \cup j}_{ij_tt}(\theta)$ and $Pr^{f \cup j}_{ij_tt}(\theta)$ analogously, note that $Pr^{Q \cup j}_{ij_tt}(\theta) = Pr^{Q}_{ij_tt}(\theta)$, $Pr^{CS^\star \cup j}_{ij_tt}(\theta) = Pr^{CS^\star}_{ij_tt}(\theta)$, and $Pr^{f \cup j}_{ij_tt}(\theta) = Pr^{f}_{ij_tt}(\theta)$ for any $j_t \in r_{it}$, while $Pr^{Q \cup j}_{ij_tt}(\theta) \leq Pr^{CS^\star \cup j}_{ij_tt}(\theta)$ and $Pr^{CS^\star \cup j}_{ij_tt}(\theta) \leq Pr^{f \cup j}_{ij_tt}(\theta)$ for any $j_t \notin r_{it}$. Using these facts, we can then complement condition (E.4) for those $j_t \notin r_{it}$ and propose choice probability bounds for all (i, j_t, t) combinations:

$$Pr^{Q \cup j}_{ij_tt}(\theta) \leq Pr^{CS^\star \cup j}_{ij_tt}(\theta) \leq Pr^{f \cup j}_{ij_tt}(\theta). \tag{E.6}$$

Condition (E.6) can be used to construct bounds for functions of individual choice probabilities, such as average choice probabilities or elasticities. The average choice probability of alternative j_t for a certain group of individuals $i = 1, \ldots, I_t$ can be bounded by:

$$I_t^{-1} \sum_{i=1}^{I_t} Pr^{Q \cup j}_{ij_tt}(\theta) \leq I_t^{-1} \sum_{i=1}^{I_t} Pr^{CS^\star \cup j}_{ij_tt}(\theta) \leq I_t^{-1} \sum_{i=1}^{I_t} Pr^{f \cup j}_{ij_tt}(\theta). \tag{E.7}$$

With indirect utilities that are linear in price, individual i's own- and cross-price elasticities are:

$$\begin{aligned}
\xi_{it}^{jj}(X_{it},\theta) &= \beta_p p_{j_t t}(1 - Pr_{ij_t t}^{CS^\star \cup j}(\theta)) \\
&= \beta_p p_{j_t t}\left(1 - \frac{\exp(\delta_{j_t} + X_{ij_t t}\beta)}{\sum_{m \in c_{it} \cup \{j_t\}} \exp(\delta_m + X_{imt}\beta)}\right), \\
\xi_{it}^{jk}(X_{it},\theta) &= \beta_p p_{k_t t} Pr_{ik_t t}^{CS^\star \cup j}(\theta) \\
&= -\beta_p p_{k_t t}\left(\frac{\exp(\delta_{k_t} + X_{ik_t t}\beta)}{\sum_{m \in c_{it} \cup \{j_t\}} \exp(\delta_m + X_{imt}\beta)}\right),
\end{aligned} \tag{E.8}$$

where $p_{j_t t}$ is j_t's price in choice situation t and β_p is the price coefficient. As (E.8) makes clear, even though we may have a consistent estimator of $\delta = [\delta_1, \ldots, \delta_j, \ldots, \delta_J]$ and β, we still do not know the exact $CS_{it}^\star \cup \{j_t\} = c_{it} \cup \{j_t\}$ for each i and t, and thus the true $Pr_{ij_t t}^{CS^\star \cup j}(\theta)$, $\forall j_t \in CS_{it}^\star \cup \{j_t\} = c_{it} \cup \{j_t\}$. Given (E.8), (E.6), and $\beta_p < 0$, we obtain the following bounds on the elasticities for any j_t, k_t, X_{it}, δ, and β:

$$\begin{aligned}
\underbrace{\beta_p p_{j_t t}(1 - Pr_{ij_t t}^{f \cup j}(\theta))}_{\text{Lower (in abs. value) Bound}} &\leq \xi_{it}^{jj}(X_{it},\theta) \leq \underbrace{\beta_p p_{j_t t}(1 - Pr_{ij_t t}^{Q \cup j}(\theta))}_{\text{Upper (in abs. value) Bound}} \\
\underbrace{-\beta_p p_{k_t t} Pr_{ik_t t}^{Q \cup j}(\theta)}_{\text{Lower Bound}} &\leq \xi_{it}^{jk}(X_{it},\theta) \leq \underbrace{-\beta_p p_{k_t t} Pr_{ik_t t}^{f \cup j}(\theta)}_{\text{Upper Bound}}.
\end{aligned} \tag{E.9}$$

The same bounds in equation (E.3) imply the ability to bound consumer surplus. Let the true consumer surplus of individual i in t be:

$$W_{it}\left(X_{it} | \theta, CS_{it}^\star = c_{it}\right) = \zeta + \frac{1}{\alpha} \ln\left(\sum_{m \in c_{it}} \exp\left(\delta_m + X_{imt}\beta + \alpha p_{mt}\right)\right), \tag{E.10}$$

where ζ is Euler's constant. Then, for any X_{it} and θ:

$$W_{it}\left(X_{it}|\,\theta, f_t\left(Y_i\right)=r_{it}\right) \leq W_{it}\left(X_{it}|\,\theta, CS^{\star}_{it}=c_{it}\right) \leq W_{it}\left(X_{it}|\,\theta, Q_t=q_{it}\right). \tag{E.11}$$

E.1 Confidence Intervals for Elasticity Bounds

As an example of how to conduct inference on the identification regions described above, we construct confidence intervals for the elasticity bounds following Imbens and Manski (2004). For notational simplicity, we limit our discussion to a single elasticity term $\xi^{jk}_{it}(X_{it},\theta)$, although the same ideas can be extended to the collection of all elasticities. Refer to the upper and lower bounds of $\xi^{jk}_{it}(X_{it},\theta)$ in (E.9) as to $\overline{\xi^{jk}_{it}}(X_{it},\theta)$ and $\underline{\xi^{jk}_{it}}(X_{it},\theta)$, respectively. Denote the elasticity *bounds* of $\xi^{jk}_{it}(X_{it},\theta)$ by the 2×1 vector $B\left(\xi^{jk}_{it}(X_{it},\theta)\right)=\left[\underline{\xi^{jk}_{it}}(X_{it},\theta),\overline{\xi^{jk}_{it}}(X_{it},\theta)\right]'$ and the corresponding elasticity *interval* from (E.9) by $IN\left(\xi^{jk}_{it}(X_{it},\theta)\right)$. Then, given X_{it} and our consistent $\widehat{\theta}$, we can estimate the elasticity bounds $B\left(\xi^{jk}_{it}(X_{it},\theta)\right)$ by $B\left(\xi^{jk}_{it}(X_{it},\widehat{\theta})\right)$. We derive the corresponding $100\left(1-\alpha\right)$ percent confidence interval $CI_{1-\alpha}$ from condition:

$$\inf_{\xi^{jk}_{it}\in IN\left(\xi^{jk}_{it}(X_{it},\theta)\right)}\left\{\lim_{I\to\infty}\Pr\left[\xi^{jk}_{it}\in CI_{1-\alpha}\right]\right\}\geq 1-\alpha. \tag{E.12}$$

Since our estimator is consistent and asymptotically normal, i.e., $\widehat{\theta}\sqrt{I}\xrightarrow{d}\mathcal{N}\left(\theta,V_{\theta}\right)$, by the delta-method:

$$B\left(\xi^{jk}_{it}(X_{it},\widehat{\theta})\right)\sqrt{I}\xrightarrow{d}\mathcal{N}\left(B\left(\xi^{jk}_{it}(X_{it},\theta)\right),\frac{\partial B\left(\xi^{jk}_{it}(X_{it},\theta)\right)}{\partial\theta'}V_{\theta}\frac{\partial B\left(\xi^{jk}_{it}(X_{it},\theta)\right)'}{\partial\theta'}\right). \tag{E.13}$$

Refer to the 2×2 asymptotic variance–covariance matrix of $B\left(\xi^{jk}_{it}(X_{it},\widehat{\theta})\right)$ as to $\Sigma_{B\left(\xi^{jk}_{it}\right)}$. It follows that, whenever $f_t\left(Y_i\right)\cup\{j_t\}=r_{it}\cup\{j_t\}$ is a strict subset of $Q_{it}\cup\{j_t\}=q_{it}\cup\{j_t\}$, so that for any X_{it} and θ, $\underline{\xi^{jk}_{it}}(X_{it},\theta)<\overline{\xi^{jk}_{it}}(X_{it},\theta)$, condition (E.12) is satisfied by:

$$CI_{1-\alpha}=\left[\underline{\xi^{jk}_{it}}(X_{it},\widehat{\theta})-q_{1-\alpha}\sqrt{\Sigma^{11}_{B\left(\xi^{jk}_{it}\right)}},\overline{\xi^{jk}_{it}}(X_{it},\widehat{\theta})+q_{1-\alpha}\sqrt{\Sigma^{22}_{B\left(\xi^{jk}_{it}\right)}}\right], \tag{E.14}$$

where $q_{1-\alpha}$ is the $\left(1-\alpha\right)^{th}$ quantile of the standard normal distribution.

In the extreme case in which $f_t(Y_i) \cup \{j_t\} = r_{it} \cup \{j_t\} = Q_{it} \cup \{j_t\} = q_{it} \cup \{j_t\}$, $\underline{\xi_{it}^{jk}}(X_{it}, \theta) = \overline{\xi_{it}^{jk}}(X_{it}, \theta)$ for any X_{it} and θ, and (E.14) is invalid. This is due to a discontinuity at $\underline{\xi_{it}^{jk}}(X_{it}, \theta) = \overline{\xi_{it}^{jk}}(X_{it}, \theta)$, since in that case the coverage of the interval is only $100\,(1-2\alpha)\,\%$ rather than the nominal $100\,(1-\alpha)\,\%$. (See Imbens and Manski (2004) for a modification of (E.14) that overcomes this problem.) However, note that *(a)* both $f_t(Y_i) \cup \{j_t\} = r_{it} \cup \{j_t\}$ and $Q_{it} \cup \{j_t\} = q_{it} \cup \{j_t\}$ are always perfectly observed by the econometrician, so that the appropriate $CI_{1-\alpha}$ can always be implemented and that *(b)* in our empirical application $f_t(Y_i) \cup \{j_t\} \subset Q_{it} \cup \{j_t\}$ for every i and t.

F Monte Carlo Evidence on Performance of SSL Model

In this Appendix, we report the results of Monte Carlo simulations evaluating the practical performance of MNL and SSL models in the presence of various forms of unobserved choice set heterogeneity. In table F.1, we directly vary the extent of choice set heterogeneity by randomly removing alternatives from choice sets, independently of the indirect utilities or the product characteristics of the removed alternatives. Differently, in table F.2 we implement two more economically relevant choice set formation processes: a model of screening on product characteristics (such as price) and a model of costly search. Here, our aim is to illustrate that even when choice set heterogeneity is the outcome of selection processes involving the alternatives' systematic utilities and/or product characteristics, the proposed SSL models work well without requiring the econometrician to know much about such possibly complex processes.

The first column of table F.1 reports results showing the bias in a MNL model from incorrectly assuming that all individuals in all choice situations have access to the full choice set, made of five alternatives. The second column reports estimates of the true MNL model, i.e. the model that correctly assigns the true choice set facing each individual in each choice situation. There is of course no estimation bias in this case. The remaining three columns report estimates from, respectively, the Full Purchase History (FPH), the Past Purchase History (PPH), and the Choice Permutation (CP) SSL models.

The top panel of table F.1 shows the lack of bias in the absence of unobserved choice set heterogeneity. The following two panels show, in turn, the bias arising from, first, increasing the share of individuals with restricted choice sets and, second, increasing the severity of the restriction on choice

sets. Overall, the results show that there is significant bias when we incorrectly assume full choice sets (the first column), but that there is no average bias when relying on any of these three SSL models for estimation. Along these lines, Appendix A presents further evidence of the magnitudes of biases that can arise when researchers mistakenly assign to individuals choice sets larger than the true ones.

Table F.1: *Performance of Sufficient Set Logits*

	MNL, full %|Bias|	MNL, true %|Bias|	FPH SSL %|Bias|	PPH SSL %|Bias|	CP SSL %|Bias|
Baseline					
100% full choice set	0.3%	0.3%	0.6%	0.8%	1.2%
Increasing share of individuals with a random product removed from choice set					
10% constrained	11.2%	0.2%	0.4%	0.9%	0.8%
30% constrained	26.3%	0.3%	0.6%	1.2%	0.6%
50% constrained	36.0%	0.4%	0.7%	1.0%	0.9%
Increasing share of products randomly removed from choice set					
30% have 4 of 5	26.3%	0.3%	0.6%	1.2%	0.6%
30% have 3 of 5	36.0%	0.7%	1.0%	1.3%	1.2%
30% have 2 of 5	57.0%	0.4%	0.5%	0.7%	0.5%

We consider a population of 1,000 individuals making a sequence of choices over 10 choice situations. On each choice situation, they choose between a maximum of five alternatives. The indirect utility of each alternative is specified as in equation (2.1). The systematic utility is $V(X_{ij_tt}, \theta) = \delta_{j_t} + X_{ij_tt}\beta$, *and the unobserved portion of utility,* ϵ_{ij_tt}, *is distributed i.i.d. Gumbel.* X_{ij_tt} *is drawn from a normal distribution with mean 0 and variance 5,* $\delta_{j_t} = 0$ *for all* j_t*'s, and* $\beta = 2$. *The table reports averages of the percentage absolute bias of the estimates,* $|(\widehat{\beta} - \beta)/\beta| \times 100$. *In the top panel, all individuals make choices from the full choice set. In the central panel, an increasing share of individuals make choices from a choice set of four randomly selected alternatives. In the bottom panel, 30% of individuals make choices from a choice set of two, three, or four randomly selected alternatives. In the central and bottom panels, choice sets differ across individuals but are constant across choice situations within individual. We simulate and average results over 20 replications per scenario. To speed up computations, the CP SSL is estimated by sampling at random (uniformly), for each individual, 5000 permutations of the observed sequence of choices, as suggested by D'Haultfœuille and Iaria (2016).*

Table F.2 reports results for two economically relevant choice set formation processes: a model of screening on product characteristics in the central panel and a model of costly search in the bottom panel. Both the models of screening and of costly search are simple. In these simulations, our aim is not to implement the most realistic screening and search models that have appeared in the literature, but rather to study the performance of MNL and SSL models when choice set heterogeneity is the outcome of non-trivial selection processes involving the alternatives' systematic utilities and/or product characteristics.

The first column of table F.2 reports results for a MNL with a full choice set of five alternatives. The second column reports results for the MNL with true choice sets, as if one could perfectly observe the outcomes of the screening and costly search for each individual in each choice situation. Both

models of screening and of costly search generate choice sets that are weakly growing over choice situations, compatibly with the assumptions of the PPH sufficient set. The third column reports the estimates of a PPH SSL model.

The central panel of table F.2 reports results for a choice set formation model of screening on product characteristic X_{ij_tt}. Each individual i has a maximum threshold $\overline{X}_i$ for the value of X_{ij_tt} they are willing to consider.[44] In $t = 1$, $CS^\star_{i1}$ contains those alternatives for which $X_{ij_11} \leq \overline{X}_i$.[45] Denote by $\overline{CS}^\star_{it}$ the collection of alternatives not in $CS^\star_{it}$. Once an alternative is considered in t, it will also be in $CS^\star_{it'}$ for $t' > t$. Accordingly, in any $t > 1$, individual i checks whether any of the alternatives in $\overline{CS}^\star_{i,t-1}$, i.e. those *not* already in $CS^\star_{i,t-1}$, has an acceptable value of X_{ij_tt} and includes in $CS^\star_{it}$ all those for which $X_{ij_tt} \leq \overline{X}_i$. In other words, in each $t > 1$, $CS^\star_{it}$ is the union between $CS^\star_{i,t-1}$ and those alternatives from $\overline{CS}^\star_{i,t-1}$ that pass the $\overline{X}_i$ screening.

The bottom panel of table F.2 reports results for a choice set formation model of costly search over alternatives. Each individual i in every t, given the set of alternatives already in their choice set from $t-1$, $CS^\star_{i,t-1}$, considers whether to incur a search cost of c_{ij} to include any *new* alternative in $CS^\star_{it}$ (i.e., any alternative belonging to $\overline{CS}^\star_{i,t-1}$). When considering whether to add or not an additional alternative to the choice set, individuals perfectly observe all the X_{ij_tt}'s and search costs, but need to form expectations about the ϵ_{ij_tt} error terms (according to Assumption 1, the choice set formation process cannot depend on the realizations of the error terms).[46] In $t = 1$, $CS^\star_{i1}$ contains those alternatives for which $V(X_{ij_11}, \theta) - c_{ij} \geq 0$.[47] In any $t > 1$, individual i is assumed to be able to add to their choice set at most one alternative from $\overline{CS}^\star_{i,t-1}$ (i.e., either add one alternative or nothing). Similar to the model of screening, once an alternative is considered in t, it will also be in $CS^\star_{it'}$ for $t' > t$. In any $t > 1$, individual i decides to search for an additional alternative to be included in CS_{it} only when the expected *net* benefit from searching is greater than the expected maximal utility from $CS^\star_{i,t-1}$ (i.e., what can be achieved without any additional search):

[44] Each X_{ij_tt} is distributed normal with mean 0 and variance 5. The individual-specific threshold $\overline{X}_i$ is distributed standard normal.

[45] To prevent $CS^\star_{i1}$ from being empty, a randomly selected alternative is included in $CS^\star_{i1}$ when $X_{ij_11} > \overline{X}_i$ for all alternatives.

[46] Each X_{ij_tt} is distributed normal with mean 0 and variance 5. The individual-alternative specific search cost c_{ij} is distributed log-normal with mean 3 and variance 1. Each ϵ_{ij_tt} error term is distributed Gumbel. Individuals have correct beliefs about the distribution of the error terms when computing expected utilities.

[47] When computing expected utilities, we ignore the Euler constant. This is an approximation only in $t = 1$, for any $t > 1$ the constant does indeed drop out of rule (F.1). To prevent $CS^\star_{i1}$ from being empty, a randomly selected alternative is included in $CS^\star_{i1}$ when $V(X_{ij_11}, \theta) - c_{ij} < 0$ for all alternatives.

$$\max_{j_t \in \overline{CS}^{\star}_{i,t-1}} \{V(X_{ij_t t}, \theta) - c_{ij}\} \geq \ln \left[\sum_{k_t \in CS^{\star}_{i,t-1}} \exp\left(V(X_{ik_t t}, \theta)\right) \right], \tag{F.1}$$

where $\overline{CS}^{\star}_{i,t-1}$ is the collection of alternatives not included in $CS^{\star}_{i,t-1}$, the choice set including all the alternatives searched for in the previous choice situations. When i decides to search in t, then the alternative in $\overline{CS}^{\star}_{i,t-1}$ corresponding to the largest expected net benefit from searching is included in $CS^{\star}_{it}$.

Overall, table F.2 shows that when the choice set formation process is a function of the alternatives' systematic utilities and/or product characteristics, mistakenly ignoring it may have detrimental effects on the estimation of preference parameters (first column). Clearly, if one had data on the true choice sets faced by each individual in each choice situation, then neither choice set generating process would cause any estimation problem given that Assumption 1 still holds (second column). Finally, the third column shows that the PPH SSL performs virtually as well as the true MNL (second column), with the advantage of not requiring the econometrician to have any additional data on true choice sets or to know much about the potentially complex details of the choice set generating process.

Table F.2: *A Model of Screening and a Model of Search*

	MNL, full %|Bias|	MNL, true %|Bias|	PPH SSL %|Bias|
Baseline			
100% full choice set	0.3%	0.3%	0.8%
Increasing share of individuals who screen sequentially			
30% screens	39.5%	0.2%	0.4%
50% screens	50.3%	0.1%	0.1%
90% screens	63.5%	0.1%	0.4%
Increasing share of individuals who search sequentially			
30% searches	52.8%	0.5%	1.6%
50% searches	64.3%	0.2%	1.6%
90% searches	77.3%	0.3%	2.6%

We consider a population of 1,000 individuals making a sequence of choices over 10 choice situations. On each choice situation, they choose between a maximum of five alternatives. The indirect utility of each alternative is specified as in equation (2.1). The systematic utility is $V(X_{ij_t t}, \theta) = \delta_{j_t} + X_{ij_t t}\beta$, *and the unobserved portion of utility,* $\epsilon_{ij_t t}$, *is distributed i.i.d. Gumbel.* $X_{ij_t t}$ *is drawn from a normal distribution with mean 0 and variance 5,* $\delta_{j_t} = 0$ *for all* j_t*'s, and* $\beta = 2$. *The table reports averages of the percentage absolute bias of the estimates,* $|(\widehat{\beta} - \beta)/\beta| \times 100$. *In the top panel, all individuals make choices from the full choice set. In the central panel, an increasing share of individuals make choices from a choice set formed sequentially by screening over the* $X_{ij_t t}$*'s (see text for detail). In the bottom panel, an increasing share of individuals make choices from a choice set formed sequentially by searching over alternatives (see text for detail). In the central and bottom panels, choice sets differ across individuals and evolve across choice situations within individual. We simulate and average results over 20 replications per scenario.*

G Specification Tests: Choosing Among Sufficient Sets

In this Appendix we first describe how Ruud (1984)'s *Factorization Theorem* can be used to construct specification tests (in the spirit of Hausman and McFadden (1984)) for SSL and ISSL models that are helpful to discriminate among different sufficient sets. Second, we illustrate with some concrete examples how to use these statistics to test for features of the choice set formation process and of unobserved preference heterogeneity. To keep notation simple, in what follows we focus on the SSL model with the understanding that the same results apply almost verbatim to the ISSL model.

Suppose that Assumption 1 holds, and that sufficient sets f_L and f_Z satisfy Condition 1, that $f_Z(Y_i) \subset f_L(Y_i)$, $Y_i \in \mathcal{CS}_i^\star = c$, and that $i = 1, \ldots, I$. Define $l_L(\theta)$ and $l_Z(\theta)$ as the log-likelihood functions corresponding to the SSL models with sufficient sets $f_L(Y_i)$ and $f_Z(Y_i)$, and denote by $\widehat{\theta}_L$ and $\widehat{\theta}_Z$ the corresponding MLEs. Then the following results hold:

1. The log-likelihood function $l_L(\theta)$ can be written as $l_L(\theta) = l_Z(\theta) + l_\triangle(\theta)$.
2. Provided that θ is identified in $l_\triangle(\theta)$, so that $\widehat{\theta}_\triangle$ is a well defined MLE, then:
   (a) $\widehat{\theta}_Z$ and $\widehat{\theta}_\triangle$ are asymptotically independent, and
   (b) $\widehat{\theta}_L$ is more efficient than $\widehat{\theta}_Z$.
3. Given result (2), then:
   (a) All Hausman tests based on pairwise estimator comparisons among $\widehat{\theta}_L$, $\widehat{\theta}_Z$, and $\widehat{\theta}_\triangle$ are equivalent,
   (b) The Likelihood Ratio statistic $LR = 2[l_Z(\widehat{\theta}_Z) + l_\triangle(\widehat{\theta}_\triangle) - l_L(\widehat{\theta}_L)]$ is asymptotically equivalent to the Hausman statistic comparing $\widehat{\theta}_L$ and $\widehat{\theta}_Z$, and
   (c) $\text{Var}\left(\widehat{\theta}_L - \widehat{\theta}_Z\right) = \text{Var}\left(\widehat{\theta}_Z\right) - \text{Var}\left(\widehat{\theta}_L\right)$.

Proof of Result (1). Starting from (3.5), we can re-write for every i the probability of the observed choice sequence j given $f_Z(Y_i) = z \subset f_L(Y_i) = l$ as:

$$\begin{aligned}
\Pr\left[Y_i = j | f_L(Y_i) = l, \theta\right] &= \frac{\prod_{t=1}^{T} \exp(V(X_{ij_t t}, \theta))}{\sum_{k \in f_L(Y_i)=l} \prod_{t=1}^{T} \exp(V(X_{ik_t t}, \theta))} \\
&= \Pr\left[Y_i = j | f_Z(Y_i) = z, \theta\right] \left(\frac{\Pr\left[Y_i = j | f_L(Y_i) = l, \theta\right]}{\Pr\left[Y_i = j | f_Z(Y_i) = z, \theta\right]} \right) \\
&= \frac{\prod_{t=1}^{T} \exp(V(X_{ij_t t}, \theta))}{\sum_{q \in f_Z(Y_i)=z} \prod_{t=1}^{T} \exp(V(X_{iq_t t}, \theta))} \frac{\sum_{q \in f_Z(Y_i)=z} \prod_{t=1}^{T} \exp(V(X_{iq_t t}, \theta))}{\sum_{k \in f_L(Y_i)=l} \prod_{t=1}^{T} \exp(V(X_{ik_t t}, \theta))} \\
&= \Pr\left[Y_i = j | f_Z(Y_i) = z, \theta\right] \Pr\left[Y_i \in f_Z(Y_i) = z | f_L(Y_i) = l, \theta\right],
\end{aligned}$$

where $\Pr\left[Y_i \in f_Z(Y_i) = z | f_L(Y_i) = l, \theta\right]$ is the probability that a choice sequence belongs to the "smaller" set z relative to the "larger" set l. By multiplying $\Pr\left[Y_i = j | f_L(Y_i) = l, \theta\right]$ across all individuals and by taking the logarithm, result (1) follows with $l_{\triangle}(\theta) = \sum_{i=1}^{I} \ln\left(\Pr\left[Y_i \in f_Z(Y_i) = z | f_L(Y_i) = l, \theta\right]\right)$.

Proof of Result (2). Given result (1) above, results (2a) and (2b) follow from the *Factorization Theorem* of (Ruud, 1984, result (1), p.24).

Proof of Result (3). Given result (1) above, result (3a) follows from the *Factorization Theorem* of (Ruud, 1984, result (3), p.24), while result (3b) follows from (Ruud, 1984, pp.28-9). Result (3c) can be proved as follows. (Ruud, 1984, result 2, p.24) shows that $\widehat{\theta}_L$ is asymptotically equivalent to $\mathrm{Var}\left(\widehat{\theta}_L\right) \mathrm{Var}\left(\widehat{\theta}_Z\right)^{-1} \widehat{\theta}_Z + \mathrm{Var}\left(\widehat{\theta}_L\right) \mathrm{Var}\left(\widehat{\theta}_{\triangle}\right)^{-1} \widehat{\theta}_{\triangle}$. This implies that $\mathrm{Cov}\left(\widehat{\theta}_L, \widehat{\theta}_Z\right) = \mathrm{Cov}\left(\mathrm{Var}\left(\widehat{\theta}_L\right) \mathrm{Var}\left(\widehat{\theta}_Z\right)^{-1} \widehat{\theta}_Z, \widehat{\theta}_Z\right) = \mathrm{Var}\left(\widehat{\theta}_L\right) \mathrm{Var}\left(\widehat{\theta}_Z\right)^{-1} \mathrm{Var}\left(\widehat{\theta}_Z\right) = \mathrm{Var}\left(\widehat{\theta}_L\right)$, where the first equality follows from result (2a). Consequently, $\mathrm{Var}\left(\widehat{\theta}_L - \widehat{\theta}_Z\right) = \mathrm{Var}\left(\widehat{\theta}_L\right) + \mathrm{Var}\left(\widehat{\theta}_Z\right) - 2\mathrm{Cov}\left(\widehat{\theta}_L, \widehat{\theta}_Z\right) = \mathrm{Var}\left(\widehat{\theta}_Z\right) - \mathrm{Var}\left(\widehat{\theta}_L\right)$.

The Likelihood Ratio statistic LR from result (3b) allows one to compare different SSL models derived from alternative assumptions on sufficient sets. It consists of the difference between an unrestricted log-likelihood function, $l_Z\left(\widehat{\theta}_Z\right) + l_{\triangle}\left(\widehat{\theta}_{\triangle}\right)$, and a restricted one, $l_L\left(\widehat{\theta}_L\right)$.[48] Even though LR requires

[48] As developed more fully in Ruud (1984), this form is common to many econometric tests, including incremental over-identifying (or Sargan) tests commonly used to investigate the validity of subsets of instruments (Arellano, 2003, Section 5.4.4).

the computation of a third estimator, $\widehat{\theta}_{\triangle}$, it is simpler to implement than other Hausman statistics based on quadratic forms. For instance, the statistic LR is always non-negative, bypassing the practical inconvenience of some estimated covariance matrices that fail to be positive definite. In contrast to some other Hausman statistics, LR also makes very transparent the computation of the degrees of freedom of the corresponding χ^2 distribution: they equal the number of parameters in $\widehat{\theta}_L$. Result (3c) is of practical convenience, it implies that the computation of Var $\left(\widehat{\theta}_L - \widehat{\theta}_Z\right)$, necessary for classical Hausman statistics, can proceed as in the standard case in which one of the compared estimators is fully efficient under the null hypothesis, even though no such efficiency assumption is required here.

G.1 Practical Examples of Testing Procedures

For simplicity of exposition, we limit our examples to the SSL model with the understanding that similar ideas readily apply also to the ISSL model, for which the IIA is only assumed *within* each individual (but not *across* individuals). In the context of SSL models, the examples of sufficient sets introduced in section (4) rely on the following economic assumptions:

- f_{CP}: Choice set stability across T choice situations *and* possibility of IIA violations in the form of individual-alternative specific fixed effects, δ_{ij_t}.
- f_{FPH}: Choice set stability across T choice situations *and* IIA property.
- f_{PPH}: Choice set evolution in the form of weakly growing choice sets (or, symmetrically, weakly shrinking choice sets) across T choice situations *and* IIA property.[49]

There are two possibilities for making comparisons across SSL models based on different sufficient sets f's, and each presents ways of implicitly testing for some of the maintained economic assumptions embedded in the compared sufficient sets. The first possibility is to compare f_{CP}, f_{FPH}, and f_{PPH} for choice sequences of constant length T. The second possibility is to fix a specific f, say f_{CP}, and to compare choice sequences with some of their *sub*-sequences: for example, the sequence $1, 2, \ldots, T^L$ can be split into two mutually exclusive sub-sequences $1, 2, \ldots, T^Z$ and $T^Z + 1, \ldots, T^L$, and this gives rise to different f_{CP}'s, f_{CP}^Z and f_{CP}^L such that $f_{CP}^Z(Y_i) \subset f_{CP}^L(Y_i)$ for any $Y_i \in \mathcal{CS}_i^\star = c$. We now illustrate with some examples each testing possibility in turn.

[49] Importantly, the IIA requirement follows from the SSL model and it is not intrinsic in the f_{FPH} and f_{PPH} sufficient sets. Neither sufficient set relies on the IIA property *across* individuals when employed in more general models such as the ISSL and the SSML discussed in subsection 3.2.2.

G.1.1 Comparisons of Different f's with Constant T

For choice sequences of a given length T, $f_{CP}(Y_i) \subseteq f_{FPH}(Y_i)$ and $f_{PPH}(Y_i) \subseteq f_{FPH}(Y_i)$ for any $Y_i \in \mathcal{CS}_i^\star = c$. Suppose $Y_i = (1,3)$. Then $f_{CP}(1,3) = \mathcal{P}(1,3) = \{(1,3),(3,1)\}$, $f_{FPH}(1,3) = \{1,3\} \times \{1,3\} = \{(1,1),(3,3),(1,3),(3,1)\}$, and $f_{PPH}(1,3) = \{1\} \times \{1,3\} = \{(1,1),(1,3)\}$. Note that there is no clear "inclusion" relationship between $f_{CP}(Y_i)$ and $f_{PPH}(Y_i)$. Given the *Factorization Theorem*, the above relationships among sufficient sets lead to two possible classes of tests. The first is about choice set stability and the second about deviations from the IIA property.

Choice Set Stability (given IIA property). In the context of SSL models, both f_{FPH} and f_{PPH} rely on the IIA property. However, they rely on different assumptions regarding the evolution of choice sets across choice situations: f_{FPH} assumes that unobserved choice sets do not change along the whole choice sequence, while f_{PPH} allows for the entry of new alternatives in the unobserved choice set while comparing choice situation t to $t+1$. On the one hand, if unobserved choice sets were stable, then both f's would give rise to consistent estimators $\widehat{\theta}_{FPH}$ and $\widehat{\theta}_{PPH}$, but result (2b) above tells us that $\widehat{\theta}_{FPH}$ would be more efficient than $\widehat{\theta}_{PPH}$. On the other hand, if unobserved choice sets were growing over choice situations, then only $\widehat{\theta}_{PPH}$ would be consistent: f_{FPH} would not satisfy Condition 1, inducing violations of the IIA property as discussed in subsection 2.2. It follows that, under the maintained assumption of the IIA property, a test for H_0: *(choice set stability in* $1,2,\ldots,T$*)* is $LR = 2\left[l_{PPH}\left(\widehat{\theta}_{PPH}\right) + l_{\triangle}\left(\widehat{\theta}_{\triangle}\right) - l_{FPH}\left(\widehat{\theta}_{FPH}\right)\right]$.

Departures from IIA property (given Choice Set Stability). The sufficient sets f_{FPH} and f_{CP} are both based on the same assumption of unobserved choice set stability in $1,2,\ldots,T$. However, in the context of SSL models, they rely on different assumptions regarding unobserved preference heterogeneity: f_{FPH} relies on the IIA property, while f_{CP} allows for individual-alternative specific fixed effects. On the one hand, if the IIA property held, then both f's would give rise to consistent estimators $\widehat{\theta}_{FPH}$ and $\widehat{\theta}_{CP}$, but result (2b) above tells us that $\widehat{\theta}_{FPH}$ would be more efficient than $\widehat{\theta}_{CP}$. On the other hand, if the IIA property were violated in ways encompassed by individual-alternative specific fixed effects, then only $\widehat{\theta}_{CP}$ would be consistent. It follows that, under the maintained assumption of choice set stability in $1,2,\ldots,T$, a test for H_0: *(IIA property)* is $LR = 2\left[l_{CP}\left(\widehat{\theta}_{CP}\right) + l_{\triangle}\left(\widehat{\theta}_{\triangle}\right) - l_{FPH}\left(\widehat{\theta}_{FPH}\right)\right]$.

G.1.2 Comparisons of Same f with Different Choice Sub-sequences

It is always possible to split choice sequences of length $1, 2, \ldots, T^L$ into two (or more) *mutually exclusive* sub-sequences $1, 2, \ldots, T^Z$ and $T^Z + 1, \ldots, T^L$. Then $f_{CP}^Z(Y_i) \subset f_{CP}^L(Y_i)$ for any $Y_i \in \mathcal{CS}_i^\star = c$. The same holds also for f_{FPH} and f_{PPH}. This method of making comparisons allows one to test for choice set stability in several alternative ways, but it does not enable one to test for departures from the IIA property (the two SSL models compared are always either both satisfying or both violating the IIA property).

Choice Set Stability: f_{CP} Example. In what follows we will show with an example that $f_{CP}^Z(Y_i) \subset f_{CP}^L(Y_i)$ for any $Y_i \in \mathcal{CS}_i^\star = c$ and afterward we will discuss how to use this fact to construct tests of choice set stability.

Suppose $J = 5$, $T^L = 4$, and that individual i is observed to make the choice sequence $Y_i = (j_1, j_2, j_3, j_4) = (3, 5, 5, 4)$.[50] By considering the observed choice sequence "at once," $Y_i = (3, 5, 5, 4)$ can be re-ordered in 12 different choice sequences.[51] Collect these sequences into the set $f_{CP}^L(Y_i) = l$. Assume that $V_i(X_{ij_tt}, \theta) = \delta_{ij_t} + X_{ij_tt}\beta$. Then, i's likelihood contribution given $f_{CP}^L(Y_i) = l$ is:

$$\begin{aligned} &\Pr\left[Y_i = (3,5,5,4) \middle| f_{CP}^L(Y_i) = l, \beta\right] \\ &= \frac{\exp\left(\left(X_{i31} + X_{i52} + X_{i53} + X_{i44}\right)\beta\right)}{\sum_{(j_1,j_2,j_3,j_4) \in f_{CP}^L(Y_i) = l} \exp\left(\left(X_{ij_11} + X_{ij_22} + X_{ij_33} + X_{ij_44}\right)\beta\right)}. \end{aligned} \tag{G.1}$$

Differently, by splitting i's observed choice sequence into two mutually exclusive pairs of choices $Y_{i1} = (3, 5)$ and $Y_{i3} = (5, 4)$, we get $f_{CP}^Z(Y_i) = f_{CP}^Z(Y_{i1}) \times f_{CP}^Z(Y_{i3})$ where $f_{CP}^Z(Y_{i1}) = \{(3,5), (5,3)\}$ and $f_{CP}^Z(Y_{i3}) = \{(5,4), (4,5)\}$. Then, i's likelihood contribution given $f_{CP}^Z(Y_{i1}) = z_1$ and $f_{CP}^Z(Y_{i3}) = z_3$ is:

[50] Alternative three in the first choice situation, alternative five in the second choice situation, etc.

[51] These sequences are: $(3,5,5,4)$, $(5,3,5,4)$, $(5,5,3,4)$, $(5,5,4,3)$, $(4,3,5,5)$, $(3,4,5,5)$, $(3,5,4,5)$, $(5,3,4,5)$, $(5,4,3,5)$, $(5,4,5,3)$, $(4,5,3,5)$, and $(4,5,5,3)$.

$$\Pr\left[Y_i = (3,5,5,4) | \, f_{CP}^{Z}(Y_i) = z_1 \times z_3, \beta\right]$$

$$= \frac{\exp\left(\left(X_{i31} + X_{i52}\right)\beta\right)}{\exp\left(\left(X_{i31} + X_{i52}\right)\beta\right) + \exp\left(\left(X_{i51} + X_{i32}\right)\beta\right)} \tag{G.2}$$

$$\times \frac{\exp\left(\left(X_{i53} + X_{i44}\right)\beta\right)}{\exp\left(\left(X_{i53} + X_{i44}\right)\beta\right) + \exp\left(\left(X_{i43} + X_{i54}\right)\beta\right)}.$$

By multiplying the binomial logits in (G.2), we get:

$$\Pr_i\left[Y_i = (3,5,5,4) | \, f_{CP}^{Z}\left(Y_i\right) = z, \beta\right] =$$

$$\frac{\exp\left(\left(X_{i31} + X_{i52} + X_{i53} + X_{i44}\right)\beta\right)}{\sum\limits_{(j_1,j_2,j_3,j_4) \in f_{CP}^{Z}(Y_i) = z} \exp\left(\left(X_{ij_1 1} + X_{ij_2 2} + X_{ij_3 3} + X_{ij_4 4}\right)\beta\right)}, \tag{G.3}$$

where $f_{CP}^{Z}\left(Y_i\right) = z$ collects sequences: $(3,5,5,4)$, $(3,5,4,5)$, $(5,3,5,4)$, and $(5,3,4,5)$. Consequently $f_{CP}^{Z}\left(Y_i\right) = z \subseteq f_{CP}^{L}\left(Y_i\right) = l$. In this example, f_{CP}^{Z} only uses information about 4 of the 12 possible choice sequences in f_{CP}^{L}. This implies that if unobserved choice sets were stable, then estimator $\widehat{\beta}_{CP}^{L}$ would be more efficient than $\widehat{\beta}_{CP}^{Z}$.

Moreover, the CP SSL estimated on choice sub-sequences may "discard" some choice situations: in the current example of sub-sequences of length two, whenever $j_t = j_{t+1}$ in $Y_{it} = (j_t, j_{t+1})$, then "fragment" Y_{it} of Y_i will not be used in estimation. For example, if i were observed to choose the sequence $Y_i = (3,4,5,5)$, then only $Y_{i1} = (3,4)$ would contribute to the likelihood function $l_{CP}^{Z}\left(\beta\right)$, while $l_{CP}^{L}\left(\beta\right)$ would still use the whole sequence $Y_i = (3,4,5,5)$. More precisely, if $Y_i = (3,4,5,5)$ were observed, then $f_{CP}^{L}\left(3,4,5,5\right) = f_{CP}^{L}\left(3,5,5,4\right) = l$ would still contain the same 12 choice sequences, while model (G.2) would collapse to:

$$\Pr\left[Y_i = (3,4,5,5) | \, f_{CP}^{Z}\left(Y_i\right) = h_1 \times h_3, \beta\right]$$

$$= \frac{\exp\left(\left(X_{i31} + X_{i42}\right)\beta\right)}{\exp\left(\left(X_{i31} + X_{i42}\right)\beta\right) + \exp\left(\left(X_{i41} + X_{i32}\right)\beta\right)}$$

$$\times \frac{\exp\left(\left(X_{i53} + X_{i54}\right)\beta\right)}{\exp\left(\left(X_{i53} + X_{i54}\right)\beta\right)} \tag{G.4}$$

$$= \frac{\exp\left(\left(X_{i31} + X_{i42} + X_{i53} + X_{i54}\right)\beta\right)}{\exp\left(\left(X_{i31} + X_{i42} + X_{i53} + X_{i54}\right)\beta\right) + \exp\left(\left(X_{i41} + X_{i32} + X_{i53} + X_{i54}\right)\beta\right)}$$

$$= \Pr\left[Y_i = (3,4,5,5) | \, f_{CP}^{Z}\left(Y_i\right) = h, \beta\right],$$

which is also equivalent to $\Pr\left[Y_{i1} = (3,4) | \, f_{CP}^{Z}\left(Y_{i1}\right) = h_1, \beta\right]$. In this case, then, $f_{CP}^{Z}\left(Y_{i1}\right) = h_1 \subset f_{CP}^{Z}\left(Y_i\right) = z \subset f_{CP}^{L}\left(Y_i\right) = l$. By result (2b) above, we can rank the corresponding estimators in terms of their relative efficiency. As a consequence, by splitting up choice sequences into mutually exclusive sub-sequences, one can face also this further loss of efficiency.

Model (G.1) requires stronger assumptions than model (G.3) for its consistent estimation. Consistent estimation of model (G.1) requires that alternatives $\{3,4,5\} \subseteq CS_{it}^{\star} = c_t$, $t = 1,2,3,4$. However, consistent estimation of model (G.3) only requires that $\{3,5\} \subseteq CS_{it}^{\star} = c_t$, $t = 1,2$ and that $\{4,5\} \subseteq CS_{it}^{\star} = c_t$, $t = 3,4$. In this example, if $4 \notin CS_{it}^{\star} = c_t$, $t = 1$ or 2, or $3 \notin CS_{it}^{\star} = c_t$, $t = 3$ or 4, then estimation of model (G.1) would not be consistent, while estimation of model (G.3) would.

These differences in consistency and relative efficiency suggest a Hausman test for unobserved choice set stability. If $\{3,4,5\} \subseteq CS_{it}^{\star} = c_t$, $t = 1,2,3,4$, then estimation of both model (G.1) and model (G.3) would be consistent. However, estimation of model (G.1) would be more efficient than estimation of model (G.3). If $4 \notin CS_{it}^{\star} = c_t$, $t = 1$ or 2 or $3 \notin CS_{it}^{\star} = c_t$, $t = 3$ or 4, then only estimation of model (G.3) would be consistent. It follows that, under the maintained assumption of unobserved preference heterogeneity in a form encompassed by individual-alternative specific fixed effects, a test for H_0: *(choice set stability in* $1,2,3,$ *and* $4)$ is $LR = 2\left[l_{CP}^{Z}\left(\widehat{\beta}_{PPH}^{Z}\right) + l_{\triangle}\left(\widehat{\beta}_{\triangle}\right) - l_{CP}^{L}\left(\widehat{\beta}_{CP}^{L}\right)\right]$.

H Specification Tests: Nested Logit and IIA

In what follows, we illustrate that with similar assumptions to those required by the MNL, sufficient sets can also be used for the consistent estimation of the *within-nest* part of a nested logit model when choice sets are unobserved, and that this is enough to implement a test for departures of the IIA along the lines of Hausman and McFadden (1984).

Suppose that the full collection of J alternatives is partitioned into N mutually exclusive nests $nest_n$ and that any individual i's choice set $CS_{it}^{\star}$ can be partitioned in N subsets of the N original nests, so that: $CS_{it}^{\star} = nest_{i1} \cup ... \cup nest_{in} \cup ... \cup nest_{iN}$, where for any n, $nest_{in}$ is either $nest_{in} \subseteq nest_n$ or empty. The econometrician knows $nest_n$, $n = 1, ..., N$, but does *not* know $nest_{in}$, $n = 1, ..., N$, for any i. Note that, for simplicity, we are assuming that both the original nests and individual i's nest subsets are constant over t. At the expense of some additional notation, this can be relaxed as in the case of the MNL. Denote by $Y_i = (Y_{i1}, ..., Y_{iT})$ individual i's sequence of chosen *alternatives* and by $Q_i^{\star} = (Q_{i1}^{\star}, ..., Q_{iT}^{\star})$ i's sequence of chosen *nests*, with $Y_{it} \in Q_{it}^{\star}$ and $Q_{it}^{\star} \in \{nest_{i1}, ..., nest_{iN}\}$. Define $\mathcal{Q}_i^{\star} = \times_{t=1}^{T} Q_{it}^{\star}$ and note that $\mathcal{Q}_i^{\star} \subseteq \mathcal{CS}_i^{\star} = \times_{t=1}^{T} CS_{it}^{\star}$, so that for each t one has $Q_{it}^{\star} \subseteq CS_{it}^{\star}$. The econometrician observes the realization of Y_i and knows to which of the original nests each Y_{it} belongs, for example $Y_{it} \in nest_n$, but does not know much about $Q_{it}^{\star}$ beyond the facts that $Y_{it} \in Q_{it}^{\star}$ and that $Q_{it}^{\star} = nest_{in} \subseteq nest_n$.

Assumption 4. Conditional on all $V(X_{ij_tt}, \theta) = X_{ij_tt}\theta$'s and on $\mathcal{CS}_i^{\star} = c$, $\Pr[Y_i = j, Q_i^{\star} = q | \mathcal{CS}_i^{\star} = c, \theta, \lambda] = \prod_{t=1}^{T} \Pr[Y_{it} = j_t, Q_{it}^{\star} = q_t | CS_{it}^{\star} = c_t, \theta, \lambda]$ is a product of T per-period nested logits as in equation (H.1) below with $\lambda = (\lambda_1, ..., \lambda_N)$ being the nesting parameters associated to each nest.

Suppose that $Y_i = j$, $Q_i^{\star} = q$, and $\mathcal{CS}_i^{\star} = c$. The nested logit model can be expressed as:

$$
\begin{aligned}
\Pr\left[Y_i = j, Q_i^\star = q \middle| \mathcal{CS}_i^\star = c, \theta, \lambda\right] &= \prod_{t=1}^{T} \overbrace{\Pr\left[Y_{it} = j_t, Q_{it}^\star = q_t \middle| CS_{it}^\star = c_t, \theta, \lambda\right]}^{\text{per-period nested logit}} \\
&= \prod_{t=1}^{T} \Pr\left[Y_{it} = j_t \middle| CS_{it}^\star = c_t, Q_{it}^\star = q_t, \theta, \lambda\right] \times \Pr\left[Q_{it}^\star = q_t \middle| CS_{it}^\star = c_t, \theta, \lambda\right] \\
&= \overbrace{\prod_{t=1}^{T} \Pr\left[Y_{it} = j_t \middle| Q_{it}^\star = q_t, \theta, \lambda\right]}^{\text{within-nest MNL}} \times \underbrace{\prod_{t=1}^{T} \Pr\left[Q_{it}^\star = q_t \middle| CS_{it}^\star = c_t, \theta, \lambda\right]}_{\text{between-nest MNL}} \\
&= \prod_{t=1}^{T} \frac{\exp\left(X_{ij_tt}\theta/\lambda_{q_t}\right)}{\sum_{v_t \in q_t} \exp\left(X_{iv_tt}\theta/\lambda_{q_t}\right)} \times \prod_{t=1}^{T} \Pr\left[Q_{it}^\star = q_t \middle| CS_{it}^\star = c_t, \theta, \lambda\right],
\end{aligned}
\tag{H.1}
$$

which is a function of the unobserved realizations q and c. When all the nesting parameters equal one, $(\lambda_1, ..., \lambda_N) = 1$, then the nested logit in (H.1) simplifies to a standard MNL. In order to test for this hypothesis, it is enough to obtain a consistent estimator of the $(\theta/\lambda_n)_{n=1}^N$ parameters of the *within-nest* MNL model:

$$
\begin{aligned}
\Pr\left[Y_i = j \middle| \mathcal{Q}_i^\star = \times_{t=1}^T q_t, \theta, \lambda\right] &= \prod_{t=1}^{T} \Pr\left[Y_{it} = j_t \middle| Q_{it}^\star = q_t, \theta, \lambda\right] \\
&= \prod_{t=1}^{T} \frac{\exp\left(X_{ij_tt}\theta/\lambda_{q_t}\right)}{\sum_{v_t \in q_t} \exp\left(X_{iv_tt}\theta/\lambda_{q_t}\right)}
\end{aligned}
\tag{H.2}
$$

and check whether $\theta/\lambda_m = \theta/\lambda_n$ for all $m \neq n$.[52] We define as a sufficient set for the within-nest MNL any correspondence that satisfies the following condition.

Condition 3. Given any choice sequence $Y_i \in \mathcal{Q}_i^\star \subseteq \mathcal{CS}_i^\star$, the correspondence f is such that $Y_i \in f(Y_i)$ and $f(Y_i) \subseteq \mathcal{Q}_i^\star$, with $f(Y_i) = \times_{t=1}^T f_t(Y_i)$ and each $f_t(Y_i)$ so that $Y_{it} \in f_t(Y_i) \subseteq Q_{it}^\star$.

[52] The first equality in (H.2) follows from equation (3.6), because $\mathcal{Q}_i^\star = \times_{t=1}^T q_t$.

In words, given any sequence of choices Y_i, a sufficient set f enables the econometrician to define a corresponding sequence of nest *subsets* $f(Y_i) = \times_{t=1}^{T} f_t(Y_i)$, where each $f_t(Y_i)$ is a subset of the specific $nest_{in}$ chosen by i in t. Given Assumption 4 and Condition 3, the within-nest MNL from (H.2) conditional on $f(Y_i) = r$ simplifies to:

$$\begin{aligned}
\Pr\left[Y_i = j \middle| \mathcal{Q}_i^{\star} = \times_{t=1}^{T} q_t, f(Y_i) = r, \theta, \lambda\right] &= \Pr\left[Y_i = j \middle| f(Y_i) = r, \theta, \lambda\right], \text{ because } r \subseteq \times_{t=1}^{T} q_t \\
&= \prod_{t=1}^{T} \Pr\left[Y_{it} = j_t \middle| f_t(Y_i) = r_t, \theta, \lambda\right], \text{ because of equation (3.6)} \\
&= \prod_{t=1}^{T} \frac{\exp\left(X_{ij_t t}\theta/\lambda_{r_t}\right)}{\sum_{v_t \in r_t} \exp\left(X_{iv_t t}\theta/\lambda_{r_t}\right)},
\end{aligned} \tag{H.3}$$

which, a part from the parameters of interest, only depends on observed quantities. Given the assumption of choice set stability, a sufficient set compatible with Condition 3 is the Within(-nest) Full Purchase History, or WFPH. This is similar to the FPH sufficient set, but now one should separately keep track of the alternatives purchased by i over choice situations within each of the N nests. Define the set of alternatives ever purchased by i in $nest_n$ by $H_i^n = \cup_{t=1}^{T} \{Y_{it} | Y_{it} \in nest_n\}$ for $n = 1, ..., N$. Note that, for any $Y_{it} \in nest_{in}$, $H_i^n \subseteq Q_{it}^{\star} = nest_{in}$. Then any observed sequence of chosen alternatives Y_i will enable one to construct the sufficient set $f_{WFPH}(Y_i) = \times_{t=1}^{T} \{H_i^n | Y_{it} \in nest_n\}$.

The idea of the IIA test is then simple. For a given sufficient set f from Condition 3, one can estimate a variant of the SSL with a different $\theta_n = \theta/\lambda_n$ for each nest $n = 1, ..., N$. Similarly, one can estimate N separate SSL models, each from the observed choices within each of the N nests. Then, if the estimated $\theta_m \neq \theta_n$ for at least two nests $n \neq m$, the econometrician will have evidence of violations of the IIA property. The validity of this testing procedure, similar to the original one proposed by Hausman and McFadden (1984), rests on the maintained assumption that f is a valid sufficient set.

I Data Appendix

In Section 6 we present an illustrative empirical example; here we describe the data used in that empirical example in greater detail.

I.1 Purchase data

We use data from the Kantar Worldpanel (see Leicester and Oldfield (2009), Dubois et al. (2016), and Dubois et al. (2019)). Kantar collects data on purchases made on-the-go from a random selection of individuals in the households that participate in the Worldpanel. The Kantar Worldpanel on-the-go survey is collected from individuals who record purchases that they make on-the-go for immediate consumption using their mobile phone.

I.2 Advertising data

To measure advertising exposure we convert weekly advertising ("flows") into an advertising "stock"; advertising stocks are the depreciated accumulation of the flows. We use advertising data collected by AC Nielsen on TV advertising. TV advertising accounts for 61.8% of total expenditure on chocolate bar advertising over this period.

For each TV ad, we have information on the time the ad was aired, the brand that was advertised, the TV station, the duration of the ad, the cost of the ad, and the TV shows that immediately preceded and followed the ad. The time path of advertising varies across brands, and all brands have some periods of zero advertising expenditure. These non-smooth strategies are rationalised in the model of Dubé et al. (2005) when the effectiveness of advertising can vary over time. This variation in the timing of adverts, coupled with variation in TV viewing behaviour, generates household level variation in exposure to brand level advertising.

Our advertising measure follows Goeree (2008) and Dubois et al. (2016) and measures advertising exposure at the individual level. We use detailed information about when individual adverts were aired on television matched with self-reported viewing information to construct individual level measures of exposure to brand advertising. We use data from the Kantar media survey, an annual survey asking the main shopper in the household about their TV subscriptions and TV viewing behaviour. Households are asked "How often do you watch ...?" for 206 different TV shows, and can choose to

answer Never, Hardly Ever, Sometimes or Regularly. At least one ad for chocolate is shown before, during, or after 112 of these shows (many of the shows with no chocolate advertising are on BBC channels, which are prohibited from showing ads). From this information we define the variable:

$$w_{is} = \begin{cases} 1 & i \text{ reports they "regularly" or "sometimes" watch show } s \\ 0 & \text{otherwise} \end{cases} \tag{I.1}$$

Households are also asked "How often do you watch ...?" 65 different TV channels and when they usually watch TV. In particular, for weekdays, Saturday, and Sunday and for 9 different time periods,[53] households are asked questions like "Do you watch live TV on Saturdays at breakfast time (6.00-9.30am)?" In each case, the household can answer Never, Hardly Ever, Sometimes or Regularly. We use this information, along with information on where the household lives (some TV channels are regional), to construct the variable:

$$w_{ikc} = \begin{cases} 1 & i \text{ says they "regularly" or "sometimes" watch on the day and time slot } k \\ & \quad \text{and "regularly" or "sometimes" watch channel } c \\ & \quad \text{and they live in the region in which } c \text{ is aired (or the channel is national)} \\ 0 & \text{otherwise} \end{cases} \tag{I.2}$$

We combine the data on household viewing behaviour with the detailed data on individual ads to create a household-specific measure of exposure to advertising. Variation in TV viewing behaviour creates considerable variation in the timing and extent of exposure an individual household has to ads of a specific brand. This leads to cross-household variation in advertising exposure that is plausibly unrelated to idiosyncratic shocks to demand for chocolate products.

Denote by T_{bskct} the duration of time that an ad for brand b is shown during show s on day and time slot k on channel c during week t. From the viewing data, we construct an indicator variable of whether household i was likely to be watching channel c on day and time slot k during show s, w_{iskc}. If show s is among the 206 specific shows households were asked for viewing information we set $w_{iskc} = w_{is}$, otherwise we set $w_{iskc} = w_{ikc}$. From this we define the household's total exposure to

[53] Breakfast time 6.00am-9.30am, Morning 9.30am-12.00 noon, Lunchtime 12.00 noon-2.00pm, Early afternoon 2.00pm-4.00pm, Late afternoon 4.00pm-6.00pm, Early evening 6.00pm-8.00pm, Mid evening 8.00pm-10.30pm, Late evening 10.30-1.00am and Night time 1.00am-6.00am.

advertising of brand b during week t (their weekly advertising "flow") as:

$$s_{ibt} = \sum_{s,c,k} w_{iskc} T_{bskct}. \tag{I.3}$$

We define a household's accumulated advertising *stock* to brand b in week t as the depreciated accumulation of these advertising flows:

$$stock_{ibt} = \sum_{k=0}^{t} \eta^k s_{ibt-k} \tag{I.4}$$

where $\eta = 0.75$

This stock is measured in seconds (and is divided by 1000 when included in the regression). It is 0 for individuals that do not watch TV, or only watch public TV (the BBC), and has a mean of 10 minutes of cumulated exposure to adverts for a particular brand.

Finally, we follow Dubé et al. (2005) and allow for diminishing returns to advertising by transforming the stock of advertising, $stock_{ibt}$, using the log inverse hyperbolic sine function, $\ln a_{ibt} = ln\left(stock_{ibt} + \sqrt{1 + stock_{ibt}^2}\right)$.