UNIVERSITÀ DEGLI STUDI DI PALERMO
Dottorato in Scienze Economiche e Statistiche
Dipartimento di Scienze Economiche, Aziendali e Statistiche
Scienza delle Finanze (SECS-P/03)

Three Essays in Microeconometrics

IL DOTTORE
Rosario Guaia

IL COORDINATORE
Prof. Andrea Consiglio

IL TUTOR
Prof. Valentino Dardanoni

CICLO XXXII
ANNO CONSEGUIMENTO TITOLO 2020

Three Essays in Microeconometrics

a dissertation presented

by

Rosario Guaia

to

The Department of Economics, Business and Statistics

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Economics and Statistics

Università degli Studi di Palermo

Palermo, Italy

February 2020

Thesis advisor: Professor Valentino Dardanoni Rosario Guaia

Three Essays in Microeconometrics

Abstract

This dissertation examines three distinct issues using microeconometric techniques. The first two chapters fall in the realm of discrete choice models and try to make allowance for limited attention. The third chapter focuses on firm behavior and investigates the impact of ownership concentration on productivity.

Chapter 1 predominantly builds on the consideration capacity model in Dardanoni, Manzini, Mariotti and Tyson (2019). In the attempt to behavioralize rational choice theory, their model identifies the distribution of cognitive characteristics in a population of agents who are observed choosing repeatedly from a single menu. By exploiting algebraic arguments, we first generalize the identification result. Then, we propose an Expectation-Maximization algorithm which is able to recover the distribution of individuals' cognitive characteristics in a non-parametric framework. The algorithm is applied to both simulated and real market data. The first application is meant to show that model primitives can be estimated with a high degree of accuracy. In the second one, instead, it is shown that a substantial fraction of individuals is either low or moderate attentive thereby contradicting full rationality which would require subjects to pay attention to all the alternatives of a given menu.

Chapter 2 resorts to a parametric setup and investigates asset allocation choices in defined contribution plans of a sample of U.S. workers. We propose a multinomial logit model in which the choice of a given financial instrument is preceded by a probabilistic consideration set formation process. Our results show that failing to recognize the relevance of limited attention can induce misleading evaluation of the effects of demographic and job-specific characteristics on the process through which workers decide how to allocate their contributions.

Chapter 3 analyzes the relationship between ownership structure and firm performance using total factor productivity (TFP) to measure firm value. Adopting a structural approach à la Olley and Pakes (1996), we proposes a semi-parametric model which controls for firms' unobserved heterogeneity and for the endogeneity of not only input factors but also of a relevant corporate governance variable, namely ownership concentration. The method is applied to Italian manufacturing data coming from the ORBIS dataset which are enriched with information on ownership provided by the Italian Securities Commission (CONSOB). Results highlight the presence of a non-monotonic relation between productivity and the degree of ownership concentration (an inverted U-shaped relationship). We argue that our findings depend on the interplay between the monitoring dimension and shareholder conflict dimension associated with ownership concentration.

Contents

Listing of figures

List of Tables

Acknowledgments

I wish to express my gratitude to my supervisor Professor Valentino Dardanoni for him being an invaluable source of guidance, support and inspiration during my training as a PhD student. I am also indebted to Professors Walter Distaso, Marco Mariotti, Paola Manzini and Annalisa Russino for their helpful suggestions. I would also like to thank all the people at the Department of Economics, Business and Statistics of the University of Palermo for the friendly environment and the help that they always gave me. A special thank goes to my PhD colleagues who shared my enthusiasm and my frustration during these years.

My greatest thank goes to my family and Cristiana for their patience and unconditional love.

Declarations

This thesis is the result of research I have conducted as a PhD student in the Department of Economics, Business and Statistics at the University of Palermo. Chapter 1 is written in co-authorship with Professor Valentino Dardanoni. Chapter 3 is written in co-authorship with Professor Valentino Dardanoni and Professor Annalisa Russino. Chapter 2 is sole-authored. During the development of all the work I benefited from my supervisor's useful advice and comments.

1

Non-parametric estimation of a consideration capacity model

1.1 Introduction

Revealed preference theory has proved extremely influential in economics and has been applied to several fields. As pioneered by Samuelson (1938), such theory holds that a given alternative a is revealed to be preferred to another alternative b if and only if a is chosen when b is also available. This argument rests upon the implicit assumption that, when choosing, a decision maker (henceforth, DM) takes into consideration each and every feasible alternative. Assuming that DMs have (potentially) unlimited capability of evaluating and comparing alternatives is hard to justify in both market and experimental settings. As a consequence, standard revealed preference theory is unable to reconcile frequently observed choice reversals with a rational utility-maximizing behavior. Following Luce (1959), let $p(a, \{a, b, c\})$ be the probability distribution that alternative a is chosen when the menu consists of alternatives a, b and c. Choice reversal would emerge whenever $p(a, \{a, b, c\}) > p(b, \{a, b, c\})$ and $p(b, \{a, b\}) > p(a, \{a, b\})$. Choice reversals are incompatible with standard Random Utility Maximization and clearly contradicts Luce rule which postulates that the probability of choosing a over b depends on the relative utility of a compared to that of b. Hence, in standard revealed preference theory choice reversal would be inevitably classified as irrational behavior. This would be the case also for the popular multinomial logit model by McFadden (1973) where agents maximize their random utility which is made of a deterministic component and an additive error term following a standard Gumbel distribution. While making the model easily tractable, such a specification of the error term is unable to make allowance for the fact that agents typically consider just a (strict) subset of the full menu of available alternatives. Limited attention qualifies as a typical source of choice error. A DM who prefers a over b might well choose b even if her menu contains a simply because she is unaware that a

is also available. As argued by Aumann (2005), this behavior is still consistent with (bounded) rationality. When making online search, for instance, a DM is likely to consider just the results provided by the first couple of web pages which do not necessarily include her most preferred alternative. Then, she maximizes her preference by choosing her best alternative among the ones she pays attention to. Most of the recent literature about nonstandard models of choice behavior conceptualizes the act of choice as a two-step process. In the first stage, cognitive limitations induce the DM to intentionally or unintentionally filters some alternatives out of the full menu so as to construct her own *consideration set*[1]. Then, in the second stage, the DM maximizes a binary preference relation as in standard models of choice behavior. In order to avoid overload of her cognitive capacity, a DM might apply different heuristics.

In Rubinstein and Salant (2006), for example, consideration set formation is sensitive to the order of presentation and/or to the frequency of appearance of a given alternative. That is to say that a DM pays attention to the first say N elements of a certain list and/or to the N elements which are more intensively advertised. Manzini and Mariotti (2007) propose a model in which a DM sequentially creates a shortlist of alternatives that are undominated according to some asymmetric binary relation and finally chooses among the alternatives which withstand that filtration process. The basic version of their model contemplates two *rationales* (i.e. two preference relations): the first one identifies a shortlist of candidates from which the second rationale identifies the final choice. To give

[1]The notion originated in the marketing literature to represent the subset of alternatives that succeed in the competition for consumers' attention (see, Wright and Barbour (1977))

an example, think of an employer who is looking for an employee with a specific skill. The employer will first discard all candidates who do not posses that skill and then select the candidate to be hired. The so called ”all or nothing” behavior documented by Gourville and Soman (2005), can be thought of as a combination of the two models above. Under such a framework, a DM considers the top N alternatives according to several rankings. When buying a new washing machine, for example, one may consider only the five cheapest and most energy-efficient machines in the market.

Notice that all the models described thus far depicts the process of forming a consideration set as a deterministic rule. In doing so, they implicitly assume that a DM is aware of all available alternatives and intentionally eliminates several of them when forming her consideration set. As a result, they are not applicable to situations where limited attention arises in the form of unawareness of some alternatives which may well stem from limited cognitive capacity. More importantly, they make it impossible to pin down attention and preferences by only observing choice data.

The two limitations outlined above are overcome by Manzini and Mariotti (2014) (henceforth, MM) where the composition of the consideration set is made stochastic. That is to say that a DM considers each alternative a with a certain probability $\gamma(a)$, the so-called *attention parameter*. Given a menu A, assume a DM has a strict preference ordering $\succ$ on it e let γ be a map $\gamma : X \to (0, 1)$, with X being a non-empty finite set of alternatives. Then, the random choice rule $p_{\succ,\gamma}$

defined by MM is such that

$$p_{\succ,\gamma}(a, A) = \gamma(a) \prod_{b \in A : b \succ a} (1 - \gamma(b)), \tag{1.1}$$

for A in the domain of menus and for all a in A. Rule (1.1) implies that the preference relation $a \succ b$ is (uniquely) revealed by $p(b, A \setminus a) > p(b, A)$. That is to say that if removing a increases the probability of choosing b, then a must have a better rank than b.

MM and much of the other recent advancements in the theory of preference analysis with limited attention share the drawback of being extremely "data hungry". That is to say that identification of primitives, namely attention and preference, is achieved by assuming that a DM is observed choosing from a large number of different overlapping menus. In Masatlioglu et al. (2012), for instance, a DM is required to choose among all possible menus drawn from a universal set of alternatives. Variability of menus, while achievable in experimental settings, is often unlikely to be encountered in a market context where either a DM chooses infrequently or the menu is slow to change, or both.

The model in Dardanoni et al. (2019), which is the focus of our estimation technique, allows to identify the distribution of cognitive characteristics in a population of DMs when observing only aggregate choice behavior from a single menu. In their general framework, a DM is assigned a specific *cognitive type* which is meant to capture the cardinality of her information set and thus represents the maximum number of alternatives that a DM can actively evaluate at any given time. Contrary to previous literature, this model enjoys a high degree of empirical feasibility given that it does not require any variability in the menu of alternatives and identification of primitives (namely, preferences and cognitive heterogeneity) is achieved even when agents are observed choosing from a single, fixed menu. By exploiting algebraic arguments and results from the literature on *latent-class models*, the present work generalizes the statistical non-parametric identifiability of the model in Dardanoni et al. (2019) and proposes an estimation strategy which is applied to both simulated and market data.

The rest of this chapter is organized as follows. Section 1.2 gives an overview of the main discrete choice models incorporating limited attention that have been elaborated by recent literature. In Section 1.3, we introduce the consideration capacity model in Dardanoni et al. (2019). Section 1.4 lays out our generic identification result. Section 1.5 discusses our estimation strategy and validates it through simulations and an application to real data. Section 1.6 concludes.

1.2 Limited attention models

Consideration set models generalize standard discrete choice models by relaxing the assumption that agents take into consideration each feasible alternative. The marketing literature contemplates several models which specify a probability that single options or subsets of them are considered by DMs. In these models, choice sets are latent meaning that they cannot be defined with certainty on the basis of observational data. Following Manski (1977), such models assign to the choice problem the following probabilistic specification:

$$P_i(j) = \sum_{C \in X(j)} P_i(j|C) P_i(C|X), \tag{1.2}$$

where :

- $P_i(j)$ is the probability of individual i choosing alternative j;
- $X(j)$ is the set of all possible choice sets containing alternative j;
- $P_i(j|C)$ is the probability of individual i choosing j given that her choice set is C;
- $P_i(C|X)$ is the probability of C being the choice set of individual i.

The main challenge when trying to implement (1.2) is the proliferation of possible choices sets C as the number of alternatives grows. Hauser and Wernerfelt (1990), for example, describe a sequential process in which a DM includes an additional brand to her consideration set if the expected incremental utility of choosing from a richer set at each consumption occasion exceeds the incremental cost arising from searching and evaluating a new brand. Ben-Akiva and Boccara (1995) formulate a constraint-based approach to choice set formation which postulates that individuals exclude from further consideration available alternatives not meeting some given criteria.

Accumulated empirical literature typically relies on cross-sectional dataset where a given sample of individuals is observed choosing among several alternatives in just one occasion. The most challenging task in estimating consideration set models is to show the identifiability of the parameters of interest, namely attention and preference parameters. Identifiability can naturally be obtained by using auxiliary data (typically, survey data) on what products are and are not considered by consumers (e.g., Draganska and Klapper (2011) and Honka and Chintagunta (2017)).

Goeree (2008), instead, develops a model of limited information for the market of Personal Computers where advertising determines the set of products entering a DM's consideration set. Thus, identification rests on the questionable assumption that advertising impacts attention but does not exercise any influence on consumers' utility. Other examples of models where similar exclusion restrictions are imposed in order to achieve identification can be found in airport choice (Basar and Bhat (2004)), retail electricity choice (Hortaçsu et al. (2017)) and health insurance plan switching (Heiss et al. (2013)).

Abaluck and Adams (2017) prove that preference and attention parameters can be retrieved by just exploiting results from economic theory without any need for exclusion restrictions. In their model, all observables are allowed to have an impact on both attention and utility. In a model of limited attention, in fact, the so-called Slustky symmetry, i.e. the symmetry of the matrix of cross-derivatives of choice probabilities with respect to characteristics of rival goods, is expected to hold only when conditioning on a specific choice set. When looking at unconditional probabilities (i.e. market shares) such a symmetry is showed to be violated and these deviations can be fruitfully exploited to constructively identify consideration probabilities. In proving identification of the model parameters, Abaluck and Adams (2017) do not make any parametric assumption even though they recognize that nonparametric estimation would be extremely data demanding in terms of markets and individuals and, as such, some parametric structure would need to be imposed especially when the number of alternatives is high.

The model in Dardanoni et al. (2019) adds to previous literature on limited attention in that it is not primarily concerned with modelling the probability of a given alternative being considered but rather focuses on the distribution of cognitive characteristics. Indeed, consideration capacity is modelled in the form of the cognitive burden that a DM is able to bear as proxied by the number of different alternatives that she is able or willing to include in her consideration set.[2] Moreover, recent empirical models on limited attention typically require the existence of a default option[3] while the model in Dardanoni et al. (2019) can easily accommodate situations where such a default does not exist.[4] Empirical feasibility is further increased by the fact that identification of the distribution of cognitive characteristics is obtained prior to any econometric specification and, contrarily to other contributions (e.g., Abaluck and Adams (2017), Barseghyan et al. (2019)), without any need for variation in the observable characteristics of available alternatives.

[2]Dardanoni et al. (2019) show that their consideration capacity model encompasses as a special case the consideration probability model in Manzini and Mariotti (2014).

[3]The default option is assigned to a DM when she does not select any of the available alternatives.

[4]This is usually the case when dealing with market data at the individual level (e.g., scanner data) which contains only information on what individuals actively choose.

1.3 The consideration capacity model

This section is intended to give a pedagogical overview of the model in Dardanoni et al. (2019). In order to stress on the applicability of the model to market data we do not contemplate a default option throughout the exposition. The generalization with a default option can be found in Dardanoni et al. (2019).

1.3.1 Homogeneous preferences

Assuming the standard framework of discrete choice models, let X be the finite universal set of alternatives and A any non-empty *menu* such that $A \subseteq X$. X is required to be collectively exhaustive while alternatives in A are mutually exclusive. This means that, when facing A , a DM chooses exactly one of the alternative in A. The number of possible consideration sets (i.e. menus) grows exponentially with the cardinality of X and is equal to $(2^{|X|} - 1)$.

In order to model cognitive heterogeneity, each DM is assigned a cognitive type $\gamma \in \Gamma = \{1, 2, ..., |X|\}$, drawn independently from a distribution F. Since our interest is in modelling cognitive capacity, we assume that when $1 \leq \gamma < |X|$, a DM is equally likely to consider each $A \subset X$ such that $|A| = \gamma$. When $\gamma \geq |X|$, instead, we can be sure that a DM's consideration set coincides with the full set of available alternatives X. For the time being, assume that all decision makers share the same linear preference order $\succ$ on X,[5] which without loss of generality is assumed to be of the form $1 \succ 2 \succ ... \succ n$. This preference relation allows us to sort alternatives from the most preferred to the least preferred one for every menu

[5]We will relax this assumption later by introducing preference heterogeneity.

$A \subseteq X$, such that the kth best option occupies the kth position. Then, a menu A is ordered according to $A = \{1_A, 2_A, ..., k_A, ..., |A|\}$, while the ordered universal set reads as $X = \{1, 2, ..., k, ..., |X|\}$. Letting $|X| = n \geq 2$,[6] the assumption of common preference implies that type-conditional choice frequencies are given by:

$$p_\gamma(k) = \begin{cases} \binom{n-k}{n-1} & \text{if } \gamma \geq n, \\ \binom{n-k}{\gamma-1} / \binom{n}{\gamma} & \text{if } 1 \leq \gamma < n. \end{cases} \tag{1.3}$$

For $\gamma \geq n$,[7] the consideration set coincides with X and the maximization of $\succ$ leads full attentive DMs to always choose option $k = 1$ (i.e. the most preferred alternative in X).

For $1 \leq \gamma < n$, DMs select option k whenever no other option l preferred to k (i.e. $l < k$) is available. That is to say that, among all the possible $\binom{n}{\gamma}$ candidate sets of cardinality γ, k is chosen whenever it is undominated by any other alternative which occurs $\binom{n-k}{\gamma-1}$ of the times. Note that for $\gamma > n - k + 1$, k is never chosen since a DM has enough consideration capacity to always include alternatives better than k in her consideration set. The probabilistic nature of the population cognitive characteristics allows us to write aggregate shares (or, equivalently, unconditional choice frequencies) as

$$p(k) = \sum_{\gamma=1}^{n-k+1} \frac{\binom{n-k}{\gamma-1}}{\binom{n}{\gamma}} \pi(\gamma). \tag{1.4}$$

The proportion in the population of type γ individuals is represented by $\pi(\gamma)$ and

[6]The case $|X| = 1$ is clearly of no interest.

[7]Notice that $\gamma > n$ is observationally indistinguishable from $\gamma = n$ in this model.

is given by the following probability masses:

$$\pi(\gamma) = \begin{cases} F(1) & \text{if } \gamma = 1, \\ F(\gamma) - F(\gamma - 1) & \text{if } 1 < \gamma < n, \\ 1 - F(\gamma - 1) & \text{if } \gamma \geq n. \end{cases} \tag{1.5}$$

Proposition 1 (Dardanoni et al., 2019, p. 11) *Under homogeneous tastes, the preference relation $\succ$ is fully revealed by aggregate choice data, i.e. $p(k) \geq p(k+1)$, with $1 \leq k < n$.*

Exploiting (1.4), one obtains

$$p(k) - p(k+1) = \frac{\pi(n-k+1)}{\binom{n}{n-k+1}} + \sum_{\gamma=2}^{n-k} \frac{\binom{n-k-1}{\gamma-2}}{\binom{n}{y}} \pi(\gamma). \tag{1.6}$$

The difference in (1.6) is always non-negative being the sum of non-negative terms. Moreover, allowing $\sum_{\gamma=2}^{n-k+1} \pi(\gamma) > 0$ implies that $p(k) > p(k+1)$ always holds true. Indeed, the two addends in (1.6) are evocative of the effects which are at work when we take one step up in the preference scale established by $\succ$. The first addend in (1.6) says that for $\gamma = n-k+1$, only alternative k has a (possibly) positive probability of being chosen. As for the second addend, we have that k is never chosen less frequently than option $k+1$ for $2 \leq \gamma \leq n-k$. This is so because, being $k \succ k+1$, at any level of cognitive capacity between 2 and $n-k$ the number of alternatives inferior to k always exceeds the number of alternatives inferior to $k+1$. Finally note that it is enough to assume that $\pi(2) > 0$ for the preference relation to be strictly revealed by aggregate choice frequencies. Indeed, (1.6) implies that if $\pi(2) > 0$ then $p(1) > p(2) > ... > p(n)$ is always satisfied.

Unconditional frequencies in (1.4) are linear functions of the probability masses $\pi(\gamma)$ and can be conveniently expressed in matrix form as

$$\underbrace{\begin{bmatrix} p(1) \\ \vdots \\ p(k) \\ \vdots \\ p(n) \end{bmatrix}}_{\mathbf{p}} = \underbrace{\begin{bmatrix} \frac{1}{n} & \cdots & \frac{\gamma}{n} & \cdots & 1 \\ \vdots & & \vdots & & \vdots \\ \frac{1}{n} & \cdots & \frac{\binom{n-k}{\gamma-1}}{\binom{n}{\gamma}} & \cdots & 0 \\ \vdots & & \vdots & & \vdots \\ \frac{1}{n} & \cdots & 0 & \cdots & 0 \end{bmatrix}}_{\mathbf{C}} \underbrace{\begin{bmatrix} \pi(1) \\ \vdots \\ \pi(\gamma) \\ \vdots \\ \pi(n) \end{bmatrix}}_{\boldsymbol{\pi}}, \tag{1.7}$$

where the generic γ-th column of $\mathbf{C}$ represents the vector of choice frequencies conditional on being of type γ.

Proposition 2 (Dardanoni et al., 2019, pp. 17-18) *Under homogeneous tastes, probability masses $\langle\pi(\gamma)\rangle_{\gamma=1}^{n}$ specifying the distribution of cognitive capacity among the population are uniquely identified by aggregate choice frequencies $\langle p(k)\rangle_{k=1}^{n}$.*

The matrix $\mathbf{C}$ is always invertible being an upper anti-triangular matrix with nonzero entries on the main anti-diagonal. $\mathbf{C^{-1}}$ turns to be a lower anti-triangular matrix with generic entry ${\mathbf{C^{-1}}}_{(\gamma,k)}$ given by :

$${\mathbf{C^{-1}}}_{(\gamma,k)} = \begin{cases} \binom{n}{\gamma}(-1)^{(\gamma-1)-(n-k)}\binom{\gamma-1}{n-k} & \text{if } k > n-\gamma, \\ 0 & \text{otherwise.} \end{cases} \tag{1.8}$$

Hence, we can always retrieve $\boldsymbol{\pi}$ from $\boldsymbol{\pi} = \mathbf{C^{-1}p}$ and its generic element $\pi(\gamma)$

reads as

$$\pi(\gamma) = \binom{n}{\gamma} \sum_{k=n-\gamma+1}^{n} (-1)^{(\gamma-1)-(n-k)} \binom{\gamma-1}{n-k}. \tag{1.9}$$

1.3.2 Heterogeneous but known preferences

The previous section assumes that all DMs share the same preference relation which, without loss of generality, is assumed to be of the form $1 \succ 2 \succ ... \succ n$. Here we relax this assumption by allowing preference heterogeneity. That is to say that, given a universal set of alternatives X with cardinality $|X| = n$, each DM chooses by maximizing one of the possible $n!$ preference orderings. Given $X = \{1, 2, ..., n\}$, let us define $\varphi : X \to \{1, 2, ..., n\}$ as the map that associates each option in X with its preference rank. Each of the possible $\langle \varphi_h \rangle_{h=1}^{n!}$ rankings can be conveniently expressed by means of a suitable $n \times n$ permutation matrix $\mathbf{P}(h)$. For instance, let $n = 3$ and let $\varphi_2 = (2, 1, 3)$ (i.e. $2 \succ 1 \succ 3$). The unique permutation matrix associated to φ_2 is

$$\mathbf{P}(2) = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \tag{1.10}$$

where each k-th column has a one in position l if $\varphi_2(l) = k$.

Each map φ_h implies a unique matrix of type-conditional choice frequencies $\mathbf{C}_h$

which turns to be given by

$$\mathbf{C}_h = \mathbf{P}(h)\mathbf{C}. \tag{1.11}$$

Thus, letting τ_h be the (known) frequency with which ranking φ_h appears in the population, unconditional choice frequencies in (1.7) can be rewritten as

$$\begin{aligned} \mathbf{p} &= \widehat{\mathbf{C}}\boldsymbol{\pi} \\ &= \left[\sum_{h=1}^{n!} \tau_h \mathbf{C}_h\right] \boldsymbol{\pi} \\ &= \left[\sum_{h=1}^{n!} \tau_h \mathbf{P}(h)\right] \mathbf{C}\boldsymbol{\pi} \end{aligned} \tag{1.12}$$

In order to identify the distribution of cognitive types $\boldsymbol{\pi}$ from aggregate choice shares, matrix $\widehat{\mathbf{C}}$ needs to be non-singular. Since $\mathbf{C}$ is always invertible, non-singularity of $\widehat{\mathbf{C}}$[8] amounts to non-singularity of $\mathbf{B} = \widehat{\mathbf{C}}\mathbf{C}^{-1} = \sum_{h=1}^{n!} \tau_h \mathbf{P}(h)$.[9] The latter is a doubly stochastic matrix[10] being a convex combination of permutation matrices.[11] Matrix $\mathbf{B}$ is clearly not invertible for every $\boldsymbol{\tau}$. Assuming, for instance,

[8] $\widehat{\mathbf{C}}$ generalizes $\mathbf{C}$ to the case of heterogeneous preferences. As for $\mathbf{C}$, the generic element $\widehat{\mathbf{c}}_{ij}$ of $\widehat{\mathbf{C}}$ is the frequency with which alternative i is chosen conditional on being of type j.

[9] Note that $\mathbf{B}$ can be seen as a sort of "average" permutation matrix. Its generice element b_{ij} equals $\sum_{h:\varphi_h(i)=j} \tau_h$ which corresponds to the frequency with which alternative i is ranked in position j.

[10] A doubly stochastic matrix (also called bistochastic) is a square matrix $A = (a_{ij})$ of non-negative real numbers, each of whose rows and columns sums to 1, i.e.

$$\sum_i a_{ij} = \sum_j a_{ij} = 1$$

.

[11] The Birkhoff–von Neumann theorem states that the set of $n \times n$ doubly stochastic matrices

$\langle\tau\rangle_{h=1}^{n!} = \frac{1}{n!}$, we obtain a singular $\mathbf{B}$ matrix with all entries equal to $\frac{1}{n}$. However, standard results in measure theory legitimate us to conclude that matrix $\mathbf{B}$ is generically invertible.

Proposition 3 (Dardanoni et al., 2019, p. 25) *Under known preference heterogeneity, probability masses* $\langle\pi(\gamma)\rangle_{\gamma=1}^{n}$ *are uniquely identified by aggregate choice frequencies* $\langle p(k)\rangle_{k=1}^{n}$ for almost all $\boldsymbol{\tau}$.

To support last proposition, note that $\det(\mathbf{B})$ can be seen as a real-valued polynomial function $f(\boldsymbol{\tau})$ on a Euclidean space, with $\boldsymbol{\tau} \in [0,1]^{n!}$. This function is either identically zero or nonzero almost everywhere (see, e.g., Caron and Traynor (2005)). It suffices to resort to the case of homogeneous preferences (i.e. $\tau_h = 1$ and $\tau_{h'} = 0$ for all $h' \in \{1, 2, ..., n!\} \setminus h$) for obtaining invertibility of $\mathbf{B}$ (or, equivalently, $\det(\mathbf{B}) \neq 0$). Being $f(\boldsymbol{\tau}) = \det(\mathbf{B})$ non identically zero,[12] this determinant is different from zero almost everywhere and we achieve non-singularity of $\mathbf{B}$ for almost all $\boldsymbol{\tau}$.

1.3.3 Heterogeneous and unknown preferences: the need for a dynamical model

The assumption of known taste distribution is mostly innocuous in contexts where the researcher can separately elicit individual preferences.[13] However it is hard to

forms a convex polytope whose vertices are the $n \times n$ permutation matrices.

[12]That is to say that the zero set of the polynomial has a Lebesgue measure equal to zero.

[13]Think, for instance, of a lab experiment where the researcher can "constraint" inattention by adopting some mechanism that forces DMs to become aware of all available alternatives.

justify in most experimental and market settings where it is precisely the interplay between (limited) attention and preferences to determine individual choice behavior. The ultimate goal, here, is to retrieve in a non-parametric fashion both preferences and cognitive types distributions from aggregate shares alone. However, the following example clarifies that aggregate shares from a single choice occasion are clearly insufficient to identify model primitives.

Example 1. Let $n = 3$ and φ_h be of the form:

$$\begin{cases} \varphi_1 = (1,2,3), \\ \varphi_2 = (1,3,2), \\ \varphi_3 = (2,1,3), \\ \varphi_4 = (2,3,1), \\ \varphi_5 = (3,1,2), \\ \varphi_6 = (3,2,1). \end{cases}$$

Then, (1.12) becomes:

$$\mathbf{p} = \left[\sum_{h=1}^{6} \tau_h \mathbf{P}(h)\right] \mathbf{C}\boldsymbol{\pi} = \begin{bmatrix} \frac{1}{3} & \frac{2}{3}(\tau_1+\tau_2)+\frac{1}{3}(\tau_3+\tau_4) & (\tau_1+\tau_2) \\ \frac{1}{3} & \frac{2}{3}(\tau_3+\tau_5)+\frac{1}{3}(\tau_1+\tau_6) & (\tau_3+\tau_5) \\ \frac{1}{3} & \frac{2}{3}(\tau_4+\tau_6)+\frac{1}{3}(\tau_2+\tau_5) & (\tau_4+\tau_6) \end{bmatrix} \begin{bmatrix} \pi(1) \\ \pi(2) \\ \pi(3) \end{bmatrix}. \tag{1.13}$$

(1.13) is an underdetermined system in which separately identifying $\boldsymbol{\tau}$ and $\boldsymbol{\pi}$ is clearly hopeless. Indeed, the elements of $\boldsymbol{\pi}$ completely exhaust the two degrees of freedom[14] provided by aggregate shares $\mathbf{p}$.

1.4 Non-parametric identifiability of the multi-occasion consideration capacity model

In order to gain some knowledge on both $\boldsymbol{\tau}$ and $\boldsymbol{\pi}$, we need to abandon the single-occasion framework adopted thus far and let individuals choose across multiple occasions. As in standard dynamic discrete probability models, we assume the analyst has access to a random sample of individuals, each of them choosing exactly one option from the universe of available alternatives on several occasions. Throughout the paper, we mantain the assumption that the distribution of cognitive types F is fixed across occasions. Our typical dependent variable is given by the joint distribution of choices over time. Indexing occasions by $i = \{1, ..., I\}$ and letting $X = \{1, ..., k, ..., n\}$[15] represent the universe of available alternatives, the tensor of (joint) aggregate choice frequencies turns to be given by

$$\mathbf{p} = \langle p(k_1 \cdots k_I) \rangle_{k=1}^{n}, \tag{1.14}$$

with $\langle p(k_1 \cdots k_I) \rangle_{k=1}^{n}$ being the joint probability of choosing option k_i on each occasion i. Thus, $\mathbf{p}$ is in all respect equivalent to the realization of a multino-

[14] In both $\boldsymbol{\pi}$ and $\mathbf{p}$, one element is residually determined since they need to sum up to one.

[15] For ease of exposition, we assume that the cardinality of X and the alternatives in it remain constant across occasions i. Note that none of these assumptions is vital to our identification result and can be easily relaxed.

mial random variable[16] which is completely specified by a unique distribution of cognitive characteristics F and a "trial-specific" distribution of tastes $\boldsymbol{\tau}_i$, with $i = 1, ..., I$. Hence, random vectors $\boldsymbol{\tau}_i$ are assumed independent across choice occasions. Moreover, they determine (joint) aggregate choice frequencies conditional on being of a given type γ, i.e. $\mathbf{p}_\gamma = \langle p_\gamma(k_1 \cdots k_I) \rangle_{k=1}^n$ while observed shares $\mathbf{p}$ arise as

$$\mathbf{p} = \int_\Gamma \mathbf{p}_\gamma dF = \sum_{\gamma=1}^{n} \pi(\gamma)\mathbf{p}_\gamma, \tag{1.15}$$

where last equality is due to the discreteness of F. Each $\boldsymbol{\tau}_i$ gives rise to an occasion-specific "average" permutation matrix $\mathbf{B}_i = \sum_{h=1}^{n!} \tau_{ih}\mathbf{P}(h)$ and its associated matrix of conditional frequencies $\widehat{\mathbf{C}}_i = \mathbf{B}_i\mathbf{C}$. The assumption of independence of $\boldsymbol{\tau}_i$ across occasions propagates to $\widehat{\mathbf{C}}_i$ and allows us to rewrite (1.15) as

$$\begin{aligned} \mathbf{p} &= \sum_{\gamma=1}^{n} \pi(\gamma)\mathbf{p}_\gamma \\ &= \sum_{\gamma=1}^{n} \pi(\gamma) \otimes_{i=1}^{I} \widehat{\mathbf{C}}_i \mathbf{e}_\gamma, \end{aligned} \tag{1.16}$$

[16]Being X constant across occasions, this random variable has dimension n^I.

where the column vector $\mathbf{e}_\gamma$ has a 1 in the γ-th position and 0's elsewhere.[17] Letting $\otimes_{i=1}^{I}\widehat{\mathbf{C}}_i\mathbf{e}_\gamma = \mathbb{P}_\gamma$, (1.16) becomes

$$\mathbf{p} = \sum_{\gamma=1}^{n} \pi(\gamma)\mathbb{P}_\gamma, \tag{1.17}$$

with $\mathbb{P}_\gamma$ being a I-order tensor with $n \times \cdots \times n = n^I$ entries. From (1.17), we can interpret $\mathbf{p}$ as the distribution of a finite mixture of finite measure products, with a known number of components $n = \max_{\gamma=\{1,\ldots,n\}} \gamma$. The $\pi(\gamma)$ term represents the probability that a draw from the population is in the γ-th class. $\mathbb{P}_\gamma = \mathbf{p}_\gamma = \langle p_\gamma(k_1 \cdots k_I)\rangle_{k=1}^{n}$, instead, gives the joint distribution of the random variables $k_1, \cdots, k_I$ conditional on being of type γ. Conditioning on the type, the I observable variables $k_1, \cdots, k_I$ are independent. However, since types are not observable, independence breaks down in the unconditional distribution $\mathbf{p}$.[18] Let $|k_i|$ be the cardinality of k_i. Following Allman et al. (2009), we refer to the model in (1.17) as the n-class, I-feature model with state space $\{1,\ldots,|k_1|\} \times \cdots \times \{1,\ldots,|k_I|\}$ and denote it by $\mathcal{M}(n; |k_1|,\ldots,|k_I|)$. Note that the assumption of the menu X being fixed across occasions implies that $|k_i| = n$ for each $i = 1,\ldots,I$.[19] Let us identify the parameter space of the model

[17] Note that $\otimes_{i=1}^{I}\widehat{\mathbf{C}}_i\mathbf{e}_\gamma$ performs column-wise Kronecker product between all γ-th columns of the $\widehat{\mathbf{C}}_i$ matrices, $i = \{1,\ldots,I\}$.

[18] Hence, one-dimensional marginalizations of $\mathbf{p}$ describing the I variables k_i are generally not independent.

[19] Each variable k_i takes exactly one value in $\{1,\ldots,n\}$.

$\mathcal{M}(n; |k_1|, \dots, |k_I|)$ with a subset $\mathbf{\Theta}$ of $[0,1]^L$ with

$$\begin{aligned} L &= (n-1) + (n-1)\sum_{i=1}^{I}(|k_i| - 1) \\ &= (n-1) + (n-1)^2 I, \end{aligned} \tag{1.18}$$

being the number of free parameters. The first addend in (1.18) is the number of free parameters in $\boldsymbol{\pi}$. The second addend is the number of free parameters determining the matrices $\widehat{\mathbf{C}}_i$, for $i = \{1, \dots, I\}$, whose columns span tensors $\mathbb{P}_\gamma$, for $\gamma = \{1, \dots, n\}$. Each $\widehat{\mathbf{C}}_i$ has $(n-1)^2$ free parameters since it is a $n \times n$ left stochastic (i.e. columns add up to 1) matrix and all elements on the first column (namely, choice frequencies conditional on being of type 1) are, by construction, all equal to $\frac{1}{n}$. Moreover, the parametrization map for $\mathcal{M}(n; |k_1|, \dots, |k_I|)$ is

$$\mathbf{\Psi}_{n,I,(|k_i|)} : \ \mathbf{\Theta} \to [0,1]^{\prod_{i=1}^{I}|k_i|} = [0,1]^{n^I}. \tag{1.19}$$

In order to show strict identifiability of parameters governing finite mixture models of the form given in (1.17), we make use of the following lemma which extends *Lemma 1* in Dardanoni et al. (2019). Our lemma rests upon *Theorem 3* in Sidiropoulos and Bro (2000) which generalizes to I-order tensors ($I \geq 3$) the result of Kruskal's theorem (see, Kruskal (1977)) on the uniqueness of the decomposition of a 3-dimensional array.

Lemma 1 Consider the I-order tensor

$$\mathbf{T} = \sum_{j=1}^{n} \otimes_{i=1}^{I} \mathbf{M}_i \mathbf{e}_j \tag{1.20}$$

spanned by a collection of matrices $\langle \mathbf{M}_i \rangle_{i=1}^{I}$ each of them being of dimension $m_i \times n$. Let $\kappa_{\mathbf{M}_i}$ be the Kruskal-rank[20] of $\mathbf{M}_i$. Then, provided that

$$\sum_{i=1}^{I} \kappa_{\mathbf{M}_i} \geq 2n + (I-1), \tag{1.21}$$

matrices $\langle \mathbf{M}_i \rangle_{i=1}^{I}$ are unique up to permutation and scaling of columns. That is to say that, for any collection of matrices $\langle \widehat{\mathbf{M}}_i \rangle_{i=1}^{I}$ such that $\mathbf{T} = \sum_{j=1}^{n} \otimes_{i=1}^{I} \widehat{\mathbf{M}}_i \mathbf{e}_j$, there exists a collection of three scaling (invertible) diagonal matrices $(\mathbf{D}_{S_1}, \mathbf{D}_{S_2}, \mathbf{D}_{S_3})$ and a permutation matrix $\mathbf{P}$ such that $\widehat{\mathbf{M}}_i = \mathbf{M}_i \mathbf{D}_{S_p}^{\frac{1}{|S_p|}} \mathbf{P}$, with $\langle S_p \rangle_{p=1}^{3}$ being a suitable tripartition of the set $\{1, \ldots, I\}$, $i \in S_p$ and $\mathbf{D}_{S_1} \mathbf{D}_{S_2} \mathbf{D}_{S_3} = \mathbf{I}_n$. Setting, for instance, $I = 4$ and imposing *irreducibility*[21] of $\mathbf{T}$ in (1.20), let us assume without loss of generality that $\kappa_{\mathbf{M}_1} \geq \kappa_{\mathbf{M}_2} \geq \kappa_{\mathbf{M}_3} \geq \kappa_{\mathbf{M}_4}$. Then, *Theorem 3* in Sidiropoulos and Bro (2000) shows that (1.20) can be equivalently expressed as

$$\mathbf{T} = \sum_{j=1}^{n} \otimes_{S_p=S_1}^{S_3} \mathbf{N}_{S_p} \mathbf{e}_j, \tag{1.22}$$

[20]Given $\mathbf{A} \in \mathbb{R}^{I \times F}$, $r_{\mathbf{A}} := rank(\mathbf{A}) = r$ iff it contains *at least* a collection of r linearly independent columns, and this fails for $r+1$ columns. $\kappa_{\mathbf{A}}$(the Kruskal-rank of $\mathbf{A}$) $= r$ iff *every* r columns are linearly independent, and this fails for at least one set of $r+1$ columns ($\kappa_{\mathbf{A}} \leq r_{\mathbf{A}} \leq \min(I, F)$ for all $\mathbf{A}$).

[21]*Irreducibility* of $\mathbf{T}$ means that the generic element of the 4-way array implied by $\mathbf{T}$ cannot be expressed using fewer than $I = 4$ components. *Irreducibility* is trivially satisfied if $\kappa_{\mathbf{M}_i} \geq 1$ for each i.

with $\mathbf{N}_{S_1} = \mathbf{M}_1$, $\mathbf{N}_{S_2} = \mathbf{M}_2$ and $\mathbf{N}_{S_3} = \mathbf{M}_3 \odot \mathbf{M}_4$.[22] Moreover, (1.22) implies a tripartition of $\{1, 2, 3, 4\}$ of the form $S_1 = \{1\}$, $S_2 = \{2\}$ and $S_3 = \{3, 4\}$. If *irreducibility* holds true, one can recursively apply the reasoning above in order to obtain a trilinear representation as the one in (1.22) of any I-order tensor with $I > 4$. Kruskal's result guarantees the uniqueness of matrices $\langle \mathbf{N}_{S_p} \rangle_{S_p=S_1}^{S_3}$ up to permutation and scaling of columns. Then, applying *Theorem 3* in Rhodes (2010), we can re-express (1.22) as

$$\mathbf{T} = \sum_{j=1}^{n} \otimes_{S_p=S_1}^{S_3} \widehat{\mathbf{N}}_{S_p} \mathbf{e}_j, \tag{1.23}$$

with $\widehat{\mathbf{N}}_{S_p} = \mathbf{N}_{S_p} \mathbf{D}_{S_p} \mathbf{P}$. Under *irreducibility* of $\mathbf{T}$, uniqueness of $\langle \mathbf{N}_{S_p} \rangle_{S_p=S_1}^{S_3}$ implies uniqueness of $\langle \mathbf{M}_i \rangle_{i=1}^{I}$ (see, Sidiropoulos and Bro (2000)). Appropriately using $\mathbf{D}_{S_p}$ and $\mathbf{P}$, we can retrieve $\langle \widehat{\mathbf{M}}_i \rangle_{i=1}^{I}$ as $\widehat{\mathbf{M}}_i = \mathbf{M}_i \mathbf{D}_{S_p}^{\frac{1}{|S_p|}} \mathbf{P}$ for $i \in S_p$. Indeed, in the case of $I = 4$ depicted above, we have $\widehat{\mathbf{M}}_1 = \mathbf{M}_1 \mathbf{D}_{S_1} \mathbf{P}$ and $\widehat{\mathbf{M}}_2 = \mathbf{M}_2 \mathbf{D}_{S_2} \mathbf{P}$ since $S_1 = \{1\}$, $S_2 = \{2\}$ and $|S_1| = |S_2| = 1$. As for $\widehat{\mathbf{M}}_3$ and $\widehat{\mathbf{M}}_4$, instead, we have

$$\begin{aligned} \widehat{\mathbf{N}}_{S_3} &= \mathbf{N}_{S_3} \mathbf{D}_{S_3} \mathbf{P} \\ &= (\mathbf{M}_3 \odot \mathbf{M}_4) \mathbf{D}_{S_3} \mathbf{P} \\ &= (\mathbf{M}_3 \mathbf{D}_{S_3}^{\frac{1}{2}} \mathbf{P} \odot \mathbf{M}_4 \mathbf{D}_{S_3}^{\frac{1}{2}} \mathbf{P}), \end{aligned} \tag{1.24}$$

which implies $\widehat{\mathbf{M}}_3 = \mathbf{M}_3 \mathbf{D}_{S_3}^{\frac{1}{2}} \mathbf{P}$ and $\widehat{\mathbf{M}}_4 = \mathbf{M}_4 \mathbf{D}_{S_3}^{\frac{1}{2}} \mathbf{P}$.

[22] $\odot$ is the Khatri-Rao product, i.e. the column-wise Kronecker product, between $\mathbf{M}_3$ and $\mathbf{M}_4$. Thus, $\mathbf{N}_{S_3}$ has dimension $(m_3 m_4) \times n$.

Lemma 1 is immediately applicable to our multi-occasion consideration capacity model $\mathcal{M}(n; |k_1|, \ldots, |k_I|)$. Writing $\mathbf{D}(\boldsymbol{\pi})$ for the diagonal matrix with entries $\boldsymbol{\pi} = \langle \pi(\gamma) \rangle_{\gamma=1}^{n}$, we can impose $\mathbf{M}_1 = \widehat{\mathbf{C}}_1 \mathbf{D}(\boldsymbol{\pi})$ and $\mathbf{M}_i = \widehat{\mathbf{C}}_i$ for each $i \geq 2$. Then, using (1.20), we can equivalently express (1.16) as

$$\mathbf{p} = \sum_{j=1}^{n} \otimes_{i=1}^{I} \mathbf{M}_i \mathbf{e}_j. \tag{1.25}$$

Provided $\boldsymbol{\pi} \gg \mathbf{0}$, $\langle \mathbf{M}_i \rangle_{i=1}^{I}$ are all $n \times n$ invertible[23] matrices and condition (1.21) becomes[24]

$$In \geq 2n + (I - 1), \tag{1.26}$$

Proposition 4. If $I \geq 3$, then $\mathbf{M}_1 = \widehat{\mathbf{C}}_1 \mathbf{D}(\boldsymbol{\pi})$ and $\mathbf{M}_i = \widehat{\mathbf{C}}_i$ with $i = \{2, ..., I\}$ are unique up to permutation and scaling of columns.

Let us restate (1.26) as

$$I \geq 2 + \frac{1}{n-1}. \tag{1.27}$$

The case $n = 1$ is clearly of no interest. For $n \geq 2$, $\frac{1}{n-1} \leq 1$ and thus $I \geq 3$ suffices to guarantee the applicability of Lemma 1 to the model $\mathcal{M}(n; |k_1|, \ldots, |k_I|)$.

Proposition 5 In the consideration capacity model with unknown preference heterogeneity $\mathcal{M}(n; |k_1|, \ldots, |k_I|)$, if $\boldsymbol{\pi} \gg \mathbf{0}$ and $I \geq 3$ then for almost all taste

[23]Since, by Proposition 3, $\langle \widehat{\mathbf{C}}_i \rangle_{i=1}^{I}$ are invertlible for almost all distribution of $\langle \boldsymbol{\tau}_i \rangle_{i=1}^{I}$.
[24]If $\mathbf{A} \in \mathbb{R}^{n \times n}$ is invertible, then $\kappa_{\mathbf{A}} = r_{\mathbf{A}} = n$.

distributions $\langle \boldsymbol{\tau}_i \rangle_{i=1}^{I}$ matrices $\mathbf{M}_1 = \widehat{\mathbf{C}}_1\mathbf{D}(\boldsymbol{\pi})$ and $\mathbf{M}_i = \widehat{\mathbf{C}}_i$ with $i = \{2, ..., I\}$ are uniquely determined by the joint choice shares $\mathbf{p} = \langle p(k_1 \dots k_I) \rangle_{k=1}^{n}$.

Proof. Lemma 1 legitimates us to rewrite (1.25) as

$$\mathbf{p} = \sum_{j=1}^{n} \otimes_{i=1}^{I} \widehat{\mathbf{M}}_i \mathbf{e}_j, \tag{1.28}$$

with

$$\widehat{\mathbf{M}}_1 = \widehat{\widehat{\mathbf{C}}}_1 \mathbf{D}(\overline{\boldsymbol{\pi}}) = \widehat{\mathbf{C}}_1 \mathbf{D}(\boldsymbol{\pi}) \mathbf{D}_{S_p:1\in S_p}^{\frac{1}{|S_p|}} \mathbf{P}, \tag{1.29}$$

and

$$\widehat{\mathbf{M}}_i = \widehat{\widehat{\mathbf{C}}}_i = \widehat{\mathbf{C}}_i \mathbf{D}_{S_{p'}:i\in S_{p'}}^{\frac{1}{|S_{p'}|}} \mathbf{P}, \quad i = \{2, \dots, I\}. \tag{1.30}$$

To be compatible with our model, $\langle \widehat{\widehat{\mathbf{C}}}_i \rangle_{i=1}^{I}$ must be left-stochastic. Thus, pre-multiplying (1.30) by $\mathbf{1}^T = \left[\sum_{j=1}^{n} \mathbf{e}_j\right]^T$ we get

$$\mathbf{1}^T = \mathbf{1}^T \mathbf{D}_{S_{p'}:i\in S_{p'}}^{\frac{1}{|S_{p'}|}} \mathbf{P}, \tag{1.31}$$

or ,equivalently,

$$\mathbf{1}^T \mathbf{P}^T = \mathbf{1}^T = \mathbf{1}^T \mathbf{D}_{S_{p'}:i\in S_{p'}}^{\frac{1}{|S_{p'}|}}. \tag{1.32}$$

For (1.32) to hold true, it must be $\mathbf{D}_{S_{p'}:i\in S_{p'}}^{\frac{1}{|S_{p'}|}} = \mathbf{I}_n = \mathbf{D}_{S_{p'}:i\in S_{p'}}$ for all S_p not including alternative 1. Thus, assuming without loss of generality $1 \in S_1$ we have, by Lemma 1, $\mathbf{D}_{S_1}\mathbf{D}_{S_2}\mathbf{D}_{S_3} = \mathbf{I}_n$ which implies $\mathbf{D}_{S_1} = \mathbf{I}_n$ being $\mathbf{D}_{S_2} = \mathbf{D}_{S_3} = \mathbf{I}_n$.

Then, recalling that $\widehat{\mathbf{C}}_i = \mathbf{B}_i\mathbf{C}$, (1.30) can be rewritten as

$$\widehat{\widehat{\mathbf{C}}}_i = \widehat{\mathbf{C}}_i\mathbf{P} = \mathbf{B}_i\mathbf{C}\mathbf{P} = \mathbf{B}_i\mathbf{C}^*. \tag{1.33}$$

For $\widehat{\widehat{\mathbf{C}}}_i$ to be compatible with the model $\mathcal{M}(n; |k_1|, \ldots, |k_I|)$, $\mathbf{C}^*$ must be upper-triangular. Since $\mathbf{C}^*$ is obtained by permuting the columns of $\mathbf{C}$, which is also upper-triangular, the only possible permutation is the one brought about by $\mathbf{P} = \mathbf{I}_n$. Uniqueness of $\langle \mathbf{M}_i \rangle_{i=1}^I$ shown above implies that ”average” permutation matrices $\langle \mathbf{B}_i \rangle_{i=1}^I$ and the distribution of cognitive characteristics specified by $\boldsymbol{\pi}$ are also uniquely determined. That is to say that a panel dataset containing observed individual choices over time suffices in order to obtain the tensor of aggregate choice frequencies $\mathbf{p}$ and to uniquely infer from it some proxy of the distribution of tastes (namely, matrices $\langle \mathbf{B}_i \rangle_{i=1}^I$) along with the distribution of cognitive characteristics $\boldsymbol{\pi}$.

1.5 Estimation Strategy

In statistical terms, the model described thus far is a clear instance a latent variable model. Following Bartolucci et al. (2012), latent variable models can be defined as models *which rely on specific assumptions on the conditional distribution of the response variables, given one or more variables which are not directly observable (latent variables)*. Latent variable models typically relies on the assumption of *local independence*, according to which response variables are independent conditioning on the latent variables. Moreover, the following two components can usually be disentangled:

- *measurement model*: it specifies the conditional distribution of the response variables given the latent variables;
- *latent model*: it specifies the (unconditional) distribution of the latent variables.

With reference to a random unit drawn from the population of interest, let $K_1, \ldots, K_I$ denote a collection of categorical random variables each of them having support $K_i \in \{1, \ldots, n\}$, for $i = \{1, \ldots, I\}$. Let $K = (K_1 \ldots K_I)$ be the multinomial random variable with I trials and $\prod_{i=1}^{I} |K_i| = n^I$ mutually exclusive possible outcomes. Aggregate choice frequencies $\mathbf{p}$ defined in (1.14) can be held equivalent to the probability mass function $f_K(k) = f(K = k)$, with $k = (k_1 \ldots k_I)$ being a realization of K. Moreover, let Γ be a latent categorical random variable with support $\Gamma \in \{1, \ldots, n\}$. Then, the *latent model* is given by the *a priori* distribution of Γ which reads as $f_\Gamma(\gamma) = f(\Gamma = \gamma)$, with γ being a realization of Γ. The *measurement model*, instead, corresponds to the conditional distribution of K given Γ whose probability mass function is given by $f_{K|\Gamma}(k|\gamma) = f(K = k|\Gamma = \gamma)$. Finally, the manifest distribution of K is governed by the probability mass function

$$f_K(k) = \sum_{\gamma} f_{K|\Gamma}(k|\gamma) f_\Gamma(\gamma), \tag{1.34}$$

where the summation is taken over all possible realizations of Γ.

With reference to the multi-occasion consideration capacity model

$\mathcal{M}(n; |k_1|, \ldots, |k_I|)$, we have

$$f_\Gamma(\gamma) = \pi(\gamma), \quad \gamma = \{1, \ldots, n\}, \tag{1.35}$$

$$\begin{aligned} f_{K|\Gamma}(k|\gamma) &= \mathbb{P}_\gamma \\ &= \otimes_{i=1}^{I} \widehat{\mathbf{C}}_i \mathbf{e}_\gamma, \quad \gamma = \{1, \ldots, n\}, \end{aligned} \tag{1.36}$$

and

$$\begin{aligned} f_K(k) &= \mathbf{p} \\ &= \sum_{\gamma=1}^{n} \pi(\gamma)\mathbb{P}_\gamma. \end{aligned} \tag{1.37}$$

Note that second equality in (1.36) is a clear manifestation of the assumption of local independence. Indeed, the γ-th column of the generic matrix $\widehat{\mathbf{C}}_i$ is equivalent to the conditional distribution $f_{K_i|\Gamma}(k_i|\gamma) = f(K_i = k_i|\Gamma = \gamma)$ of the generic categorical K_i variable defined above. Thus, the joint distribution of $(K_1 \ldots K_I)|\Gamma = K|\Gamma$ obtains as the product of (conditional) marginals, i.e.

$$f_{K|\Gamma}(k|\gamma) = \prod_{i=1}^{I} f_{K_i|\Gamma}(k_i|\gamma). \tag{1.38}$$

1.5.1 Expectation-maximization (EM) algorithm

Originally derived by Baum et al. (1970) in the context of hidden Markov models and later extended and formalized by Dempster et al. (1977), EM algorithm[25] is definitely the most used tool for estimating latent variable models. It consists of a derivative-free iterative method which is able to provide maximum-likelihood or maximum a posteriori estimates of parameters in statistical models depending on unobserved latent variables.

Drawing a sample of S independent subjects, let $k^s = (k_1^s \dots k_I^s)$[26] denote the observed response configuration of a subject $s = \{1, \dots, S\}$ who is asked to choose one of the alternatives of a menu $X = \{1, \dots, n\}$ on a number of occasions I.

In our non-parametric framework where no individual covariates are accounted

[25] A complete overview of this method can be found in Watanabe and Yamaguchi (2003) and in McLachlan and Krishnan (2008).

[26] k^s can be thought as a column-vector of dimension $|X|^I$ with a 1 in the position implied by the (individual) joint response $(k_1^s \dots k_I^s)$ and 0 elsewhere. $|X|$ is the number of available alternatives and I the number of occasions. Let, for instance, $I = 3$ and $|X| = 3$. Then, if s chooses alternative 2 on occasion 1, alternative 3 on occasion 2 and alternative 3 on occasion 3, we have

$$k_1^s = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \quad k_2^s = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, \quad k_3^s = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

and k^s is uniquely determined by $k^s = k_1^s \otimes k_2^s \otimes k_3^s$.

for, (1.37) gives rise to the following model log-likelihood

$$\begin{aligned}\ell(\boldsymbol{\theta}) &= \sum_{s=1}^{S} \log \mathbf{p}_s \\ &= \sum_{s=1}^{S} \log f_K(k^s; \boldsymbol{\theta}),\end{aligned} \tag{1.39}$$

with $\boldsymbol{\theta}$ belonging to the parameter space $\boldsymbol{\Theta}$ defined above. Let $I(\cdot)$ be the indicator function that equals 1 if the argument is true and 0 otherwise and define

$$S_k = \sum_{s=1}^{S} I(k^s = k). \tag{1.40}$$

S_k is the frequency with which response configuration k is observed in the sample. Then, (1.39) can be more conveniently expressed as

$$\ell(\boldsymbol{\theta}) = \sum_{k} S_k \log f_K(k; \boldsymbol{\theta}). \tag{1.41}$$

The advantage of using (1.41) instead of (1.39) rests upon the fact, as the sample size grows larger, ever more subjects are likely to exhibit the very same response configuration. The function $\ell(\boldsymbol{\theta})$ is usually called the *incomplete-data log-likelihood* since it totally neglects the missing data problem from which our latent variable model suffers. For a maximization routine to be immediately applicable, we would need to know not only the response configuration k^s but also the realization γ^s of the latent variable Γ for each subject s in the sample. Observing the pairs

(k^s, γ^s) for $s = \{1, \ldots, S\}$ would allow us to construct the so-called *complete-data log-likelihood* which is given by

$$\ell^*(\boldsymbol{\theta}) = \sum_{s=1}^{S} \log f_{K,\Gamma}(k^s, \gamma^s; \boldsymbol{\theta}). \tag{1.42}$$

Being $f_{K,\Gamma}(k, \gamma) = f_{K|\Gamma}(k|\gamma) f_{\Gamma}(\gamma)$, (1.42) rewrites as

$$\begin{aligned} \ell^*(\boldsymbol{\theta}) &= \sum_{s=1}^{S} \log \left(f_{K|\Gamma}(k^s|\gamma^s; \boldsymbol{\theta}) f_{\Gamma}(\gamma^s; \boldsymbol{\theta}) \right) \\ &= \sum_{s=1}^{S} \log f_{K|\Gamma}(k^s|\gamma^s; \boldsymbol{\theta}) + \sum_{s=1}^{S} \log f_{\Gamma}(\gamma^s; \boldsymbol{\theta}). \end{aligned} \tag{1.43}$$

In the same spirit of simplification as the one deployed for obtaining (1.41), let us define

$$a_{k\gamma} = \sum_{s=1}^{S} I(\gamma^s = \gamma, k^s = k) \quad \text{and} \quad b_{\gamma} = \sum_{s=1}^{S} I(\gamma^s = \gamma). \tag{1.44}$$

The term b_γ is the number of subjects in the sample with latent variable equal to γ while $a_{k\gamma}$ is the number of those having both latent variable equal to γ and response configuration equal to k. Using (1.44), the *complete-data log-likelihood* becomes

$$\ell^*(\boldsymbol{\theta}) = \sum_{\gamma} \sum_{k} a_{k\gamma} \log f_{K|\Gamma}(k|\gamma; \boldsymbol{\theta}) + \sum_{\gamma} b_{\gamma} \log f_{\Gamma}(\gamma; \boldsymbol{\theta}), \tag{1.45}$$

where $\sum_k$ and $\sum_\gamma$ are taken over all possible realizations of K and Γ, respectively. From the perspective of the EM algorithm, maximization of $\ell(\boldsymbol{\theta})$ is achieved by iteratively performing the following two steps until some convergence criteria[27] are met:

- **Expectation (E) step.** It amounts to compute the expectation of $\ell^*(\boldsymbol{\theta})$ with respect to the conditional distribution of Γ given K and the current values of the parameters $\hat{\boldsymbol{\theta}}^t$, namely

$$\begin{aligned} E_{\Gamma|K,\hat{\boldsymbol{\theta}}^t}\left[\ell^*(\boldsymbol{\theta})\right] &= \sum_{\gamma}\sum_{k} E_{\Gamma|K,\hat{\boldsymbol{\theta}}^t}\left[a_{k\gamma}\right] \log f_{K|\Gamma}(k|\gamma;\boldsymbol{\theta}) + \sum_{\gamma} E_{\Gamma|K,\hat{\boldsymbol{\theta}}^t}\left[b_{\gamma}\right] \log f_{\Gamma}(\gamma;\boldsymbol{\theta}) \\ &= \sum_{\gamma}\sum_{k} \hat{a}_{k\gamma} \log f_{K|\Gamma}(k|\gamma;\boldsymbol{\theta}) + \sum_{\gamma} \hat{b}_{\gamma} \log f_{\Gamma}(\gamma;\boldsymbol{\theta}), \end{aligned} \tag{1.46}$$

with $\hat{a}_{k\gamma}$ and $\hat{b}_{\gamma}$ given by

$$\begin{aligned} \hat{a}_{k\gamma} &= \sum_{s=1}^{S} I(k^s = k) f_{\Gamma|K,\hat{\boldsymbol{\theta}}^t}(\gamma^s|k,\hat{\boldsymbol{\theta}}^t) \\ &\text{and} \\ \hat{b}_{\gamma} &= \sum_{s=1}^{S} f_{\Gamma|K,\hat{\boldsymbol{\theta}}^t}(\gamma^s|k,\hat{\boldsymbol{\theta}}^t), \end{aligned} \tag{1.47}$$

respectively. The function $f_{\Gamma|K,\hat{\boldsymbol{\theta}}^t}(\gamma|k,\hat{\boldsymbol{\theta}}^t)$ specifies the *a posteriori* distribution of the latent variable which is obtained conditioning on a certain response configuration k and a set of parameters $\hat{\boldsymbol{\theta}}^t$. At each iteration t,

[27] Letting t be the number of iteration and $\hat{\boldsymbol{\theta}}^t = \langle \hat{\theta}_h^t \rangle_{h=1}^L$ the vector of parameter estimates obtained at iteration t, convergence is typically assumed to be achieved if $\ell(\hat{\boldsymbol{\theta}}^{t+1}) - \ell(\hat{\boldsymbol{\theta}}^t) \leq 1e^{-6}$ and $\max_h |\hat{\theta}_h^{t+1} - \hat{\theta}_h^t| \leq 1e^{-6}$.

Bayes' theorem allows us to recover this probability mass function as

$$f_{\Gamma|K,\hat{\boldsymbol{\theta}}^t}(\gamma|k,\hat{\boldsymbol{\theta}}^t) = \frac{f_{K|\Gamma,\hat{\boldsymbol{\theta}}^t}(k|\gamma,\hat{\boldsymbol{\theta}}^t) f_{\Gamma|\hat{\boldsymbol{\theta}}^t}(\gamma|\hat{\boldsymbol{\theta}}^t)}{f_{K|\hat{\boldsymbol{\theta}}^t}(k|\hat{\boldsymbol{\theta}}^t)}, \tag{1.48}$$

where $f_{K|\hat{\boldsymbol{\theta}}^t}(k|\hat{\boldsymbol{\theta}}^t) = \sum_\gamma f_{K|\Gamma,\hat{\boldsymbol{\theta}}^t}(k|\gamma,\hat{\boldsymbol{\theta}}^t) f_{\Gamma|\hat{\boldsymbol{\theta}}^t}(\gamma|\hat{\boldsymbol{\theta}}^t)$.

For each iteration-specific parameters $\hat{\boldsymbol{\theta}}^t$, (1.48) provides an allocation rule according to which each subject is assigned to a certain latent variable configuration.

In the framework of the consideration capacity model $\mathcal{M}(n; |k_1|, \ldots, |k_I|)$, let $\hat{\boldsymbol{\theta}}^t = (\boldsymbol{\pi}^t, \text{vec}(\langle \widehat{\mathbf{C}}_i^t \rangle_{i=1}^I)$ be the vector of parameters at iteration t and write $\hat{c}_i^t(r,c)$ for the generic (r,c) element of matrix $\widehat{\mathbf{C}}_i^t$. Then, applying (1.48), the posterior probability of belonging to a certain latent class γ given response configuration $k = (k_1 \ldots k_I)$ and current parameters $\hat{\boldsymbol{\theta}}^t$ reads as

$$\pi(\gamma|k,\hat{\boldsymbol{\theta}}^t) = \frac{(\prod_{i=1}^I \hat{c}_i^t(k_i,\gamma))\pi(\gamma|\hat{\boldsymbol{\theta}}^t)}{\sum_{\gamma=1}^n (\prod_{i=1}^I \hat{c}_i^t(k_i,\gamma))\pi(\gamma|\hat{\boldsymbol{\theta}}^t)}. \tag{1.49}$$

- **Maximization (M) step.** In this step, parameters are updated by maximizing the expectation computed in (1.46) with respect to the "true" parameters $\boldsymbol{\theta} = (\boldsymbol{\pi}, \text{vec}(\langle \widehat{\mathbf{C}}_i \rangle_{i=1}^I)$. Thus, next-iteration parameters $\hat{\boldsymbol{\theta}}^{t+1}$ are obtained as

$$\hat{\boldsymbol{\theta}}^{t+1} = \arg\max_{\boldsymbol{\theta}} E_{\Gamma|K,\hat{\boldsymbol{\theta}}^t}\left[\ell^*(\boldsymbol{\theta})\right]$$
$$\text{s.t.}$$
$$\begin{cases} \mathbf{1}^T \mathbf{B}_i^{t+1} = \mathbf{1}^T, \\ \mathbf{B}_i^{t+1}\mathbf{1} = \mathbf{1}, \quad i = \{1, \ldots, I\} \\ \mathbf{B}_i^{t+1} \geq \mathbf{0}_n, \end{cases} \tag{1.50}$$

where $\mathbf{B}_i^{t+1} = \widehat{\mathbf{C}}_i^{t+1}\mathbf{C}^{-1}$ and $\mathbf{0}_n$ is the $n \times n$ zero matrix. Constraints in (1.50) effectively impose the requirement of bistochasticity on the matrices $\mathbf{B}_i^{t+1}$ and guarantee that the estimated *measurement model* belongs to the same model space as the one of $\mathcal{M}\left(n; |k_1|, \ldots, |k_I|\right)$ at each and every iteration. Note that the first constraint in (1.50) is redundant being always satisfied. Indeed, we have

$$\mathbf{1}^T \mathbf{B}_i^{t+1} = \mathbf{1}^T \widehat{\mathbf{C}}_i^{t+1}\mathbf{C}^{-1} = \mathbf{1}^T\mathbf{C}^{-1} = \mathbf{1}^T, \tag{1.51}$$

where the second equality is due to $\widehat{\mathbf{C}}_i^{t+1}$ being left-stochastic while last equality is a consequence of the fact that the inverse of a left-stochastic matrix (namely, $\mathbf{C}^{-1}$) has again the property that the sum of the columns add up to 1. Substituting for the expression of $\mathbf{B}_i^{t+1}$, the second constraint reads as

$$\mathbf{B}_i^{t+1}\mathbf{1} = \widehat{\mathbf{C}}_i^{t+1}\mathbf{C}^{-1}\mathbf{1} = \widehat{\mathbf{C}}_i^{t+1}\begin{bmatrix} n \\ 0 \\ \vdots \\ 0 \end{bmatrix}. \tag{1.52}$$

Last equality in (1.52) arises since, by construction, all rows of $\mathbf{C}^{-1}$ sum up to zero except for the first one which sums up to n. Hence, (1.52) shows that the second constraint is in all respect equivalent to imposing all the terms on the first column of matrices $\widehat{\mathbf{C}}_i^{t+1}$ equal to $\frac{1}{n}$, that is to say that all alternatives in the menu are chosen with the very same probabilities (namely, $\frac{1}{n}$) by type-1 individuals. Finally, vectorizing $\mathbf{B}_i^{t+1}$ gives

$$\text{vec}(\mathbf{B}_i^{t+1}) = \text{vec}(\widehat{\mathbf{C}}_i^{t+1}\mathbf{C}^{-1}) = ((\mathbf{C}^{-1})^\top \otimes \mathbf{I}_n)\text{vec}(\widehat{\mathbf{C}}_i^{t+1}). \tag{1.53}$$

(1.53) allows us to more conveniently formulate last constraint in (1.50) as a system of linear inequalities in the parameters to be estimated, namely

$$((\mathbf{C}^{-1})^\top \otimes \mathbf{I}_n)\text{vec}(\widehat{\mathbf{C}}_i^{t+1}) \geq \mathbf{0}, \tag{1.54}$$

with $\mathbf{0} = \text{vec}(\mathbf{0}_n)$ being the zero column vector of length $n \times n$.

1.5.2 Simulations

The simulation reported here assumes individuals choosing among 5 mutually exclusive alternatives (i.e. $X = \{1, 2, 3, 4, 5\}$) on 3 occasions. After fixing the "true" distribution of cognitive characteristics $\boldsymbol{\pi}$, we generate 100 samples of $S = 10,000$ subjects and draw the cognitive type γ^s of each subject s independently from $\boldsymbol{\pi}$. For each occasion i, we generate a random vector $\boldsymbol{\tau}_i$ of dimension 5! specifying the probability of each of the possible $\langle \varphi_h \rangle_{h=1}^{5!}$ rankings. Then, for each sample, we make S independent draws from each distribution $\boldsymbol{\tau}_i$ so as to assign every individual in the sample with a specific preference ordering at each choice occasion. Once attention and preference have been assigned, choices are retrieved by independently drawing from the following occasion-specific vector of individual choice probabilities

$$\mathbf{p}_i^s = \mathbf{P}(h_i^s)\mathbf{C}\mathbf{e}_{\gamma^s}, \tag{1.55}$$

where h_i^s identifies the preference ordering of subject s at occasion i drawn from $\boldsymbol{\tau}_i$. Drawing from $\mathbf{p}_i^s$, we obtain individual choice at each occasion (namely, k_i^s)[28] and the response variable, that is the tensor of (joint) aggregate choices, is obtained as

$$\boldsymbol{Y} = \sum_{s=1}^{1e4} \otimes_{i=1}^{3} k_i^s. \tag{1.56}$$

[28]In these simulations, k_i^s is a column-vector of length $|X| = 5$ with a 1 in the position implied by the individual choice at occasion i and 0 elsewhere.

Table 1.1: Simulations: γ-types distribution

	$\pi(1)$	$\pi(2)$	$\pi(3)$	$\pi(4)$	$\pi(5)$
true value	.0415	.2033	.4067	.2228	.1257
estimated value	.0399	.2035	.3809	.2248	.1509
bias	-.0016	.0002	-.0258	.0019	.0252
RMSE	.0042	.0172	.0491	.0295	.0463

Table 1.1 reports the EM average estimates of the parameters governing the distribution of cognitive characteristics along with their associated bias and root mean squared error. Table 1.2, instead, compares the true matrices of type-conditional probabilities with the average of the ones obtained through the EM algorithm. For both the predicted *latent* and *measurement* models, estimates are substantially close to the true values. Note that, in generating the sample, we achieve the highest possible degree of variability in our data. This is so since individual choices are not obtained drawing directly from type-conditional probabilities but are constructed from (1.55) which depends on the specific preference ordering $\varphi(h_i^s)$ attached to individual s at occasion i. In applications involving real data, instead, it is likely the case that the true distribution of preferences $\boldsymbol{\tau}_i$ is rather sparse, meaning that many of the possible preference orderings never appear in the population.

Table 1.2: Simulations: type-conditional choice probabilities

true (i=1)

$$\begin{bmatrix} .200 & .191 & .102 & .014 & .017 \\ .200 & .022 & .010 & .011 & .012 \\ .200 & .379 & .548 & .702 & .843 \\ .200 & .112 & .037 & .048 & .060 \\ .200 & .296 & .303 & .226 & .068 \end{bmatrix}$$

estimated (i=1)

$$\begin{bmatrix} .200 & .194 & .105 & .018 & .013 \\ .200 & .024 & .011 & .010 & .011 \\ .200 & .374 & .537 & .694 & .845 \\ .200 & .112 & .040 & .046 & .055 \\ .200 & .298 & .307 & .232 & .077 \end{bmatrix}$$

true (i=2)

$$\begin{bmatrix} .200 & .273 & .260 & .173 & .000 \\ .200 & .204 & .118 & .021 & .001 \\ .200 & .048 & .065 & .083 & .104 \\ .200 & .116 & .020 & .006 & .000 \\ .200 & .358 & .537 & .716 & .896 \end{bmatrix}$$

estimated (i=2)

$$\begin{bmatrix} .200 & .272 & .266 & .180 & .016 \\ .200 & .200 & .118 & .030 & .006 \\ .200 & .051 & .064 & .082 & .099 \\ .200 & .116 & .021 & .005 & .003 \\ .200 & .361 & .531 & .703 & .876 \end{bmatrix}$$

true (i=3)

$$\begin{bmatrix} .200 & .101 & .002 & .001 & .000 \\ .200 & .002 & .001 & .001 & .000 \\ .200 & .301 & .302 & .204 & .007 \\ .200 & .398 & .596 & .795 & .993 \\ .200 & .199 & .099 & .000 & .000 \end{bmatrix}$$

estimated (i=3)

$$\begin{bmatrix} .200 & .101 & .004 & .001 & .001 \\ .200 & .003 & .001 & .001 & .000 \\ .200 & .301 & .305 & .212 & .024 \\ .200 & .396 & .590 & .782 & .972 \\ .200 & .199 & .102 & .005 & .003 \end{bmatrix}$$

1.5.3 An application to market data: ERIM dataset on sugar

The ERIM dataset[29] contains data collected by the now-defunct ERIM division of A.C. Nielsen on panels of households in two midsized cities in the U.S.. Information is available on the purchases of households in a number of product categories along with household demographic information. In this application we focus on sugar purchases. Among available categories, sugar is, in our opinion, the one with the highest degree of substitutability between brands and, as such, qualifies as an ideal candidate for the extrapolation of households' attention capacity in our fully non-parametric framework. ERIM data on sugar collects daily scanner data from 42 supermarkets for the period between the 5-th week of 1985 and the 23-rd week of 1987. In this application we select the supermarket with the highest number of transactions and consider only the 8 most chosen sugar brands which represent more than 75% of the market. Next we identify 3 different shop occasions by aggregating successive weeks over which the very same prices and display activity prevail for the 8 brands under consideration. Then, we further refine our dataset by only considering households who are observed choosing on all the 3 occasions defined above. Doing so, we are able to identify 225 households and construct the tensor of (joint) aggregate choices which has $8^3 = 512$ entries. As expected, this tensor turns to be quite sparse having 416 zero entries. The (relative) market shares for the 8 products in our 3-occasion sub-samples are reported in Table 1.3.

Table 1.4 shows that the distribution of γ-types is essentially tri-modal with

[29] Full documentation of this dataset is available at: https://www.chicagobooth.edu/research/kilts/datasets/erim

Table 1.3: ERIM data: Market shares

	1	2	3	4	5	6	7	8
i = 1	.089	.084	.089	.044	.449	.004	.107	.133
i = 2	.102	.098	.107	.058	.404	.040	.080	.111
i = 3	.093	.053	.084	.044	.471	.027	.089	.138

around 16% of the individuals having a low attention capacity ($\gamma = 2$), around 38% having a moderate attention capacity ($\gamma = 5$) and around 46% being full attentive ($\gamma \geq 8$). Standard errors in Table 1.4 are computed by using the observed information matrix, denoted by $\mathbf{I}(\boldsymbol{\theta})$. Following Bartolucci et al. (2012), $\mathbf{I}(\boldsymbol{\theta})$ is obtained as minus the numerical derivative of the score vector $\mathbf{s}(\boldsymbol{\theta})$. The score vector $\mathbf{s}(\boldsymbol{\theta})$, in turn, is obtained as the numerical first derivative of the complete-data log-likelihood $\ell^*(\boldsymbol{\theta})$, evaluated at $\boldsymbol{\theta}$ equal to the value of the estimated parameter vector $\hat{\boldsymbol{\theta}}$. Finally, standard errors are computed as the square root of the corresponding diagonal element of $\mathbf{I}(\hat{\boldsymbol{\theta}})^{-1}$.

Table 1.5, instead, reports estimated type-conditional choice probabilities. Note that, as we move rightward over the columns of the matrices in Table 1.5, predicted probability mass functions tend to get more concentrated around the alternative that has the highest market share in all the 3 occasions (namely, alternative 5). Clearly, this is a consequence of the fact that, as consideration capacity increases, individuals become more likely to choose what they actually prefer.

Table 1.4: ERIM data: γ-types distribution

	$\pi(1)$	$\pi(2)$	$\pi(3)$	$\pi(4)$	$\pi(5)$	$\pi(6)$	$\pi(7)$	$\pi(8)$
etimated value	.000	.161	.000	.000	.381	.000	.000	.458
s.e.	.005	.017	.005	.007	.057	.005	.005	.040

1.6 Conclusion

Relaxing the strong assumption that subjects take into consideration all the options available to them before choosing is a common feature of discrete choice models with limited attention. In such models, identification has typically been achieved either by relying on additional information on what alternatives are actually considered by individuals or by assuming that some variables impact attention or utility but not both. One exception in this respect is represented by Abaluck and Adams (2017) who constructively identify consideration set probabilities from the asymmetries in the matrix of cross-derivatives of choice probabilities with respect to the characteristics of rival goods. As argued by the same authors, however, their model is likely to get particularly data ”hungry” when estimated in a non-parametric setup since it requires to observe a large number of individuals and exogenous variation in the characteristics of rival goods (typically, the price).

The model in Dardanoni et al. (2019) complements the literature to date given that it links directly cognitive heterogeneity to observed (aggregate) choices. To our knowledge, their model is the only one in which the distribution of cognitive characteristics can be determined *per se* without requiring variation in alternative-

Table 1.5: ERIM data: type-conditional choice probabilities

$$\textbf{estimated}\ (\mathrm{i{=}1})$$
$$\begin{bmatrix}
.125 & .160 & .186 & .197 & .184 & .146 & .085 & .000 \\
.125 & .143 & .107 & .057 & .018 & .000 & .000 & .000 \\
.125 & .005 & .007 & .009 & .012 & .014 & .016 & .019 \\
.125 & .090 & .039 & .013 & .009 & .010 & .012 & .014 \\
.125 & .235 & .334 & .426 & .514 & .602 & .694 & .793 \\
.125 & .093 & .042 & .014 & .008 & .010 & .012 & .013 \\
.125 & .190 & .207 & .189 & .147 & .093 & .040 & .000 \\
.125 & .084 & .077 & .093 & .109 & .124 & .141 & .161
\end{bmatrix}$$

$$\textbf{estimated}\ (\mathrm{i{=}2})$$
$$\begin{bmatrix}
.125 & .170 & .213 & .229 & .217 & .177 & .110 & .015 \\
.125 & .110 & .065 & .037 & .032 & .038 & .045 & .051 \\
.125 & .170 & .176 & .171 & .174 & .193 & .225 & .257 \\
.125 & .104 & .075 & .066 & .069. & .078 & .090 & .102 \\
.125 & .128 & .193 & .257 & .321 & .385 & .449 & .514 \\
.125 & .076 & .052 & .045 & .032 & .020 & .013 & .015 \\
.125 & .134 & .146 & .133 & .103 & .066 & .028 & .000 \\
.125 & .108 & .080 & .063 & .051 & .043 & .040 & .046
\end{bmatrix}$$

$$\textbf{estimated}\ (\mathrm{i{=}3})$$
$$\begin{bmatrix}
.125 & .102 & .128 & .138 & .131 & .107 & .067 & .010 \\
.125 & .136 & .115 & .101 & .105 & .123 & .143 & .164 \\
.125 & .140 & .104 & .059 & .026 & .013 & .016 & .018 \\
.125 & .057 & .023 & .022 & .025 & .030 & .035 & .040 \\
.125 & .128 & .192 & .256 & .319 & .383 & .447 & .511 \\
.125 & .074 & .023 & .007 & .007 & .008 & .010 & .011 \\
.125 & .198 & .227 & .221 & .189 & .139 & .080 & .021 \\
.125 & .165 & .187 & .198 & .197 & .196 & .202 & .225
\end{bmatrix}$$

and/or individual-specific characteristics.

On a theoretical ground, this paper extends the identification of the model in Dardanoni et al. (2019) to the generic case where subjects are observed choosing in more than three occasions. From an empirical standpoint, it develops an estimation strategy that rests on an Expectation-Maximization algorithm where suitable (linear) constraints are imposed. This methodology, we believe, is flexible enough to be adapted to a parametric context which explicitly models the interaction between cognitive and taste heterogeneity and allows to isolate the (possibly different) effects that some given regressors have on utility and attention. In such a context, one could easily extend standard normative analysis and evaluate, for instance, the welfare gain that individuals might experience if they were full attentive.

The implicit assumption that all consideration sets with the same cardinality are equally likely is the one that future research should try to relax in order to generalize the model in Dardanoni et al. (2019) and incorporate saliency effects.

2

Asset Allocation in Defined Contribution Plans with Limited Attention

2.1 Introduction

Defined-contribution plans (henceforth, DCPs), and in particular the 401(k) scheme introduced in US in the late 1970s, brought about a significant alteration in the relationship between private employers and their workers as far as retirement savings are concerned. Before the approval of the Employee Retirement Income Security Act of 1974 (ERISA), defined-benefit plans (henceforth, DBPs) were commonplace among U.S. private businesses' employees. Under a DBP, employers typically provide workers with financial security in retirement via a pension. Thus, the employer is responsible for all investment decisions regarding financial assets allocated for retirement income. In addition, DBPs usually make employers responsible for certain guarantees regarding minimum income distributions through the use of annuities (e.g., guaranteed lifetime income representing a certain percentage of the last years salary of employment). Under such a scheme, employees worried neither about how their retirement savings were invested nor if the returns generated by those investments were sufficient to provide for adequate retirement income.

The adoption of the 401(k) and other defined-contribution schemes fundamentally modified the employer-employee relationship, given that investment management decisions are transferred from companies to individual workers, who furthermore find themselves without the guarantees associated with a DBP pension. The employees have become not only responsible for their own investment account allocations, but also for investment performance and for managing the disbursement of income upon retirement.

The trend of switching from DBPs to DCPs over the past three decades has been remarkable. In 2013 only 2% of private-sector employees enjoyed a company directed defined-benefit pension plan for their retirement security, while in

1979 that number amounted to around 62%. According to the US Department of Labor, by 2009 more than two-thirds of all employees relied entirely on a DCP.[1] Nowadays, the majority of private-sector workers no longer relies on their employer for making decisions concerning their retirement income security; rather, employees are left to voluntarily contribute to their own retirement savings and make their own decisions as to how those contributions are invested. The transition outlined above has raised relevant questions as to what drives workers' decision process concerning their saving contributions which will determine income upon retirement.

Analyzing the structure of 401(k) plans, previous literature has extensively documented the importance of defaults and their non-negligible impact on saving behavior. Choi et al. (2002) assess the impact of several 401(k) plan features, such as automatic enrollment, employer matching provisions and investment options on saving behavior documenting a strong employees tendency to follow a path of least resistance. That is to say that workers tend to stick to the defaults of their 401(k) plan. By exploiting the switch from automatic non-enrollment to automatic enrollment in 401(k) plan made by some companies, the authors find evidence that automatic enrollment substantially increases the 401(k) participation rate of newly hired workers and this effect is the largest for those individuals least likely to participate in the first place: namely younger employees, lower-paid employees, Blacks and Hispanics. With regards to the contribution rate and asset allocation, Choi et al. (2002) also show that employees tend to passively accept the

[1] https://www.dol.gov/sites/default/files/ebsa/researchers/statistics/retirement-bulletins/historicaltables.pdf

defaults offered by their plan's sponsor. Thus, many workers end up contributing either an excessively high or low fraction of their income and investing in assets which may well be not adequate to their risk attitude. Additionally, the event studies conducted by these authors give some insights on the relevance of one more common feature of 401(k) plans, that is employer match. Employer match implies that for each dollar contributed by the employee to the plan, the employer contributes a matching amount up to a certain threshold (e.g., 50 percent of the employee contribution up to 6 percent of compensation). Partially contradicting previous literature (e.g., Papke and Poterba (1995)), Choi et al. (2002) find that varying the match threshold has no significant effect on 401(k) participation but it does impact contribution rates. Indeed, match threshold exercises an anchoring effect on the decision of which contribution rate to select with many participants clustering at the threshold.

As far as risk diversification is concerned, Thaler and Benartzi (2001) suggest that naive strategies for diversifying across investment options cause many investors to allocate part of their contributions to employer stock simply because it is available in the 401(k) menu. Benartzi (2001), VanDerhei et al. (2010) and Brown et al. (2006) find that discretionary contributions to employer stock are higher in firms where the employer directs matching contributions into employer stock than in firms where employer stock is simply available as another investment option. Investing too much in employer stocks does not represent a sensible strategy given that the value of employer stocks are likely to be positively correlated with employees' labour income. All in all, a growing body of literature finds that

401(k) savings outcomes are strongly affected by the features of the plan, even when those features do not explicitly restrict employee choices.

In this chapter we depart from previous literature since we use a rich cross-sectional dataset and analyze investing decisions under the lenses of a multinomial model which accounts for limited attention. To our knowledge, Sunden and Surette (1998) is the only paper using discrete choice modeling techniques in the context of allocation choices within DCPs. Sunden and Surette (1998) are primarily concerned with the evaluation of gender differences in investment decisions. Here, instead, we link the decision process to several individual and job-specific characteristics that have been found to shape asset allocation by previous literature. Taking limited attention into account essentially amounts to assume that individuals do not necessarily consider all available options when making their choices. This lack of attention may be the consequence, for instance, of search cost or of bounded rationality. Thus, our model specifies a probabilistic consideration set formation process for each of the possible subsets of options.[2]

The rest of this chapter is organized as follows. Section 2.2 describes the survey used to construct the cross-sectional dataset to which our model is applied. Section 2.3 presents the model whose (local) identifiability is shown in Appendix A. Section 2.4 discusses the empirical results and highlights the relevance of (in)attention by comparing the estimates of our model with the ones of a standard multinomial logit model. Section 2.5 concludes.

[2]This approach has been extensively applied in the marketing literature (see, e.g., Shocker et al. (1991)) and has also gained popularity in both theoretical (see, e.g., Manzini and Mariotti (2014)) and applied economics (see, e.g., Goeree (2008) and Abaluck and Adams (2017)).

2.2 Data

The empirical analysis performed here relies on the Survey of Income and Program Participation (henceforth, SIPP) conducted by the United States Census Bureau. SIPP collects data related to various types of income, labor force participation, social program participation and eligibility, and general demographic characteristics to measure the effectiveness of existing federal, state, and local programs. The survey is articulated in topical modules run on a yearly basis and containing specific questions on socio-economic issues. The source of data employed here is the Retirement and Pension Plan Coverage module run in 2012 which contains individual-level information on pension plan contribution, asset allocation and demographic characteristics.

To make our data suitable for the implementation of a multinomial logit model, we restrict our attention to those individuals who decide to allocate all their contributions in just one asset category. Perhaps surprisingly, this sample (5729 units) represents the very large majority of individuals who report to be completely free to determine how to allocate the money invested in their DCPs (8671 units).

The dependent variable of interest is represented by the asset category chosen while the explanatory variables refer to demographic and job-specific characteristics. In order to establish some comparisons with the results accumulated in previous literature, we adopt as explanatory variables[3] age, gender (= 1 if male), the percentage of salary contributed by the worker, a dummy variable for university education and a further dummy coding the presence of employer matching

[3] The two continuous variables, namely age and percentage contribution, are standardized.

contribution.

2.3 Model

In our setting individuals face a decision among six mutually exclusive asset classes. The choice set is considered to be unordered since our data provides no information on alternative-specific characteristics. Having information on the risk-return profile of the asset classes, for instance, would have allowed us not only to establish a precise ranking in terms of asset riskiness but also to incorporate our model in a mixed logit framework where both individual and alternative-specific features affect a subject's decision. Unordered-choice models can be motivated by a random utility model where the i-th individual who faces a menu $X = \{0, 1, .., J\}$ of alternatives derives from choice j utility

$$U_{ij} = \boldsymbol{\beta}_j' \mathbf{x}_i + \epsilon_{ij}. \tag{2.1}$$

$\mathbf{x}_i$ is a vector of individual-specific characteristics while ϵ_{ij} is an i.i.d. error term following a type 1 extreme value distribution. If j is chosen by individual i, then we assume that U_{ij} is the maximum among the J utilities. Hence, the statistical model is driven by the probability of choosing alternative j, which is

$$Prob(U_{ij} > U_{ik}) \quad \text{for all other } k \neq j \tag{2.2}$$

In our setting, the menu of available alternatives and individual regressors are respectively given by:

1. *Investment*: 0 = money market funds and other liquid investments, 1 = government securities, 2 = diversified stock and bond funds, 3 = corporate bonds and bond funds, 4 = stock funds, 5 = employer company stock.

2. $\mathbf{x}_i$: constant, age, gender, college education, employer matching contribution, employee % contribution.

Given that errors are i.i.d. and follow a type 1 extreme value distribution, allocation probabilities have the following multinomial logit specification:

$$p_i^{MNL}(j \mid X) = \frac{e^{\boldsymbol{\beta}_j' \mathbf{x}_i}}{\sum_{k \in X} e^{\boldsymbol{\beta}_k' \mathbf{x}_i}}, \quad j \in X \tag{2.3}$$

Adopting the convenient normalization $\boldsymbol{\beta}_0 = \mathbf{0}$, probabilities can be rewritten as

$$p_i^{MNL}(j \mid X) = \frac{e^{\boldsymbol{\beta}_j' \mathbf{x}_i}}{1 + \sum_{k \in X \mid k \neq 0} e^{\boldsymbol{\beta}_k' \mathbf{x}_i}}, \quad j \in X \quad \textit{and} \quad \boldsymbol{\beta}_0 = \mathbf{0} \tag{2.4}$$

The multinomial logit model implicitly assumes that individuals consider all available options when choosing. Consideration set models generalize standard discrete choice models by relaxing the assumption of individuals considering all goods. These models specify a probability that each subset of options is considered (Manski (1977)) so that the choice problem typically assumes the following probabilistic specification:

$$P_i(j) = \sum_{C \in X(j)} P_i(j|C) P_i(C|X),$$

where :

- $P_i(j)$ is the probability of individual i choosing alternative j;
- $X(j)$ is the set of all possible choice sets containing alternative j;
- $P_i(j|C)$ is the probability of individual i choosing j given that her choice set is C;
- $P_i(C|X)$ is the probability of C being the choice set of individual i.

Consideration sets might arise, for instance, due to bounded rationality (e.g., Treisman and Gelade (1980)), search costs (e.g., Caplin et al. (2018)), or because consumers face (unobserved) constraints on what options can be chosen (e.g., Gaynor et al. (2016)). As an extension of the standard multinomial logit described in (2.4), we develop a limited attention model which includes a probabilistic choice set generation in the first stage followed by the choice of financial instrument from a given choice set. This extended model formulation assumes that workers can choose any of the available asset categories though not all of them may be considered by each individual. Under such a model individual utility writes as

$$U_{ij} = \boldsymbol{\beta}_j' \mathbf{x}_i + K_{ij} + \epsilon_{ij}, \tag{2.5}$$

where

$$K_{ij} = \begin{cases} 0, & \text{with probability} \quad \Lambda_{ij} \\ K, & \text{with probability} \quad (1 - \Lambda_{ij}). \end{cases}$$

K represents a sufficiently large negative shock affecting utility. Λ_{ij} represents the probability that alternative j is considered by individual i and is assumed to be standard logistically distributed:

$$\Lambda_{ij} = \frac{e^{\boldsymbol{\gamma}_j' \mathbf{w}_i}}{1 + e^{\boldsymbol{\gamma}_j' \mathbf{w}_i}}. \tag{2.6}$$

where $\mathbf{w}_i$ is a column vector of observed attributes of individual i (including a constant) and $\boldsymbol{\gamma}_j$ is a corresponding column vector of coefficients to be estimated (these coefficients provide the impact of attributes on the consideration probability of alternative j). In the empirical analysis performed here $\mathbf{w}_i$ includes exactly the same variables as $\mathbf{x}_i$, that is to say $\mathbf{w}_i = \mathbf{x}_i$. As an exclusion restriction, we impose the first element of each $\boldsymbol{\gamma}_j$, namely the coefficient of the constant term, to be equal across j.[4]

The overall probability of a choice set C for individual i may then be written as:

$$\rho_i(C) = \prod_{j \in C} \Lambda_{ij} \prod_{l \notin c} (1 - \Lambda_{il}) \tag{2.7}$$

[4]The first element of $\boldsymbol{\gamma}_j$ is imposed equal across all alternatives so as to retrieve a sort of "average attention" which is common to all financial instruments regardless of individual characteristics. We show in Appendix A that such a restriction considerably facilitates parameter identification.

The choice of financial instrument from a given choice set can be written, using a multinomial logit formulation, as:

$$p_i^{MNL}(j \mid C) = \begin{cases} \frac{e^{\boldsymbol{\beta}_j' \mathbf{x}_i}}{\sum_{k \in C} e^{\boldsymbol{\beta}_k' \mathbf{x}_i}}, & \text{if} \quad j \in C \\ 0, & \text{if} \quad j \notin C. \end{cases}$$

Dardanoni et al. (2017) show that when $K \to -\infty$ the limiting unconditional probability of choosing alternative j can be written as follows:

$$p_i\left(j \mid \boldsymbol{\beta}, \boldsymbol{\gamma}; \mathbf{x}_i\right) = \rho_i(\emptyset) p_i^{MNL}\left(j \mid X\right) + \sum_{\emptyset \neq C : j \in C} \rho_i(C) p_i^{MNL}(j \mid C) \tag{2.8}$$

where X and $\emptyset$ represent the full and the empty choice set, respectively. The empty set represents the event in which all the alternatives suffer from the adverse utility shock K. In this last instance, individuals are assumed choosing from the full menu X (i.e. $p_i^{MNL}\left(j \mid \emptyset\right) = p_i^{MNL}\left(j \mid X\right)$).

2.3.1 Estimation

Let $\mathbf{d}_i = \langle d_{ij} \rangle_{j=0}^{5}$ denote the response vector for subject i, with $d_{ij} = 1$ if subject i selects alternative j. Our dataset is a sample of i.i.d. observations of individual choices and characteristics, namely $\langle \mathbf{d}_i, \mathbf{x}_i \rangle_{i=1}^{5729}$, with $\mathbf{x}_i$ determining both attention probabilities and utilities. The model log-likelihood is given by the sum of individual contributions and reads as

$$ll\left(\boldsymbol{\beta},\boldsymbol{\gamma}\right)=\sum_{i=1}^{5729}\sum_{j=0}^{5}d_{ij}log\left(p_i\left(j\mid\boldsymbol{\beta},\boldsymbol{\gamma};\mathbf{x}_i\right)\right). \tag{2.9}$$

Parameter estimates are obtained by numerically maximizing (2.9) and associated standard errors are computed taking the square root of the diagonal of the negative inverse Hessian of the log-likelihood.

2.3.2 Identification

Let $\boldsymbol{\Psi}=[\mathbf{B},\boldsymbol{\Gamma}]$ denote the parameter space of our limited attention model. Following Lewbel (2018), the model is said to be *point identified*[5] if no distinct pairs of parameters $\boldsymbol{\psi}=[\boldsymbol{\beta},\boldsymbol{\gamma}]$ and $\tilde{\boldsymbol{\psi}}=[\tilde{\boldsymbol{\beta}},\tilde{\boldsymbol{\gamma}}]$ in Ψ are observationally equivalent. That is to say that no pairs $\boldsymbol{\psi}$ and $\tilde{\boldsymbol{\psi}}$ in $\boldsymbol{\Psi}$ exist such that $\boldsymbol{\psi}\neq\tilde{\boldsymbol{\psi}}$ and $p_i\left(j\mid\boldsymbol{\beta},\boldsymbol{\gamma};\mathbf{x}_i\right)=p_i(j\mid\tilde{\boldsymbol{\beta}},\tilde{\boldsymbol{\gamma}};\mathbf{x}_i)$ for all i and j. Under suitable assumptions, the non-sigularity of the theoretical information matrix is a necessary and sufficient condition for point identification. Unfortunately, this principle cannot be exploited here since the complex structure of our model makes the information matrix analytically intractable.

In the context of limited attention model, point identification may be easily obtained by using auxiliary data (typically, survey data) on what alternatives are and are not taken into consideration by individuals (e.g., Draganska and Klapper (2011) and Honka and Chintagunta (2016)). Goeree (2008), instead, develops

[5]Depending on context, *global identification* or *frequentist identification* are often encountered as synonyms for *point identification*.

a model of limited attention where identification achieves through an exclusion restriction. In her model, advertising determines the set of alternatives entering a subject's consideration set but is constrained to exercise no impact on subject's utility. Abaluck and Adams (2017) show that utility and consideration set probabilities can be separately identified without excluding variables from attention or utility or relying on auxiliary data. Their identification proof constructively recovers consideration probabilities from asymmetries in the responsiveness of choice probabilities to characteristics of rival goods. The central insight of their proof is that changes in consideration probabilities can be expressed as a function of observable differences in cross-derivatives and market shares. Unfortunately, their argument cannot be applied in the context of our model since regressors vary across individuals but not across alternatives and, as such, cross-derivatives do not exist. Finally, Dardanoni et al. (2019) prove in a non-parametric framework that the distribution of tastes and attention probabilities can be separately elicited as long as the model is made dynamic, that is to say that individuals are repeatedly observed choosing. Again, their identification strategy cannot be exploited here due to the cross-sectional nature of our dataset.

As an alternative to point identification, we explore *local identifiability* of our model. For given true parameters $\boldsymbol{\psi}_0$, local identification of $\boldsymbol{\psi}_0$ means that there exists a neighborhood of $\boldsymbol{\psi}_0$ such that no $\boldsymbol{\psi} \in \boldsymbol{\Psi}$ exists in this neighborhood that is different from $\boldsymbol{\psi}_0$ and observationally equivalent to $\boldsymbol{\psi}_0$ (see, e.g., Lewbel (2018)).[6] Hence, let us impose the reduced-form parameters of our model to be equal to

[6]Local identification everywhere in $\boldsymbol{\Psi}$ is clearly a necessary but not sufficient condition for point identification.

the choice probabilities defined in (2.8). For given parameters $\boldsymbol{\psi_0} = [\boldsymbol{\beta_0}\ \boldsymbol{\gamma_0}]$, the mapping in (2.8) generates the corresponding reduced-form parameters $\boldsymbol{m}$, such that (suppressing the dependence on i for convenience)

$$m_j = p\left(j \mid \boldsymbol{\beta_0}, \boldsymbol{\gamma_0}; \mathbf{x}\right), \quad 1 \leq j \leq J. \tag{2.10}$$

As emphasized by Skrondal and Rabe-Hesketh (2004), $\boldsymbol{\psi_0}$ is (locally) identified if and only if $\boldsymbol{\psi_0} = [\boldsymbol{\beta_0}\ \boldsymbol{\gamma_0}]$ is the unique solution of the system of equations in (2.10). Using standard results of calculus, identification of $\boldsymbol{\psi_0}$ is then guaranteed if the Jacobian of the mappings $\langle p\left(j \mid \boldsymbol{\beta_0}, \boldsymbol{\gamma_0}; \mathbf{x}\right)\rangle_{j=1}^{J}$ is full (column) rank. In Appendix A we provide an application of the Jacobian method for local identification of our model.

Finally, we verify *empirical identification*[7] which is based on the estimated information matrix. Following Skrondal and Rabe-Hesketh (2004), if the information matrix computed at the maximum likelihood estimates $\widehat{\boldsymbol{\psi}}$ is non-singular, the model is empirically identified. In our application, the estimated information matrix is found to be full rank and its condition number[8] is around $35,000$. Denoting with K the number of regressors used in our application, our model contemplates $(KJ) + [(K-1)(J+1)+1] = 61$ free parameters. The first addend is the number of free utility parameters while the second one is the number of free attention pa-

[7]As argued by Skrondal and Rabe-Hesketh (2004), empirical identification is a useful complement to "theoretical" identification since it verifies identification where it matters, namely at estimated parameters.

[8]That is, the ratio between the maximum and the minimum singular values of a matrix.

rameters. Hence, our (61×61) covariance matrix $\mathbf{\Sigma}$ obtained from the estimated information matrix can be conveniently rewritten as

$$\mathbf{\Sigma} = \text{diag}(\boldsymbol{\sigma})\mathbf{R}\text{diag}(\boldsymbol{\sigma}), \tag{2.11}$$

where $\boldsymbol{\sigma}$ is the vector of standard deviations, $\text{diag}(\cdot)$ the diagonal matrix with its argument on the main diagonal and $\mathbf{R}$ the correlation matrix. The extreme eigenvalues of $\mathbf{\Sigma}$ respectively satisfy

$$\lambda_{\max}(\mathbf{\Sigma}) \leq \lambda_{\max}(\text{diag}(\boldsymbol{\sigma})\text{diag}(\boldsymbol{\sigma}))\lambda_{\max}(\mathbf{R}) = \max_i \sigma_i^2 \lambda_{\max}(\mathbf{R}), \tag{2.12}$$

and

$$\lambda_{\min}(\mathbf{\Sigma}) \geq \lambda_{\min}(\text{diag}(\boldsymbol{\sigma})\text{diag}(\boldsymbol{\sigma}))\lambda_{\min}(\mathbf{R}) = \min_i \sigma_i^2 \lambda_{\min}(\mathbf{R}). \tag{2.13}$$

From the two inequalities above, the following upper bound for the condition number of covariance matrix $\mathbf{\Sigma}$, denoted $\kappa(\mathbf{\Sigma})$, can be derived

$$\kappa(\mathbf{\Sigma}) \leq M = \frac{\max_i \sigma_i^2}{\min_i \sigma_i^2} \frac{\lambda_{\max}(\mathbf{R})}{\lambda_{\min}(\mathbf{R})}. \tag{2.14}$$

In our application M is found to be of the order of 10^5 and this further reassures us on the local identifiability of our model.

2.4 Results

2.4.1 MNL Model

Table 2.1 reports the maximum likelihood estimates of the parameters of the multinomial logit model presented above. The baseline category is represented by money market funds and other liquid investments and the corresponding parameters which affect utility are imposed to be zero.

First and foremost, we observe a substantial heterogeneity across categories with regard to their responsiveness to selected regressors. Age, for instance, is found to exercise a significant and negative effect on the probability of preferring the two riskiest alternatives, namely stock funds and employer company stock, over the baseline. Such a finding is consistent with previous literature concerning life-cycle investing models (e.g., Bodie et al. (1992)). Indeed, younger workers enjoy higher flexibility to modify their labor supply and, as such, can offset more easily the changes in the value of their risky financial assets by changing the amount they work. Thus, if younger workers have more opportunity to alter their labor supply than older workers, the share of assets held as risky equity should decline with age.

As far as gender is concerned, our estimates confirm established results on gender differences in risk aversion (e.g., Sunden and Surette (1998)). Male workers appear to be more likely to opt for (risky) stock funds than their female counterparts, coeteriis paribus. As a result, higher women risk aversion could translate into large differences in the accumulation of financial wealth for retirement.

Focusing on the choice of investing in employer company stock, Choi et al. (2005) document a substantial ineffectiveness of education in reducing employer stock holdings. In our model education is coded as a dummy taking value one if the individual holds a bachelor or a higher degree. In contrast to Choi et al. (2005), we observe a significant effect of college education in attenuating the bias for directing contributions toward employer company stock.

The presence of an employer match has a significant and positive impact on the probability of choosing the three riskiest alternatives, namely employer stock, stock funds and diversified stock and bond funds. The effect is particularly acute with reference to employer stock which may signal the need to extend to DCPs the restrictions currently imposed by ERISA on the maximum amount (namely, 10% of total assets) that DBPs can invest in the stock of the employer.

Finally, the percentage of salary contributed does not seem to play a relevant role in shaping investment decisions even though it exercises a mild positive effect in favoring diversification strategies.

2.4.2 Limited Attention Model

Maximum likelihood estimates of the parameters of the limited attention model are reported in the rightmost columns of Table 2.1. Notice that attention parameters often exhibit opposite sign with respect to their utility counterparts. That is to say that individual and job-specific characteristics may have different impact on attention and preferences and, as such, subjects may well not pay attention to their preferred alternatives. Male workers, for instance, tend to disregard corporate bonds and bond funds even though being male has a positive effect on the utility arising from this asset category. College education, instead, makes investors more likely to consider bonds while having a negative impact on utility. As for stock funds, workers who contribute more tend to derive higher utility form this kind of financial assets but at the same time they are less likely to include them in their consideration set.

Estimates of the $\boldsymbol{\beta}$ coefficients obtained under the limited attention model where consideration sets are taken into account are not immediately interpretable given that they interact with the attention parameters $\boldsymbol{\gamma}$ in a non-linear fashion. Thus, in order to evaluate the impact of each regressor on the probability of choosing one of the six alternatives we refer to the average marginal effects reported in Table 2.2. The marginal effects computed under the limited attention model tend not to be very different from the ones obtained under the standard multinomial logit model.

For example, a one standard deviation increase in age makes the probability of choosing stock funds 1.34% lower under the multinomial model while this reduction amounts to 1.18% under the limited attention model. Similarly, a decrease in risk exposure is associated with an increase in age as suggested by the marginal effect of this variable on the probability of choosing diversified stock and bond funds under both model formulations.

Furthermore, we still document gender difference in risk aversion given that, under both models, male workers' probability of opting for stock funds is around

2.15% higher than the one of their female counterparts, coeteriis paribus.

Coming to education, holding a bachelor or a higher degree appears to be powerful in preventing workers from allocating their contributions in their employer stocks even under the limited attention specification (the negative impact on the probability of investing in employer stocks is around 2.5% under both models). The same holds true for the presence of an employer match which still produces an increase of 1.48% on the probability of choosing employer stocks.

All in all, the two models do not imply considerable differences in terms of marginal effects. This result may appear puzzling but it is likely to be caused by two competing forces which can be disentangled under the limited attention model. On the one side, limited attention reduces financial instruments comparisons while, on the other side, it exacerbates the utility/disutility arising from some regressors. In fact, some of the ratios of the utility parameters obtained under the limited attention model over the corresponding ones estimated in the plain multinomial logit model are of an extremely high/low magnitude. Thus, setting consideration effects aside, the two models are likely to generate significantly different preference orderings and this justifies the implementation of a further counterfactual specification discussed below.

2.4.3 Counterfactual Analysis

To isolate the effects of limited attention, we also compute marginal effects under a counterfactual specification in which we let each $\boldsymbol{\gamma_j} \to \infty$ (i.e. we resort to a standard multinomial framework where each alternative is considered with probability 1 by the decision maker) while adopting the utility parameters $\boldsymbol{\beta_j}$ obtained from the maximization of the log-likelihood function in (2.9). In a sense, this extension can be held equivalent to a fictitious policy measure which can make workers fully attentive by providing them with appropriate information. These full attention hypothetical probabilities represent a counterfactual which allows us to elicit only the "taste" components from the utility specified in (2.5).

Counterfactual marginal effects are reported in the third column of Table 2.2 and deserve some comments. First and foremost, we document a striking increase in the absolute value of marginal effects across all alternatives but employer company stock. Additionally, some effects exhibit opposite signs with respect to previous estimates.

A one standard deviation increase in age, for example, makes workers 8.08% more likely to choose corporate bonds and bond funds while decreasing their probability of investing in diversified stock and bond funds by 11.92%. Such behavior is consistent with the life-cycle investing model which postulates a decrease in risk exposure as age increases.

Moreover, the gender difference in risk aversion claimed above is not supported here where males are found to be 9.14% less likely than females to choose stock funds. Thus, female may just be less likely to include stocks in their consideration sets.

Additionally, employee percentage contribution tuns out to be a very important driver of investment decision under the full attention specification. Reasonably enough, workers who contribute more tend to invest more aggressively by decreasing their exposure to money market funds and other liquid investments while chasing higher returns through both stock funds and diversified funds.

Focusing on employer stock, the marginal effects are particularly instructive. Indeed, no more significant positive impact of the matching contribution emerges: the limited attention model is able to account for a kind of endogeneity issue in that matching contributions are usually directed toward employer stock by default (at least partially). This fact is captured by the positive impact of employer matching at the consideration stage which becomes negligible at the choice stage. As a result, a laissez-faire policy on the maximum amount that can be invested in employer stocks seems to be justified. Finally, a word of caution is necessary with respect to the effect of college education on the choice of asset categories. Under full attention, education becomes virtually irrelevant in its ability to make investors aware of the riskiness of employer stocks which may signal the urgency for campaigns promoting financial literacy.

2.5 Conclusion

Accumulated literature has almost exclusively analyzed the investing behavior in DCPs by exploiting event studies (e.g., Choi et al. (2002)). The main exception is represented by Sunden and Surette (1998) who develop a standard multinomial logit model in order to detect gender differences in investment decisions. The limited attention model described in this chapter relaxes the strong assumption that agents consider all of the options available to them before making a choice. In our model the act of choosing is conceptualized as a two-stage process. In the first stage individuals form their consideration set while in the second one they make their final choice. Since individuals do not necessarily consider what they prefer, failing to disentangle utility and attention may severely under- or overestimate the impact of a worker's characteristics on the probability of choosing a given financial instrument that enters her consideration set. This is so despite the fact that the unconditional choice probabilities implied by our limited attention model do not differ considerably from the ones implied by its standard multinomial counterpart. While remaining agnostic regarding the source of limited attention, our model can be fruitfully exploited for the twofold purpose of anticipating what financial instrument(s) a given worker is more likely to consider and, conditioning on that, of gauging a better understanding of the asset class(es) which are likely to provide her with higher utility. Note that we can serve this purpose by letting the same individual characteristics have an impact on both attention and utility and without relying on strong exclusion restrictions (see, e.g., Goeree (2008)).

Depending on data availability, the most natural extension of our model would be to also include alternative-specific regressors such as the return and/or some measures of risk of the financial asset chosen by each agent. Finally, if subjects in our sample were observed choosing repeatedly, we could have established global identifiability of our model by invoking recent results on the identification of cognitive heterogeneity in the context of discrete choice models (see, e.g., Dardanoni et al. (2019)).

Table 2.1: Estimation Results

Variable	MNL Model		MNL Model with CS			
			Consideration Stage		Choice Stage	
	Coefficients	se	Coefficients	se	Coefficients	se
Reference Category: Money Market Funds and Other Investments - Count: 1310						
constant	—	—	0.154	0.101	—	—
age	—	—	-0.206	0.093	—	—
sex	—	—	-0.389	0.234	—	—
college education	—	—	0.210	0.231	—	—
matching contribution	—	—	-0.304	0.212	—	—
employee % contribution	—	—	0.556	0.105	—	—
Category: Government Securities - Count: 198						
constant	-1.974	0.202	0.154	0.101	-5.987	2.012
age	0.063	0.077	1.436	0.439	-1.723	0.873
sex	0.022	0.153	5.191	2.911	-3.968	1.412
college education	-0.092	0.154	1.312	0.628	-0.167	1.086
matching contribution	0.141	0.191	-0.958	0.901	3.177	1.336
employee % contribution	-0.035	0.089	-0.634	0.273	0.975	0.347
Category: Diversified Stock and Bond Funds - Count: 2839						
constant	0.435	0.089	0.154	0.101	1.526	0.723
age	0.010	0.034	0.244	0.070	-0.627	0.326
sex	0.094	0.067	-0.107	0.205	-0.338	0.581
college education	0.034	0.067	0.086	0.159	0.222	0.498
matching contribution	0.341	0.084	0.866	0.634	-1.258	0.653
employee % contribution	0.079	0.037	-0.108	0.121	1.502	0.540
Category: Corporate Bonds or Bond Funds - Count: 81						
constant	-2.859	0.301	0.154	0.101	-5.277	2.069
age	0.111	0.116	-1.467	0.375	1.879	0.665
sex	0.202	0.231	-4.378	0.869	5.732	2.007
college education	-0.156	0.233	2.085	0.610	-2.629	1.118
matching contribution	0.046	0.282	1.609	0.813	-1.921	1.786
employee % contribution	0.084	0.113	-0.399	0.213	1.842	0.600
Category: Stock Funds - Count: 1025						
constant	-0.584	0.114	0.154	0.101	-1.814	0.810
age	-0.095	0.042	-0.223	0.160	-0.334	0.235
sex	0.211	0.084	0.470	0.552	-0.824	0.829
college education	-0.142	0.085	0.633	0.320	-0.675	0.636
matching contribution	0.348	0.108	-0.645	0.549	1.891	0.806
employee % contribution	-0.073	0.050	-0.608	0.162	1.804	0.468
Category: Employer Company Stock - Count: 276						
constant	-1.955	0.203	0.154	0.101	-4.424	1.592
age	-0.232	0.067	-0.480	0.264	-0.728	0.800
sex	0.033	0.134	-0.824	0.425	-0.073	0.689
college education	-0.593	0.141	-0.602	0.412	0.046	0.733
matching contribution	0.687	0.196	1.708	0.523	0.278	1.190
employee % contribution	-0.121	0.088	1.568	0.481	-0.853	0.542

Table 2.2: Average Marginal Effects

	MNL	Limited Attention	Counterfactual
Money Market Funds and Other Investments			
age	0.44%	0.39%	6.44%
sex	-2.04%	-1.96%	2.96%
college education	0.92%	1.38%	3.69%
matching contribution	-6.45%	-6.06%	16.07%
employee % contribution	-0.46%	-1.70%	-27.39%
Government Securities			
age	0.29%	0.28%	-1.62%
sex	-0.23%	-0.09%	-2.67%
college education	-0.18%	-0.22%	-0.18%
matching contribution	-0.43%	-0.36%	1.57%
employee % contribution	-0.19%	-0.62%	-0.15%
Diversified Stock and Bond Funds			
age	1.44%	1.53%	-11.92%
sex	0.23%	-0.11%	-5.26%
college education	3.71%	3.35%	12.27%
matching contribution	3.92%	3.76%	-27.59%
employee % contribution	2.89%	3.89%	16.14%
Corporate Bonds or Bond Funds			
age	0.18%	0.16%	8.08%
sex	0.16%	0.05%	13.98%
college education	-0.16%	-0.23%	-8.68%
matching contribution	-0.32%	-0.64%	-6.17%
employee % contribution	0.09%	0.04%	3.53%
Stock Funds			
age	-1.34%	-1.18%	-0.57%
sex	2.16%	2.15%	-9.14%
college education	-1.77%	-1.80%	-7.16%
matching contribution	1.51%	1.82%	15.53%
employee % contribution	-1.65%	-0.81%	9.37%
Employer Company Stock			
age	-1.01%	-1.18%	-0.41%
sex	-0.28%	-0.04%	0.13%
college education	-2.52%	-2.48%	0.06%
matching contribution	1.77%	1.48%	0.59%
employee % contribution	-0.67%	-0.80%	-1.50%

3

Productivity and Ownership Concentration: A Structural Approach

3.1 Introduction

The following paper investigates of the relationship between shareholders' ownership structure and firm's productivity. In the seminal model of Modigliani and

Miller (1958), ownership is expected to exercise no impact on firm performance. Recognizing the separation of ownership and control as a potential source of conflict of interest, however, traces back to Smith (1776). The seminal contribution of Berle and Means (1776) suggests that the diffuseness of shareholdings should be inversely related to firm performance. More recently, the principal-agent model developed by Jensen and Meckling (1976) shows how the distribution of shares between insiders and outsiders can exercise an impact on firm behavior. In particular, the agency costs related to the conflict of interests between managers and outsider owners could be mitigated through increases in ownership concentration since a more concentrated ownership structure provides to large shareholders greater incentives to engage in costly monitoring.

Even though the presence of blockholders is commonly considered as a constraint of managerial opportunism, blockholders's interests are not necessarily convergent and conflicts are likely to arise. Conflicts among blockholders, denoted as principal-principal problem, emerge because there exists private benefits of control that are not enjoyed by non-controlling shareholders.

In sum, there are at two potential sources of conflict related to the distribution of ownership that may affect firm value through their impact on the decision making process: the relationship between managers and owners and the relationship among owners. In the presence of financial constraints, requiring the participation of multiple investors, the degree of ownership concentration can be an effective internal corporate governance mechanism to align the incentives of owners and managers towards the maximization of firm value. An increase in ownership concentration can have a positive effect on firm value because of the monitoring role that can be played by large investors to prevent managerial opportunism. But at the same time, an increase in ownership concentration can be detrimental due to the expropriation of minority shareholder by large shareholders.

The empirical evidence on the relationship between ownership concentration and firm value is mixed and not conclusive. Some authors (e.g, Konijn et al. (2011)) document a positive relationship between ownership concentration and firm profitability while others (e.g., Laeven and Levine (2008) and Attig et al. (2009)) hold that a high level of concentration negatively influences corporate performance. Another stream of research points towards a non-monotonic relationship (e.g., Morck et al. (1988), De Miguel et al. (2004), Russino et al. (2019)). Others do not find any significant relationship (Demsetz and Villalonga (2001)).

The disagreement in empirical research can be traced back to contextual and methodological differences. The two dimensions related to the increase in ownership concentration (the monitoring of managers and the potential conflict among shareholders) make the analysis of the relationship between the degree of ownership concentration and firm value a complex empirical question. In countries, or more generally environments, where ownership is widely dispersed, like the U.S., the main source of conflict is managerial opportunism and we can expect that ownership concentration will be positively related to firm value. However, in settings where the fractional ownership of the largest shareholders is high, the main problem is not the alignment of interests of the managers but the potential expropriation of non-controlling shareholders. The effect of increasing ownership concentration on firm value will depend on its effect on the conflict of interests among shareholders. In addition, institutional differences, such as the level of legal protection offered to minority shareholders, and market conditions, such as the functioning of the market for corporate control, will be important factors affecting the empirical relationship between the ownership distribution and firm value.[1]

From the methodological point of view, three issues arise. The first concerns the measurement of firm value. Previous literature has relied upon accounting based measurements, such as return on assets, or market based measurements, such as Tobin Q. The use of accounting or market based ratios as proxies of firm performance has been criticized because they generally depend on accounting practices and manipulations and, additionally, market value ratios are affected by investors' sentiment and financial market characteristics (Demsetz and Villalonga (2001)). The second issue relates to the measurement of ownership concentration. As stressed by Overland et al. (2012), diverse measures can capture different dimensions related to the distribution of ownership. For instance, measures like the share of the largest block or the sum of the shares of a number of largest shareholders is better suited to represent the dimension related to the monitoring of managers, while measures representing the relative size of shareholders are more appropriate to capture the conflicts of interests among shareholders. Finally, the empirical analysis of the relationship between ownership concentration and firm value is plagued by a serious problem of endogeneity because the ownership structure may depend on some variables (such us managerial ability) that are observed

[1] The ownership structure is an internal corporate governance mechanism that will be more relevant in settings where alternative external governance mechanisms do not work.

by firms when the distribution of ownership is chosen but are unobserved by the econometrician. To overcome the biases arising from OLS estimation, previous studies have mainly implemented alternative static econometric methods such as Fixed Effects (FE) and Instrumental Variables (IV) estimation. As stressed by Roberts and Whited (2013), both these approaches have very limited application: FE can address the endogeneity problem only under the restrictive assumption that the source of endogeneity is invariant over time, while IV method relies on the identification of relevant and valid instruments (sources of truly exogenous variation correlated with the endogenous regressors) that is extremely hard in the corporate governance setting.

In this paper we study the relationship between ownership concentration and firm value using a sample of Italian listed manufacturing firms. The novelty of the paper can be summarized as follows. First, we use total factor productivity (TFP) to measure firm value, a more primitive measure of firm performance that is not affected by accounting procedures and stock market volatility. Second, we use measures of concentration that consider the entire ownership distribution allowing us to capture the interplay among shareholder and to shed light on the shareholders' conflict dimension. Additionally, we address the issue of the divergence between the distribution of control power and the distribution of cash flow rights. Finally, our choice of measuring firm performance in terms of productivity allows us to adopt a structural approach à la Olley and Pakes (1996) to analyze the relationship between ownership concentration and firm value. That is, we exploit structural modelling to deal with the endogeneity problem. We specify a semi-parametric model which controls for firms' unobserved heterogeneity and for the endogeneity of input factors and variables representing the ownership structure. The big advantage of this approach is that the parameters of the structural production function can be estimated without fully specifying the firm's decision making problem. The approach relies on timing assumptions about the inputs choices and about the firms' information sets at the time the inputs are chosen, but does not requires the solution of the corresponding complex dynamic optimization problem.

3.2 The concept of productivity

In the attempt to unveil the role of ownership structure, previous literature has relied upon financial measures of firm performance (accounting profit ratios such as return on assets and return on equity or market value ratios such as Tobin Q). The use of financial ratios as proxies of firm value has been criticized because they generally depend on accounting practices and manipulations and, additionally, market value ratios are affected by investors' sentiment and financial market characteristics (Demsetz and Villalonga (2001)).

In this paper instead of using a profitability ratio to proxy firm value we utilize firm productivity, a more primitive measure of firm performance that is not affected by accounting procedures and stock market volatility. Following Hulten (2001), productivity gauges the efficiency by which inputs are turned into outputs. A reliable index of efficiency needs to make allowance for, and attach proper weight to, the contributions of all the inputs in the production process.

Such an index is typically represented by *Total Factor Productivity* (TFP) which is defined as the ratio of total output to the (weighted) sum of associated labor and capital (factors) inputs. Thus, letting L_{it} and K_{it} denote, respectively, labour and capital input employed by firm i in period t, TFP is given by

$$\tau_{it} = \frac{Y_{it}}{f(L_{it}, K_{it})}, \tag{3.1}$$

where Y_{it} denotes total output and $f(L_{it}, K_{it})$ represents total input. From (3.1), TFP emerges as the portion of output which is not explained by the amount of inputs used. As a result, a higher level of τ_{it} is brought about, ceteris paribus, by a more efficient and intense utilization of the inputs in the production process.

Letting Y_{it} become the dependent variable, (3.1) can be rearranged as

$$\begin{aligned} Y_{it} &= \tau_{it} f(L_{it}, K_{it}) \\ &= Y(\tau_{it} f(L_{it}, K_{it})). \end{aligned} \tag{3.2}$$

which corresponds to a standard neoclassical production function technology where τ_{it} represents the Hicksian neutral shift parameter.

In particular, assuming Cobb-Douglas[2] technology, we obtain

[2] While imposing unit elasticity of substitution between inputs, Cobb-Douglas functions represent a valid first-order logarithmic approximation of more complicated specifications which are often encountered in the literature. Maddala (1979), for instance, shows that Cobb-Douglas function produces estimates which are substantially close to the ones obtained with several other functions such as generalized Leontief, homogeneous translog, and homogeneous quadratic.

$$Y(\tau_{it} f(L_{it}, K_{it})) = \tau_{it} L_{it}^{\alpha} K_{it}^{\beta}. \tag{3.3}$$

Letting $\alpha_0 = \ln(\tau_{it})$ and taking logharithms of (3.3), we obtain the following estimating equation

$$y_{it} = \alpha_0 + \alpha l_{it} + \beta k_{it} + \epsilon_{it}, \tag{3.4}$$

where lower case letters denote logarithms, e.g., $l_{it} = \ln(L_{it})$, and ϵ_{it} represents an error term.

3.3 Endogeneity issue in productivity estimation

The most pervasive problem in productivity estimation studies, arguably, is endogeneity which can loosely be defined as the presence of correlation between the error term and covariates in a regression. Indeed, the error term can be assumed to be made of some components which are known or predictable by the firm when decisions are taken. Observed inputs are typically responsible for endogeneity in that they are chosen by the firm taking into account unobservable (to the econometrician) components of production. As of our main regressor of interest, if ownership concentration may affect firm performance through its impact on business activities, it may well be the case that performance itself shapes to some extent the contracting environment impacting ownership distribution among blockholders.

Traditional static solutions adopted by previous literature, Instrumental Variable (IV) or Fixed Effects (FE) estimation, have not generally proved effective in addressing the endogeneity issue.

The IV approach tries to exploit the exogenous variation of input shifters such as input prices. However, input prices are often unavailable and, additionally, using them to instrument inputs may well prove ineffective due to insufficient variation. If firms use homogeneous inputs and operate in the same output and input markets, we should not expect to find any significant cross-sectional variation in input prices. The introduction in the model of ownership related variables makes the implementation of IV estimation much harder. Indeed, it is extremely difficult to identify relevant valid instruments for variables related to the ownership structure.

As for the FE, the underlying assumption of unobserved firm heterogeneity being time-invariant seems too restrictive. Assuming that unobserved productivity varies just across firms and excluding within-firm time-series variation is not held adequate to identify output elasticities of production factors and to drive out the relationship between efficiency and ownership. Additionally, ownership related variables are generally highly persistent and FE estimation will provide poor estimates.

Actually, state-of-the-art techniques for productivity estimation can be traced back to two alternative methodologies which are commonly referred to as the “dynamic panel” approach and the “proxy variable” approach, respectively.

Instead of exploiting the exogenous variability provided by some instruments, both sets of techniques are concerned with modelling firm behavior by means of appropriate assumptions. Such assumptions mainly concern the timing of input choices and the information set that firms face when deciding on the utilization of input factors. However, both approaches are not "fully" structural since they can be implemented without analytically solving the complex dynamic optimization problem faced by firms and, as such, parameters of interest can be estimated in a semi-parametric setup.

3.3.1 Dynamic Panel Approach

While being a quite general methodology that stems from the vast literature on panel data models (see, e.g., Chamberlain (1982), Anderson and Hsiao (1982), Arellano and Bond (1991)), Blundell and Bond (2000) provide an application of this technique to production functions. They consider a Cobb-Douglas production function of the form

$$
\begin{aligned}
y_{it} &= \alpha l_{it} + \beta k_{it} + \gamma_t + (\eta_i + m_{it} + \upsilon_{it}), \\
\upsilon_{it} &= \rho \upsilon_{it-1} + \epsilon_{it}, \quad |\rho| < 1, \\
&\epsilon_{it}, m_{it} \sim MA(0),
\end{aligned}
\tag{3.5}
$$

where γ_t denotes year fixed effect and the error component within parentheses is made of the following three terms:

- an unobserved firm-specific effect η_i which is allowed to be arbitrarily correlated with inputs;

- a further innovation term m_{it} which proxies for measurement error of the inputs;[3]

- an autoregressive shock to productivity υ_{it}.

In its original formulation, (3.5) is estimated under the timing assumption that l_{it} and k_{it} are chosen by the firm before the realization of ϵ_{it}. Nonetheless, input factors are deemed endogeneous since they likely correlate with the predictable part of the productivity shock υ_{it},[4] that is $\rho\upsilon_{it-1}$. Taking the ρ first-difference of (3.5) gives

$$\begin{aligned} y_{it} - \rho y_{it-1} = \alpha l_{it} - \alpha\rho l_{it-1} + \beta k_{it} - \beta\rho k_{it-1} + (\gamma_t - \rho\gamma_{t-1}) \\ + (\eta_i(1-\rho) + m_{it} - \rho m_{it-1} + \epsilon_{it}), \end{aligned} \tag{3.6}$$

which allows to isolate the innovation in υ_{it} and get rid of the source of endogeneity that υ_{it} potentially contains (i.e. $\rho\upsilon_{it-1}$).

However, to eliminate the firm fixed effect η_{it} a further differencing is needed (see, for instance, Arellano and Bond (1991)) which yields

[3] For the sake of exposition, we assume m_{it} to be orthogonal to inputs choices in all periods. Ackerberg (2016), for instance, shows how this assumption can be easily relaxed.

[4] In addition to being correlated with the firm-fixed effect η_i.

$$
\begin{aligned}
(y_{it} - \rho y_{it-1}) - (y_{it-1} - \rho y_{it-2}) = {} & \alpha\left[(l_{it} - \rho l_{it-1}) - (l_{it-1} - \rho l_{it-2})\right] \\
& + \beta\left[(k_{it} - \rho k_{it-1}) - (k_{it-1} - \rho k_{it-2})\right] \\
& + \left[(\gamma_t - \rho\gamma_{t-1}) - (\gamma_{t-1} - \rho\gamma_{t-2})\right] \\
& + \left[(m_{it} - \rho m_{it-1}) - (m_{it-1} - \rho m_{it-2})\right] \\
& + (\epsilon_{it} - \epsilon_{it-1}).
\end{aligned} \tag{3.7}
$$

Under the timing assumption that firm observes shocks up to ϵ_{it-1} when choosing inputs in period t,[5] the model in (3.7) can be estimated by exploiting the following moment conditions

$$
E\left[(m_{it} - \rho m_{it-1}) - (m_{it-1} - \rho m_{it-2}) + (\epsilon_{it} - \epsilon_{it-1}) \mid l_{it-1}, k_{it-1}\right] = 0. \tag{3.8}
$$

However, the model in (3.7) turns out to be particularly data demanding given that microeconomic datasets usually contemplate a high number of individuals (i.e. large N) observed over a relatively small period of time (i.e. small T). Additionally, it is often found to produce imprecise estimates. In order to overcome the two issues raised above, Blundell and Bond (2000) suggest to rely upon a further stationarity assumption which constraints the way in which the firm fixed effect η_i correlates with inputs. They assume that

[5] Ackerberg (2016) shows how timing assumptions can be further strengthened so as to let the productivity shock be orthogonal to inputs lagged more than just one period (e.g., $k_{it+\Delta}$ with $\Delta < -1$).

$$E[\eta_i \mid l_{it} - l_{it-1}, k_{it} - k_{it-1}] = 0. \tag{3.9}$$

Condition (3.9) can be interpreted as a stationarity assumption of the firm operating environment. If (3.9) is imposed, that is if η_i is allowed to be correlated with the levels of inputs but assumed orthogonal to variations of these inputs between consecutive periods,[6] no need for second differencing arises and (3.6) could directly be estimated.

3.3.2 Proxy Variable Approach

The proxy variable approach stems from the seminal contribution of Olley and Pakes (1996) and several methodological refinements have been proposed by, among others, Levinsohn and Petrin (2003), Wooldridge (2009) and Ackerberg et al. (2015). Assuming, as before, Cobb-Douglas technology, let (log of) production be governed by

$$y_{it} = \beta_0 + \alpha l_{it} + \beta k_{it} + \omega_{it} + \epsilon_{it}. \tag{3.10}$$

Equation (3.10) contains the following error terms:

- ω_{it} is a shock to productivity that, while unknown to the econometrician, is (at least partially) observable or predictable by the firm before deciding on

[6]Note that assumption (3.9) implicitly rules out the possibility for more productive firms (namely, those with a high η_i) to grow faster than less productive firms. Obviously, this is hard to justify for innovative and growing industries.

inputs at time t;

- ϵ_{it}, instead, is a true error component which possibly originates from an unpredictable productivity shock and/or serially correlated measurement error in output.

In a manufacturing process, for instance, ω_{it} might represent the number of defected items that the firm is able to foresee while ϵ_{it} would capture deviations from the expected defect rate brought about by unpredictable changes in the operating environment.

Arguably, ω_{it} is the problematic innovation term since it likely correlates with both l_{it} and k_{it} making OLS estimates of α and β in (3.10) inconsistent. Endogeneity occurs since input choices by the firm are influenced by firm's beliefs about ω_{it}.

To overcome this endogeneity issue, Olley and Pakes (1996) propose a discrete time model of firm behavior in which firms dynamically maximize the expected discounted value of future cash flows. At each period t, a firm's value function is given by the following Bellman equation[7]

$$V_t(\omega_t, k_t) = \max \left\{ \Phi_t, \sup_{i_t \geq 0} \pi_t(\omega_t, k_t) - c(i_t) + \delta E \left[V_{t+1}(\omega_{t+1}, k_{t+1} \mid I_t \right] \right\}, \quad (3.11)$$

where Φ_t is the firm's sell-off value, $\pi_t(\omega_t, k_t)$ is the firm's profit function, $c(i_t)$ is the cost of investment i_t, δ is the firm's discount rate and I_t is the firm's information set at period t. Luckily, identification of output elasticities does not require to

[7]For ease of exposition we suppress the i subscript from (3.11).

explicitly solve (3.11) and a two-stage estimation procedure can be deployed once the following assumptions are imposed.

Assumption 1: ω_{it} follows a first-order Markov process and evolves according to the distribution

$$p(\omega_{it+1} \mid I_t) = p(\omega_{it+1} \mid \omega_{it}) \tag{3.12}$$

which is known to the firm and stochastically increasing in ω_{it}.

Assumption 2: Labour factor l_{it} is non-dynamic in that it can freely be varied by the firm and, as such, it does not impact the future stream of profits. Capital k_{it}, instead, is a state variable since it accumulates according to

$$k_{it+1} = (1 - d)k_{it} + i_{it}, \tag{3.13}$$

with d being the annual depreciation rate. Thus, capital is assumed to be dynamic given that, according to (3.13), $k_{it+1} \in I_t$, or equivalently, it takes a whole period for the firm to order, receive and install new capital.

Assumption 3: ω_{it} is the only scalar unobservable and impacts firm's decision on investment through

$$i_{it} = f_t(k_{it}, \omega_{it}). \tag{3.14}$$

This assumption implies that all firms in the industry face the very same conditions as for the markets for inputs and output and, as such, no source of heterogeneity across firms other than ω_{it} is allowed.

Assumption 4: $f_t(k_{it}, \omega_{it})$ is monotonically increasing in ω_{it}.

Despite being stated as an assumption, it can be formally shown that, after imposing Assumption 1, the solution to the dynamic programming problem in (3.11) implies an optimal investment demand which is monotonically increasing in ω_{it}. Intuitively, $p(\omega_{it+1} \mid \omega_{it})$ being stochastically increasing in ω_{it} suggests that firms with higher ω_{it} enjoy higher expected marginal productivity of the fixed input factor (namely, capital) and, as such, will engage in higher investment in the future.

The immediate consequence of Assumptions 3-4 is the invertibility of the investment function. Thus, unobserved heterogeneity ω_{it} can be made a function of observable variables, that is

$$\omega_{it} = f_t^{-1}(k_{it}, i_{it}). \tag{3.15}$$

Substituting (3.15) in the estimating equation (3.10) gives

$$y_{it} = \beta_0 + \alpha l_{it} + \beta k_{it} + f_t^{-1}(k_{it}, i_{it}) + \epsilon_{it}. \tag{3.16}$$

To explicitly derive $f_t^{-1}(k_{it}, i_{it})$, one would need to make several further assumptions on model primitives and solve (3.11). Olley and Pakes (1996), instead, suggest to treat $f_t^{-1}(k_{it}, i_{it})$ non-parametrically and to estimate the following equation

$$y_{it} = \alpha l_{it} + \phi_t(k_{it}, i_{it}) + \epsilon_{it}, \tag{3.17}$$

with $\phi_t(k_{it}, i_{it})$ being a non-parametric composite term such that $\phi_t(k_{it}, i_{it}) = \beta_0 + +\beta k_{it} + f_t^{-1}(k_{it}, i_{it})$.

In the first stage, estimates $\widehat{\alpha}$ and $\widehat{\phi}_t(k_{it}, i_{it})$ can be obtained by exploiting the following moment condition

$$E[\epsilon_{it} \mid I_{it}] = E[y_{it} - \alpha l_{it} - \phi_t(k_{it}, i_{it}) \mid I_{it}] = 0. \tag{3.18}$$

Then, from the markovian assumption we can let ω_{it} evolve according to

$$\omega_{it} = E[\omega_{it} \mid \omega_{it-1}] + \xi_{it} = g(\omega_{it-1}) + \xi_{it}, \tag{3.19}$$

with $E[\xi_{it} \mid I_{t-1}] = 0$. Note that (3.19) is able to make allowance for a much more general behavior of unobserved heterogeneity ω_{it} than the AR(1) process typically assumed in the dynamic panel literature (e.g., see (3.5)).

Using first-stage estimates $\widehat{\alpha}$ and $\widehat{\phi}_t(k_{it}, i_{it})$ and being

$$\omega_{it} = \phi_t(k_{it}, i_{it}) - \beta_0 - \beta k_{it}, \tag{3.20}$$

we can plug (3.19) into (3.10) and obtain the following second-stage estimating equation

$$\begin{aligned} y_{it} &= \beta_0 + \widehat{\alpha} l_{it} + \beta k_{it} + g(\omega_{it-1}) + \xi_{it} + \epsilon_{it} \\ &= \beta_0 + \widehat{\alpha} l_{it} + \beta k_{it} + g\left(\widehat{\phi}_{t-1}(k_{it-1}, i_{it-1}) - \beta_0 - \beta k_{it-1}\right) + \xi_{it} + \epsilon_{it}. \end{aligned} \tag{3.21}$$

From (3.21) the following second-stage moment condition derives

$$E\left[\xi_{it} + \epsilon_{it} \mid I_{it-1}\right] = E\Big[y_{it} - \beta_0 + \widehat{\alpha} l_{it} + \beta k_{it} + g\left(\widehat{\phi}_{t-1}(k_{it-1}, i_{it-1}) - \beta_0 - \beta k_{it-1}\right) \mid I_{it-1}\Big] = 0, \tag{3.22}$$

which can be exploited to obtain estimates $\widehat{\beta}_0$ and $\widehat{\beta}$.

The two-stage procedure described above has been subject to several refinements. Levinsohn and Petrin (2003), for instance, suggest to overcome the problem of lumpy investments[8] by using the demand for intermediate inputs (e.g., electricity and raw materials) to proxy for ω_{it}.

Ackerberg et al. (2015), instead, question the non-dynamic nature of labour and show that identification of output elasticity of labour (namely, α) in the first stage is ensured under just a few specific data generating processes. In their model, the investment function[9] in (3.15) is replaced by $i_{it} = f_t(k_{it}, l_{it}, \omega_{it})$ which clearly prevents one from identifying α in the first stage. Thus, all parameters of interest are retrieved from the second-stage moment condition.

Finally, Wooldridge (2009) proposes a GMM procedure in which first and second-stage moment conditions (namely, equations (3.18) and (3.22)) are jointly imposed within a suitable minimization routine in order to achieve higher efficiency and easier standard error calculations.

[8] Firm's investment is typically a large and infrequent episode (see, for instance, Doms and Dunne (1998)). To make invertibility of investment possible, OP methodology forces to drop all observations such that $i_{it} = 0$.

[9] Ackerberg et al. (2015) also show how their approach can be implemented when the demand for intermediate inputs instead of investment is used to proxy for ω_{it} as suggested by Levinsohn and Petrin (2003).

3.4 The effect of ownership structure on productivity

Our model originates predominantly from the proxy variable approach. Departing from much of previous literature where output is usually assumed to depend on just labour and capital, we assume a $3-$factor Cobb-Douglas production function of the form

$$Y_{it} = \tau_{it} L_{it}^{\alpha} K_{it}^{\beta} M_{it}^{\gamma}, \tag{3.23}$$

with M_{it} representing intermediate inputs. Since our interest lies in evaluating the impact of ownership structure on firm's productivity, we explicitly model τ_{it} as a function of variables representing the distribution of ownership . More precisely, we focus on measures of ownership concentration and let them exercise a non-linear impact on (the log of) output. That is, we assume τ_{it} to be governed by

$$\tau_{it} = e^{\alpha_0 + \mathbf{H}'_{it}\boldsymbol{\delta} + \eta_{it}}, \tag{3.24}$$

where $\mathbf{H}_{it}$ is a (column) vector containing ownership-related variables[10] and relevant controls.

Thus, our estimating equation reads as

$$y_{it} = \alpha_0 + \mathbf{H}'_{it}\boldsymbol{\delta} + \alpha l_{it} + \beta k_{it} + \gamma m_{it} + \eta_{it}, \tag{3.25}$$

[10] Since we posit ownership distribution to impact performance in a non-linear fashion, throughout the analysis, we let $\mathbf{H}_{it}$ include proxies of ownership distribution and their square.

where the error term η_{it} is assumed to be equal to

$$\eta_{it} = \omega_{it} + \epsilon_{it}. \tag{3.26}$$

ω_{it} represents firm's unobserved (to the econometrician) heterogeneity and is allowed to be arbitrarily correlated with input factors and ownership variables. As in Assumption 1, we assume it to evolve according to the first-order Markov process specified in (3.19). ϵ_{it}, instead, is a pure error component assumed orthogonal to our explanatory variables.

We partly retain Assumption 2 in that we assume capital to evolve according to the process specified in (3.13). As for labour and intermediate inputs, instead, we follow Ackerberg et al. (2015) and assume that the demand for these inputs is subject to some form of rigidity and, as such, none of them can be immediately varied by the firm. In particular, the process specified in (3.13) implies that capital at time t, namely k_{it}, is chosen in period $t-1$ and, as such, it belongs to the information set in period $t-1$ (i.e. $k_{it} \in I_{t-1}$). While being more "variable" than capital, both labour and intermediate inputs in period t, namely l_{it} and m_{it}, are allowed to have dynamic implications in that they are assumed to be chosen in period $t-b$ with $0 < b < 1$. That is to say that hiring and firing costs along with the costs arising from the the modification of existing supply agreements become part of the decision problem in (3.11) and turn to have an impact on both current and future profits. While not being a proper input factor, our econometric specification recognizes that ownership structure may well have a possibly non-linear impact on a firm's TFP. In dealing with ownership-related variables, we attach to them the same timing assumption used for the capital input. Indeed, ownership structure is commonly agreed to be a highly persistent firm characteristic, and in Italy its persistence appears to be even stronger than usual (Bianchi and Bianco (2008)) due to the low stock market liquidity and to the absence of a market for corporate control. By recognizing that ownership is as slow as our fixed input factor to vary over time, we not only address the potential endogeneity of the ownership structure but also make allowance for its highly persistent nature. Note that the timing assumptions depicted above prevent us from retrieving estimates of any of our parameters of interest from the first stage of a standard procedure à la Olley and Pakes (1996).

We replace Assumption 3 with Assumption 3b.

Assmption 3b: ω_{it} is the only scalar unobservable and impacts firm's decision on intermediate inputs through

$$m_{it} = f_t(k_{it}, \omega_{it}). \tag{3.27}$$

Following Levinsohn and Petrin (2003), (3.27) lets intermediate inputs proxy for unobserved heterogeneity. Hence, we depart from the original approach in Olley and Pakes (1996) where investment is used to control for ω_{it}. Substituting investment with the demand for intermediate inputs has a clear data-driven advantage. Due to the invertibility condition described below, a proxy variable is only valid if it is strictly greater than zero. Firms in our data always report positive level of intermediate inputs while the same does not hold true for investment that often exhibits a lumpy behavior. Thus, if we had to use investment we should have dropped a considerable amount of observations from our sample. A second advantage of using intermediate inputs is related to the potential non-convex nature of investment adjustment costs.[11] As argued by Levinsohn and Petrin (2003), non-convex adjustment costs may generate kinks in the investment function which are likely to undermine the responsiveness of investment to productivity shocks. We believe intermediate inputs to be less costly to adjust than investment and to respond more fully to productivity shocks.

Finally, to guarantee invertibility of $f_t(\cdot)$ we retain Assumption 4 and let $f_t(\cdot)$ be monotonically increasing in ω_{it}. The necessity of imposing this last assumption arises from the fact that we do not explicitly solve the dynamic programming problem in (3.11).

3.4.1 A 2-stage estimation procedure

Estimating (3.25) without taking into account the potential endogeneity arising from ω_{it} is likely to produce biased estimates. In the spirit of Olley and Pakes (1996), we develop the following 2−stage algorithm. Under ω_{it} being the only scalar unobservable and appealing to the monotonicity of $f_t(\cdot)$, we can invert (3.27) and obtain

$$\omega_{it} = f_t^{-1}(k_{it}, m_{it}). \tag{3.28}$$

[11] Evidence of non-convex adjustment costs of investment is reported, for instance, in Doms and Dunne (1998) and Attanasio et al. (2000)

Plugging (3.28) into (3.25) gives us

$$y_{it} = \alpha_0 + \mathbf{H}'_{it}\boldsymbol{\delta} + \alpha l_{it} + \beta k_{it} + \gamma m_{it} + f_t^{-1}(k_{it}, m_{it}) + \epsilon_{it}. \tag{3.29}$$

None of the parameters of interest can be estimated directly from (3.29). On the one side, elasticities of capital and intermediate inputs turn to be collinear with the non-parametric term $f_t^{-1}(k_{it}, m_{it})$, being $f_t^{-1}(k_{it}, m_{it})$ unconstrained. On the other side, the endogenous nature of labour and ownership-related variables makes them potentially correlated with the predictable component of ω_{it} which is subsumed into $f_t^{-1}(k_{it}, m_{it})$.

First stage

Equation (3.29) can be rewritten in the form of the following semiparametric partially linear regression

$$y_{it} = \mathbf{H}'_{it}\boldsymbol{\delta} + \alpha l_{it} + \phi_{it}(k_{it}, m_{it}) + \epsilon_{it}, \tag{3.30}$$

where the composite non-parametric term $\phi_{it}(k_{it}, m_{it})$ is given by

$$\phi_{it}(k_{it}, m_{it}) = \alpha_0 + \beta k_{it} + \gamma m_{it} + f_t^{-1}(k_{it}, m_{it}), \tag{3.31}$$

with $f_t^{-1}(k_{it}, m_{it}) = \omega_{it}$.

For ease of exposition, let us impose $\mathbf{z}_{it} = [k_{it}, m_{it}]'$, $\mathbf{x}_{it} = [l_{it}, \mathbf{H}'_{it}]'$ and $\boldsymbol{\theta} =$

$[\alpha, \boldsymbol{\delta}']'$. Then, we can rewrite (3.30) as

$$y_{it} = \mathbf{x}_{it}'\boldsymbol{\theta} + \phi_{it}(\mathbf{z}_{it}) + \epsilon_{it}. \tag{3.32}$$

Following Robinson (1988), we apply the conditional expectation operator $E[\cdot \mid \mathbf{z}_{it}]$ to (3.32) so as to get

$$\begin{aligned} E[y_{it} \mid \mathbf{z}_{it}] &= E[\mathbf{x}_{it}'\boldsymbol{\theta} \mid \mathbf{z}_{it}] + E[\phi_{it}(\mathbf{z}_{it}) \mid \mathbf{z}_{it}] + E[\epsilon_{it} \mid \mathbf{z}_{it}] \\ &= E[\mathbf{x}_{it} \mid \mathbf{z}_{it}]'\boldsymbol{\theta} + \phi_{it}(\mathbf{z}_{it}), \end{aligned} \tag{3.33}$$

where the last equality is due to ϵ_{it} being orthogonal to the set of input factors $\mathbf{z}_{it}$.

Subtracting (3.33) from (3.32), we get rid of the non-parametric term and obtain

$$y_{it} - E[y_{it} \mid \mathbf{z}_{it}] = (\mathbf{x}_{it} - E[\mathbf{x}_{it} \mid \mathbf{z}_{it}])'\boldsymbol{\theta} + \epsilon_{it}. \tag{3.34}$$

which is equivalent to the double residual regression

$$\epsilon_{it}^{y} = \boldsymbol{\epsilon}_{it}^{\mathbf{x}\prime}\boldsymbol{\theta} + \epsilon_{it}, \tag{3.35}$$

where

$$\epsilon_{it}^{y} = y_{it} - E[y_{it} \mid \mathbf{z}_{it}] \tag{3.36}$$

and

$$\boldsymbol{\epsilon}_{it}^{\mathbf{x}} = (\mathbf{x}_{it} - E[\mathbf{x}_{it} \mid \mathbf{z}_{it}]) \tag{3.37}$$

are the conditional expectations errors from the regression of y_{it} on $\mathbf{z}_{it}$ and of $\mathbf{x}_{it}$ on $\mathbf{z}_{it}$, respectively.

This transformed equation immediately suggests an infeasible least square estimator for $\boldsymbol{\theta}$ of the form

$$\widetilde{\boldsymbol{\theta}} = \left(\sum_{i=1}^{N} \sum_{t_i=1}^{T_i} \boldsymbol{\epsilon}_{it}^{\mathbf{x}} \boldsymbol{\epsilon}_{it}^{\mathbf{x}\prime} \right)^{-1} \left(\sum_{i=1}^{N} \sum_{t_i=1}^{T_i} \boldsymbol{\epsilon}_{it}^{\mathbf{x}} \epsilon_{it}^{y} \right). \tag{3.38}$$

In order to make the estimator in (3.38) feasible, we follow Robinson (1988) and first estimate the conditional expectations of y_{it} and $\mathbf{x}_{it}$ given $\mathbf{z}_{it}$ by Nadaraya–Watson (henceforth, NW) regression where cross-validation is used for bandwidth selection. Using these estimates, we obtain the regression residuals

$$\widehat{\epsilon_{it}^{y}} = y_{it} - E\widehat{[y_{it} \mid \mathbf{z}_{it}]} \tag{3.39}$$

and

$$\widehat{\boldsymbol{\epsilon}_{it}^{\mathbf{x}}} = (\mathbf{x}_{it} - E\widehat{[\mathbf{x}_{it} \mid \mathbf{z}_{it}]}) \tag{3.40}$$

which lead to the following feasible estimator for $\boldsymbol{\theta}$

$$\widehat{\boldsymbol{\theta}} = \left(\sum_{i=1}^{N} \sum_{t_i=1}^{T_i} \widehat{\boldsymbol{\epsilon}^{\mathbf{x}}_{it}} \widehat{\boldsymbol{\epsilon}^{\mathbf{x}}_{it}}' \right)^{-1} \left(\sum_{i=1}^{N} \sum_{t_i=1}^{T_i} \widehat{\boldsymbol{\epsilon}^{\mathbf{x}}_{it}} \widehat{\epsilon^{y}_{it}} \right). \tag{3.41}$$

Finally, we turn to the non-parametric component $\phi_{it}(\mathbf{z}_{it})$ which is estimated by the following NW regression

$$y_{it} - \mathbf{x}'_{it}\widehat{\boldsymbol{\theta}} = \phi_{it}(\mathbf{z}_{it}) + \epsilon_{it} \tag{3.42}$$

where again cross-validation is used for bandwidth selection.[12] Retrieving $\widehat{\phi}_{it}(\mathbf{z}_{it})$ from (3.42) allows us to deploy the second stage of our algorithm where all parameters of interest are finally estimated.

Second stage

In the second stage we are concerned with isolating the composite error term $\eta_{it} = \omega_{it} + \epsilon_{it}$. The assumed first-order markovianity of ω_{it} allows us to write

$$\begin{aligned} \xi_{it} &= \omega_{it} - E[\omega_{it} \mid \omega_{it-1}] \\ &= \omega_{it} - g(\omega_{it-1}), \end{aligned} \tag{3.43}$$

whose conditional expectation with respect to the information set at $t-1$, namely I_{t-1}, equals zero.

[12] As pointed out by Robinson (1988), since $\widehat{\boldsymbol{\theta}}$ converges at a rate $n^{-1/2}$ which is faster than a non-parametric rate, we can pretend $\boldsymbol{\theta}$ to be known and do non-parametric regression of $y_{it} - \mathbf{x}'_{it}\widehat{\boldsymbol{\theta}}$ on $\mathbf{z}_{it}$.

As for the pure innovation ϵ_{it}, instead, we have

$$\epsilon_{it} = y_{it} - \mathbf{x}_{it}'\boldsymbol{\theta} - \phi_{it}(\mathbf{z}_{it}). \tag{3.44}$$

As a result, the moment condition to be exploited here is given by

$$\begin{aligned} E[\xi_{it} + \epsilon_{it} \mid I_{t-1}] = & \\ & E[y_{it} - \alpha_0 - \mathbf{x}_{it}'\boldsymbol{\theta} - \beta k_{it} - \gamma m_{it} - \\ & g(\phi_{it-1}(\mathbf{z}_{it-1}) - \alpha_0 - \beta k_{it-1} - \gamma m_{it-1}) \mid I_{t-1}] = \\ & = 0, \end{aligned} \tag{3.45}$$

where the argument of $g(\cdot)$ is due to

$$\omega_{it-1} = \phi_{it-1}(\mathbf{z}_{it-1}) - \alpha_0 - \beta k_{it-1} - \gamma m_{it-1}. \tag{3.46}$$

By substituting into (3.46) for the first-stage estimate $\widehat{\phi}_{it-1}(\mathbf{z}_{it-1})$, condition in (3.45) becomes

$$\begin{aligned} E[\xi_{it} + \epsilon_{it} \mid I_{t-1}] = & \\ & E[y_{it} - \alpha_0 - \mathbf{x}_{it}'\boldsymbol{\theta} - \beta k_{it} - \gamma m_{it} - \\ & g(\widehat{\phi}_{it-1}(\mathbf{z}_{it-1}) - \beta k_{it-1} - \gamma m_{it-1}) \mid I_{t-1}] = \\ & = 0. \end{aligned} \tag{3.47}$$

Then, letting $\mathbf{v}_{it-1} = [\mathbf{H}_{it-1}', l_{it-1}, k_{it}, m_{it-1}]'$ denote the full set of explanatory variables[13] belonging to the information set at time $t-1$ (i.e. $\mathbf{v}_{it-1} \in I_{it-1}$), (3.47) allows us to derive the following criterion function

$$Q(\boldsymbol{\psi}) = \sum_{i=1}^{N} \sum_{t_i=2}^{T_i} \left(\left(\xi_{it_i}(\boldsymbol{\psi}) + \epsilon_{it_i}(\boldsymbol{\psi}) \right) \mathbf{v}_{it_i-1} \right)^2, \tag{3.48}$$

where

$$\boldsymbol{\psi} = [\boldsymbol{\delta}, \alpha, \beta, \gamma]. \tag{3.49}$$

Finally, our parameters of interest are estimated as

$$\boldsymbol{\psi}^* = \text{argmin } Q(\boldsymbol{\psi}). \tag{3.50}$$

Throughout the minimization of $Q(\boldsymbol{\psi})$, we treat the deterministic component of ω_{it} (namely, $g(\omega_{it-1})$) non-parametrically and estimate it by NW regression. In particular, at each iteration of our minimization routine, we compute $g(\omega_{it-1})$ through a non-parametric regression of $\omega_{it-1} = \widehat{\phi}_{it-1}(\mathbf{z}_{it-1}) - \beta k_{it-1} - \gamma m_{it-1}$ on $\omega_{it} = \widehat{\phi}_{it}(\mathbf{z}_{it}) - \beta k_{it} - \gamma m_{it}$. This prevents us from obtaining an estimate of the intercept α_0. Given the unknown asymptotic distribution of the estimator in (3.50), standard errors are computed using $1,000$ bootstrapping replications.

To our knowledge, our econometric framework represents the first attempt to incorporate variables related to corporate governance within a structural ap-

[13]Note that $\mathbf{H}_{it-1}$ contains ownership-related variables recorded at time t since we apply to them the same timing assumptions adopted for capital.

proach to productivity estimation. Moreover, the assumptions imposed allow us to explicitly model unobserved firms heterogeneity which, thanks also to the non-parametric tools employed, is let evolve according to a much more flexible process than the one that standard dynamic panel techniques can accommodate.

3.5 Dataset & Variables Description

The econometric model described above is applied to a sample of Italian listed firms that operate in the manufacturing sector. We collect the financial data from the ORBIS platform provided by Bureau Van Dijk. Such platform manages a very rich dataset containing financial information on firms operating in different industries across 230 countries and has extensively been used in previous studies (e.g., Tian and Twite (2011) and Bloom et al. (2010)). For what concerns information on the firms' ownership structure, instead, we resort to the public dataset made available by the Italian Securities Commission (CONSOB).[14] Indeed, Italian law sets forth a disclosure requirement according to which anyone holding more than $k = 2\%$ of the voting rights[15] in an Italian listed company shall disclose her participation by notifying CONSOB and the issuer. Thus, the CONSOB dataset provides historical information on the shareholders holding at least k percent of the shares of a company for all companies listed on the national stock exchange. In particular, for each direct blockholder the dataset reports over

[14] Information regarding ownership is available at: http://www.consob.it/web/area-pubblica/quotate

[15] Precisely, the threshold is $k = 2\%$ for firms with a market capitalization greater or equal to 500 million euros, and $k = 5\%$ for small-mid capitalization firms. The threshold relative to large capitalization firms has been raised to $k = 3\%$ in 2016.

time the total participation and the percentage of shares that do not have voting rights. Additionally, the dataset contains the participations (total percentage and the fraction with no voting rights) of those who are at the top of the control chain (called ”Dichiaranti”). In our analysis we are interested in the ownership structure as an internal corporate governance mechanism. Under this perspective the distribution of ownership shapes the formation of control and the incentives of those who are in control. Accordingly, we use the data concerning the ”Dichiaranti” (ultimate owners) that provide a clearer picture about the distribution of control.

By matching the two sources of information mentioned above, we obtain an unbalanced panel data sample of 116 firms and 673 firm-year observations from 1998-2015. Economic variables of interest are appropriately selected and modified following Gal (2013) who describes how to construct firm-level total factor productivity measures using ORBIS. Information about the distribution of ownership among ultimate owners, instead, is used to construct measures of concentration of votes, of cash-flow rights, and of controlling power.

3.5.1 Economic Variables

Given the production function assumed in our econometric specification, we first have to select a reasonable dependent variable that could proxy for total output. In the productivity literature it is common to use a variable measuring revenues to represent output, since a physical measure of output in terms of number of units produced is normally unavailable. As suggested by Gal (2013), we use the

ORBIS variable ”Operating Turnover” as a measure of gross output.[16]

Coming to input factors, we follow Gal (2013) and use the reported number of full-time employees to measure labour (ORBIS variable “Employees”). Measuring the labour input is problematic. One should have information on the number of hours worked and on the different types and characteristics of employees. Unfortunately, the ORBIS database does not contain this information. The only available variables related to labour input are the number of full-time employees (that we use) and a variable representing the total cost of labour. However, as noted by Gal (2013), the variable representing total labor costs has a smaller coverage and may not properly reflect the quality and intensity of labor because is directly influenced by the regulatory environment. We therefore choose to follow Gal (2013) and use the ORBIS variable “Employees”.

Intermediate inputs, instead, are measured by using the value of material costs (ORBIS variable “Material Costs”). In the ORBIS database intermediate inputs are not differentiated across materials, energy and purchased services. As stressed by Gal (2013) the unavailability of a richer set of variables measuring intermediate inputs does not allow the estimation of a more detailed production function.

As for the capital stock, we calculate it by using the Perpetual Inventory Method. According to this method, the capital stock K_{it} in firm i at time t

[16]Note that, if we had complied with much of previous studies and used value added as a measure of production, we could have not gained any insights on the impact of intermediate inputs on productivity (see, e.g., Foster et al. (2001)) since value added is, by definition, equal to the difference between sales and intermediate inputs.

is defined as

$$K_{it} = K_{it-1}(1 - \delta_{it}) + I_{it}, \tag{3.51}$$

where δ_{it} and I_{it} denote the depreciation rate and the level of investment, respectively. To make (3.51) applicable, let us define the level of investment as

$$I_{it} = K_{it}^{BV} - K_{it-1}^{BV} + DEPR_{it}^{BV}, \tag{3.52}$$

where K_{it}^{BV} and $DEPR_{it}^{BV}$ denote the book value of fixed tangible assets and depreciation. Both K_{it}^{BV} and $DEPR_{it}^{BV}$ are retrieved from the ORBIS dataset. The depreciation rate, in turn, is given by

$$\delta_{it} = DEPR_{it}^{BV} / K_{it-1}^{BV}. \tag{3.53}$$

For the first year in which each firm is observed we impose $K_{i0} = K_{i0}^{BV}$. Then, applying (3.51) after having recovered the level of investment and the depreciation rate from (3.52) and (3.53), respectively, we construct the whole series of capital stocks for each firm in our sample. In the ORBIS database, capital goods are differentiated only to the extent of being tangible and intangible. There is no specification regarding the type of the asset. We follow Gal (2013) suggestion of using only data on total tangible fixed assets to avoid the difficulties in measuring and valuing intangibles which are poorly reported in ORBIS.

Excluding labour, all the input factors obtained thus far are expressed in thousands of euros. In order to compare values over time, we need to adjust for price changes and deflate these nominal variables through appropriate price indices. The ORBIS database does not provide firm-level price indices, thus only industry level deflators can be used. We use the 2-digit industry annual price deflators from the OECD STAN database to convert operating turnover, material costs and capital in real 2010 euros. The practice of measuring output using revenues deflated with industry-level price deflators rather than firm-level prices (due to data availability), has important implications. As underlined by Syverson (2011), neglecting price variability across firms implies that the estimated productivity will reflect more than just supply-side forces. In particular, within-industry price differences will be embodied in output and productivity measures. In markets where prices reflect idiosyncratic demand shifts, representing variation of firms' market power, then the estimated TFP will capture both technical efficiency and demand factors.

Finally, we log-transform all economic variables.[17] Summary statistics for these variables are reported in the first part of Table 3.1.

3.5.2 Ownership Variables

The main information about ownership retrieved from CONSOB pertain to the voting and non-voting shares held by the ultimate shareholders holding $k = 2\%$ or

[17]From now on we let l, k, m to denote respectively the log-transformation of the production inputs, that is $l = \ln L, k = \ln K$ and $m = \ln M$. Correspondingly, $y = \ln Y$.

more of the voting capital. Using the information about the identified blockholders, we can immediately obtain two basic measures of ownership concentration which are respectively given by the sum of the shares held by the blockholders of a firm at a given time (see, e.g., De Miguel et al. (2004) and Laeven and Levine (2008)) or by the number of blockholders themselves. While being simple, we recognize that these measures are prone to an erroneous representation of ownership concentration (see, e.g., Overland et al. (2012)). Consider, for instance, the case of a company with a majority blockholder (namely, a blockholder owning 50% or more of the shares) and two minor blockholders holding 10% and 20%, respectively. Then, an increase in the holding of either of the minority shareholders would qualify as an increase in ownership concentration if we measure it through the combined holding of blockholders while it effectively amounts to a decrease in ownership concentration since it produces a less uneven distribution of owners' stakes. To overcome the limitation outlined above, we rely on two additional measures of ownership concentration that take into account the interplay and relative position of blockholders.

The first one is represented by the Herfindahl index of the blockholders' participation. Thus, letting firm i having J blockholders at a given time t, the Herfindahl index of the blockholders' shares is defined as

$$H_{it} = \sum_{j=1}^{J} s_{ijt}^2, \tag{3.54}$$

where s_{ijt} denotes the share of the capital of firm i held by blockholder j in period t relative to the to the sum of the shares held by all blockholders. The index defined in (3.54) ranges from $\frac{1}{J}$ to 1. At the lower bound, the smallest level of ownership concentration is achieved since all blockholders hold the very same share of the capital. When a firm's capital is entirely held by a single blockholder, instead, H equals 1 and ownership concentration reaches its maximum. Computing the Herfindahl index using the total participation of each blockholder without distinguishing between voting and non-voting shares we obtain a measure of cash-flow rights' concentration (H Total), while the same index obtained using only voting shares provides a measure of voting rights concentration (H Voting). As noted by Hall and Tideman (1967), if some shares are shifted from a larger to a smaller blockholder, a company will experience a decrease in the level of ownership concentration and, in line with that, the Herfindhal index will be lower. Note that, under such circumstances, the index detects a change in the concentration level even if both the number of blockholders and the amount of shares that they collectively hold remain constant.

Our second measure of ownership concentration builds upon a game theoretic measure of power, namely the Banzhaf index (henceforth, BI) introduced in Banzhaf (1965). Identifying blockholders as the players of a yes-no voting game, BI measures the probability of individual players to be *critical*. A player is said to be critical if, with her vote, she can turn a given coalition from "loosing" to "winning", would she decide to join it.[18] Letting a blockholder B_j be a voter in

[18]Another commonly used measure of power is the so-called Shapley index which focuses on the probability of a player's vote to be *pivotal*. A vote is defined as pivotal if, by casting it,

a yes-no voting system, her total Banzhaf power, denoted by $TBP(B_j)$, is equal to the number of coalitions C satisying the following three conditions:

- $B_j \in C$,
- C is a winning coalition (i.e. the sum of the shares held by the blockholders belonging to C exceeds the majority requirement which is assumed equal to 50%),
- $C - \{B_j\}$ is not a winning coalition.

Computing the Banzhaf index of each of the shareholders of a given firm would require us to posses information on the shares held by each of them. Unfortunately, the CONSOB dataset only reports the shares held by those possessing 2% or more of the company and we classify them as blockholders. Let a firm have in total N shareholders with J of them being blockholders with a stake higher or equal to 2% (obviously, $J \leq N$). Then, the amount of capital in the hands of the $N - J$ shareholders is equal to $1 - S_{2\%}$, with $S_{2\%}$ being the combined shares of the firm's blockholders. In modelling this $1 - S_{2\%}$ fraction of the voting capital on which we do not have information, we assume it to be split in infinitesimal small parts among an infinite numbers of residual shareholders. Doing so, we achieve an ”oceanic” representation of the voting game (see, e.g., Milnor and Shapley (1978)). Dubey and Shapley (1979) show that the power indices for an oceanic game with J major

a player can turn a given coalition from ”loosing” to ”winning”. Thus, the Shapley index of a given player will be given by the number of times in which her vote is pivotal over the total number of possible voting sequences (i.e. the factorial of the number of players). Analyzing both the Shapley and the Banzhaf indeces of a sample of British companies, Leech (2002) finds that the Banzhaf index reflects much better the variations in the power of shareholders between companies.

players (namely, blockholders) with combined shares equal to S^* and a majority requirement of $q = 50\%$ are the same as for a finite game consisting only of the J major players and a modified majority requirement of $q^* = (50 - (1 - S^*)/2)\%$. Thus, using this lower majority requirement, we can finally compute the Banzhaf power of every blockholder for each of the firm-year observations in our sample. The index is obtained considering only the voting shares of each owner. The computed Banzhaf power measures allow us to construct the normalized Banzhaf power index, denoted by $BPI(B_j)$, of each blockholder B_j which turns to be given by

$$BPI(B_j) = \frac{TBP(B_j)}{\sum_{n=1}^{J} TBP(B_n)}. \tag{3.55}$$

By definition, the index defined in (3.55) ranges from 0 to 1 and is increasing in the amount of voting power effectively held by a given blockholder. Once we obtain the Banzhaf power indeces of all blockholders at each firm-year observation, we use them to construct another measure of ownership concentration in terms of control power as follows

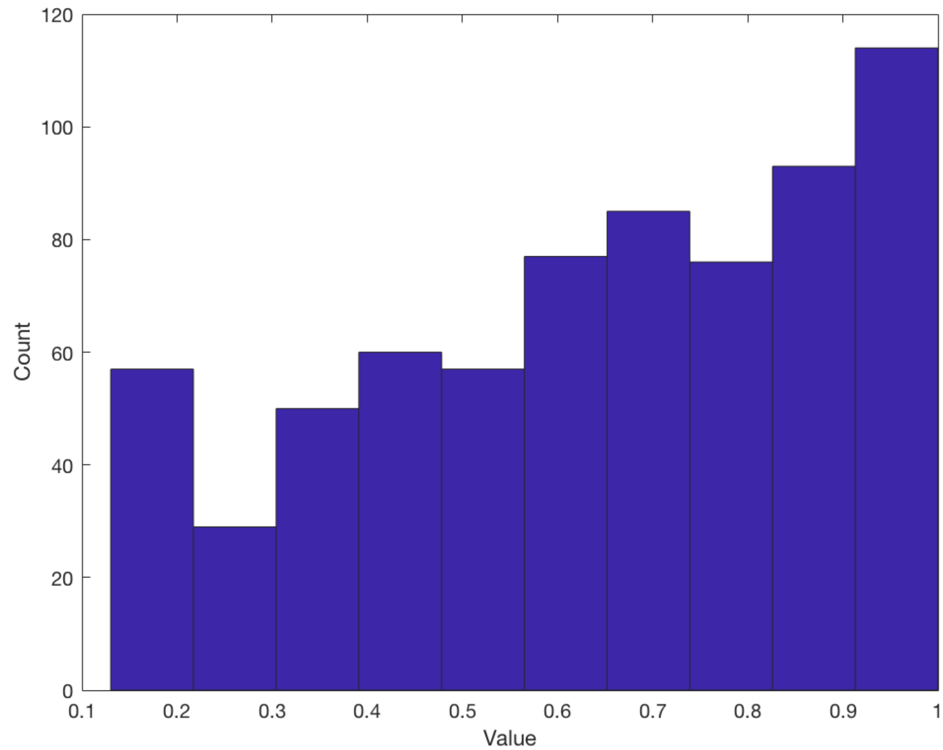
$$H_{it}^{banzhaf} = \sum_{j=1}^{J_{it}} (BPI(B_{jit}))^2, \tag{3.56}$$

where we restored the dependence on i and t and J_{it} represents the total number of blockholders of firm i in period t.

Histograms in Figures 3.1 and 3.2 provide evidence of the substantially different distributions of concentration implied by H index respectively obtained considering total participations and $H^{banzhaf}$. In particular, when focusing on the blockholders' probability of being critical in a yes-no voting game the control power tends to be much more concentrated since around 81% of our firm-year observations clusters at a value of $H^{banzhaf}$ equal 1. That is to say that for all these observations there is a single critical blockholder. The discrepancy between H and $H^{banzhaf}$ casts some doubts on the adequacy of shares *per se* to proxy for relative blockholders' power.

Finally, we introduce an index measuring the wedge between control rights and cash-flow rights. Generally speaking, control rights refer to an owner's ability to influence the business activities, while the cash-flow rights refer to the portion of the firm's profits to which an owner is entitled. Cash-flow rights can be unambiguously measured using the total fraction of shares held by each shareholder. More complex is the measure of control rights. One way is to equate control rights with voting rights. Under this approach a wedge between control rights and cash-flow rights may arise only by issuing classes of shares that differ in terms of their relative proportion of voting rights and dividend payments or allowing an owner to exercise control through a chain of other firms (pyramids). Alternatively, as suggested by Edwards and Weichenrieder (2009) control rights can be measured in terms of the ability of a shareholder to determine the outcome of a vote, given the fraction of votes required to win and the overall distribution of voting rights. The advantage of this approach is that control rights are more strictly related to

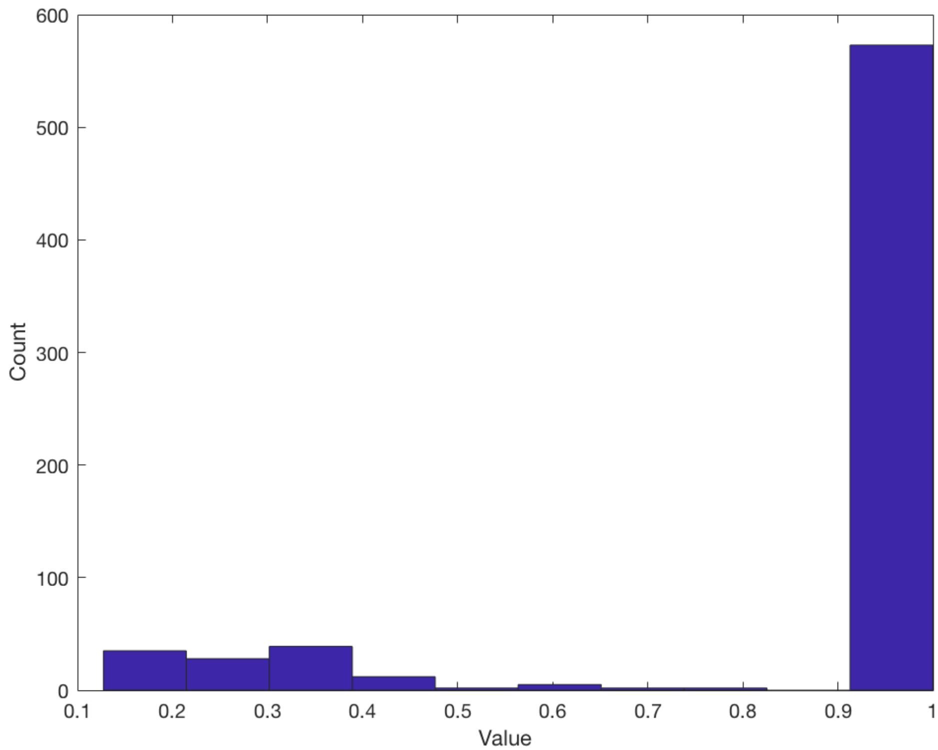
Figure 3.1: Histogram of Total H



the effective ability of owners to influence firm's decisions and that it allows to capture differences between control and cash-flow rights even when all shares have the same voting rights and no owner exerts control through a pyramid. We follow the approach of Edwards and Weichenrieder (2009) and we measure control rights in terms of voting power. Precisely, we introduce CR_{it} given by

$$CR_{it} = \frac{H_{it}^{banzhaf} - H_{it}}{H_{it}^{banzhaf}}. \tag{3.57}$$

Figure 3.2: Histogram of $H^{banzhaf}$



The ratio is a measure, for each firm and year, of the relative difference between the concentration of power and the concentration of cash-flows due to the ownership structure. Note that for computing H we use both blockholders' voting and non-voting shares thereby capturing the distribution of cash flow rights. As for $H^{banzhaf}$, instead, we only rely on voting shares so as to get a more precise representation of control rights. The ratio in (3.57) gauges the extent to which the concentration of control rights deviates from the concentration of cash-flow rights.

In addition to the variables described above, we include in our analysis a dummy variable, $Majority_{it}$, which takes value 1 if the largest shareholder owns 50% or more of a firm's shares at a given time and zero otherwise. This dummy variable should capture the differences arising between the case where control is exercised unilaterally and the case where control is obtained forming a coalition of shareholders. Since, a shareholder can have complete power on a firm even without holding the majority of the shares, we build a second dummy, $Control_{it}$, equal to 1 if the $H^{banzhaf} = 1$ and 0 otherwise. While the previous dummy is based on the identification of cases where one blockholder has formally the control of the firm (*de iure* control), this second dummy, more generally, differentiate the case where there exists one single blockholder having *de facto* control power over the firm without specifying a quota (50%) for obtaining total control.

In the second and third part of Table 3.1 we report summary statistics for ownership-related variables and the controls that we use in our empirical analysis. In Table 3.2 we describe the composition of our sample with respect to the case where there is a single shareholder having complete control of a company and cases where control is shared among different shareholders. In Table 3.3 we report the transition probabilities of the dummy variables detecting the existence of a single controlling blockholder (both de iure and de facto).

The sample of Italian listed companies is characterized by a high degree of ownership concentration. In about 62% of the total number of observation there is a single controlling blockholder holding more than 50% of the company's shares. There are small differences in the statistics computed considering total shares and voting shares, suggesting that there is not a widespread use of dual-class shares. On the contrary, in our sample, there is a substantial gap between concentration of control power and concentration of voting rights. Strikingly, about 81% of the sample observations correspond to cases where de facto a single blockoholder has complete control over the firm. The ownership structure is highly persistent: as Table 3.3 shows the probabilities of a change over time of the categorical variables indicating whether there is a controlling blockholder are very low (the diagonal elements of the transition probabilities matrices are greater than 93 percent).

3.6 Results

The last column of Tables 3.4, 3.5, and 3.6 report the parameters estimated using our econometric model where the ownership structure is respectively represented by $H^{banzhaf}$, $H\ Voting$, and CR . The first and second columns respectively report the corresponding OLS and Fixed Effects estimates. In the third column, estimates are obtained by exploiting a standard first-difference GMM estimator (see, e.g., Blundell and Bond (2000)). Throughout all the specifications, we control for firm's leverage (as proxied by the ratio between total long term liabilities and total assets) and (the log of) firm's age. Additionally, in all settings, we include year fixed effects.[19]

In order to implement the first-difference GMM estimator, we first adapt equation (3.6) to our framework. That is, we regress the log of gross output on its lagged value and on the current and lagged values of the log of inputs, ownership-

[19]For ease of exposition, year fixed effects are not reported in the Tables with results.

related variables and relevant controls. We construct suitable orthogonality conditions extending the condition in (3.9) to all input factors, ownership-related variables and leverage. Hence, our moment conditions require the error term to be orthogonal only to variations of these variables between consecutive periods. That is to say we allow for the lagged and current level of inputs, ownership characteristics and leverage to be correlated with firm-fixed effect η_i thereby addressing at least to some extent the potential endogeneity of these variables. As for age and year fixed effects, instead, we hold them to be truly exogenous and, as such, our moment conditions require the error term to be orthogonal to both their current and lagged values.

Under our econometric model, leverage is again treated as an endogenous variable. Thus, in order to allow it to correlate with unobserved heterogeneity ω_{it}, we minimize the criterion function in (3.48) by including the lagged value of leverage into our vector of explanatory variables $\mathbf{v}_{it-1}$. Moreover, the current levels of age and year fixed effects are included in $\mathbf{v}_{it-1}$, since, being deterministic, these variables can be assumed to belong to the information set in period $t-1$.

Coming to the interpretation of estimated input factors elasticities, we document a decrease in the labour coefficient when comparing OLS estimates with any of the other methods that explicitly address the endogeneity issue (e.g., in Table 3.6 its magnitude ranges between around 46 and 56 percent). This change confirms the positive OLS bias that we should expect in the presence of correlation between labour usage and firm-specific differences in productivity.

The capital coefficient, in comparison with the OLS estimates, is lower under both the fixed effects and the first difference GMM estimators and higher in the remaining case. For instance, in Table 3.6 our model predicts a level of capital elasticity which exceeds the OLS estimate by around 120 per cent. This is in line with the positive change documented in Olley and Pakes (1996) and several others. As argued by Levinsohn and Petrin (2003), the OLS estimate on capital is likely to be biased downwards if capital positively covaries with variable inputs but correlates much more weakly than variable inputs with the productivity shock.

With the exception of the fixed effect estimator, all the other specifications do not imply considerable differences in the estimated elasticity of intermediate inputs. Coherently with what we found for the other variable input, namely labour, we document an attenuation of the expected positive OLS bias in all specifications with the exception of the difference GMM one. However, the unrealistically low value obtained under the fixed effects specification, in our opinion, suggests that imposing unobserved heterogeneity to be constant over time is too restrictive.

In all specifications, we include both the measure of ownership concentration and its square as explanatory variables to detect possible non-linearities. Excluding the fixed effects estimator, the remaining three models agree in the signs of the effects that the different proxies of blockholders' concentration and their square have on firm productivity. They depict an inverted U-shaped relation between the degree of ownership concentration and performance. However, such relation is found to be significant only in our model which is the one that accounts for a flexible time-varying process for firms' heterogeneity. The difference GMM estimates exhibit the highest standard errors. This suggests that instrumenting a highly persistent variable, such as ownership concentration, using its variations between consecutive periods is likely to produce poor estimates.

Focusing on the effect of increases of concentration of power, we find that higher levels of concentration have a positive effect up to a threshold level, $H^{banzhaf*} = 0.38$, and a negative effect afterwards. In our sample about 15 percent of the observations assumes values not greater than the threshold and the identified turning point is much lower than the average value equal to 0.869. We detect a similar behavior looking at the relationship between concentration of voting rights and firm productivity. An increase in the level of concentration has a positive effect up to a threshold level, $H\ Voting = 0.49$, and a negative effect afterwards. About 30 percent of the sample observations are below the estimated turning point. The common pattern identified can be interpreted as a result of the interaction between the monitoring dimension and the shareholder conflict dimension related to increases of ownership concentration. Indeed, as power gets concentrated in the hands of large shareholders, they find themselves more incentivized to invest in acquiring information and in monitoring the activity of managers (see, e.g., Shleifer and Vishny (1986)). On the contrary, when the ownership structure is very diffuse, a free-rider problem is likely to arise: no shareholder may have incentive to engage in costly monitoring activities or to exercise her ”voice”; this, in turn, would lead to weak control on the part of shareholders on managers who may well deviate from a profit-maximizing behavior (see, e.g., Grossman and Hart (1980)). The increase in ownership concentration will have a positive effect up to the point where the monitoring dimension prevails. But, as ownership concentration reaches a certain threshold giving additional power to large shareholders starts being detrimental to firm productivity as a consequence of the shareholder conflict dimension. Indeed, when the concentration of ownership is such to guarantee that the largest shareholders have the incentives to be active blockholders and monitor business activities, a further increase in concentration may be detrimental because it exhacerbates the conflicts between controlling and non-controlling shareholders. While controlling blockholders may be tempted to exercise their power for the purpose of extracting private benefits of control rather than use it for engaging in value enhancing activities for all shareholders, the existence of sufficiently large non-controlling blockholders that put pressure on the controlling block can limit their expropriation.

In our sample, characterized by high levels of ownership concentration, most of the observations lie on the downward sloping part of the detected inverted U-shaped relationship (respectively about 81 percent of the observations when we measure ownership concentration in terms of power and about 70 when we measure it in terms of voting rights). That is, as expected, in settings where the ownership is highly concentrated the more prevalent dimension is the one related to the conflict of interests among shareholders. Interestingly, comparing de facto and de iure measure of ownership concentration, we find that the concentration of control power is much higher than the concentration of voting rights and, accordingly, the downward sloping part of the detected inverted U-shaped relationship is more prominent when we measure ownership concentration in terms of power (81 percent of the observations versus 70 percent).

We find a similar pattern when we look at the relative difference between the concentration of control power and the concentration of cash-flow rights measured by CR_{it}. Our model identifies an inverted U-shaped relationship between CR_{it} and firm productivity. The turning point of CR_{it} is about 0.36, higher than the average value of this ratio for the firms in our sample (namely, 0.25). Marginal effects are positive for roughly 68% of our firm-year observations. In the remaining 32% of the sample, where CR_{it} exceeds its predicted optimal level, marginal effects are negative. Figure 3.3 scatters the predicted marginal effects of the concentration ratio CR_{it} on the log of production for the firms in our sample. Marginal effects are given by

$$\begin{aligned} \frac{\partial \widehat{E}[y_{it} \mid CR_{it}]}{\partial CR_{it}} &= \widehat{\delta_1} + 2\widehat{\delta_2} CR_{it} \\ &= .890 - 2.482 CR_{it}, \end{aligned} \tag{3.58}$$

where $\widehat{\delta_1}$ and $\widehat{\delta_2}$ denote the estimated coefficients of CR_{it} and its square, respectively. To verify if, in our sample, an inverted U shaped relationship between firm productivity and CR_{it} exists, we compute the marginal effects at the minimum and at the maximum of CR_{it}. These marginal effects respectively amount to 1.88 and -0.88. In order to check their significance, we compute their standard errors as follows

$$Var\left(\frac{\partial \widehat{E}[y_{it} \mid CR_{it}]}{\partial CR_{it}}\right) = Var(\widehat{\delta_1}) + 4CR_{it}^2 Var(\widehat{\delta_2}) + 4CR_{it} Cov(\widehat{\delta_1}, \widehat{\delta_2}). \tag{3.59}$$

Evaluating (3.59) at $CR_{it} = min(CR_{it})$ and at $CR_{it} = max(CR_{it})$, standard errors are respectively equal to around 0.06 and 0.17. As a result, marginal effects at the extremes are both found to be significantly different from zero at all usual significance levels.[20]

The detected non-monotonic relationship between CR_{it} and firm productivity, as in the previously discussed cases of concentration of power and of voting rights, may be explained as the result of the interplay between the monitoring dimension and the shareholder conflict dimension. Indeed, CR_{it} is a measure of the wedge between the concentration of control power and the concentration of cash-flow rights. Cash-flow rights shape the shareholders incentives to intervene and monitor business activities. Additionally, they determine the alignment of controlling shareholders to the interests of non-controlling shareholders. Control power reflects the actual possibility of shareholders to influence business activities.

When CR_{it} is negative, the ownership distribution is such that concentration of control power is lower than the concentration of cash-flow rights. This certainly happen when when firms issue non-voting shares. In such cases an increase of CR_{it} will have a positive effect since it will produce a better alignment between the incentives to intervene and the power to do it.

CR_{it} will be equal to zero in all cases where there is a single majority blockholder. An increase of CR_{it}, corresponding to situations where $H\ Total$ decreases, will have a positive effect because the presence of additional blockholders monitoring the majority blockholder will have a positive alignment effect. In general, whenever there exists one blockholder having complete power over the

[20] In unreported analysis, we verify that the same also holds true for both $H^{banzhaf}$ and $H\ Voting$.

firm ($H^{banzhaf} = 1$) and the level of concentration of cash-flows is very high ($CR_{it} \leq 0.36$), an increase in the relative difference between control concentration and cash-flow rights concentration has a positive effect. Indeed, in all the previous cases, there exists one blockholder having the power to control managers to avoid self-interested behavior. A less concentrated ownership structure with non-controlling blockholders big enough to have the incentives to engage in costly monitoring of the controlling blockholder will reduce the conflict of interests among shareholders.

When the wedge between concentration of power and concentration of cash flows exceeds the estimated threshold ($CR_{it} > 0.36$), a further increase of CR_{it} has a negative effect on firm productivity. In all these cases, the ownership distribution is such that there is a substantial divergence between control power and cash-flow rights. For instance, that happens in all the cases where there is de facto one single controlling shareholder but the distribution of cash-flows is highly dispersed. These are situations where the conflict of interests among shareholders is the most relevant issue. An increase in the divergence between control power and cash-flow rights will be detrimental because will exacerbate the incentives of the controlling blockholder to expropriate non-controlling shareholders. This result is in line with the findings of Maury and Pajuste (2005). They find that the ratio of voting rights to cash flow rights of the largest shareholder has a significant negative effect on firm value measured as Tobin Q. They argue that the increase of the divergence between voting rights and cash flow rights produces an entrenchment effect: the largest shareholder has more incentives to extract private benefits not shared by all shareholders.

Our analysis shows that the relationship between control rights and cash-flow rights is more complex. There are some important methodological differences that allow us to discover the described non-monotonic relationship. First, we define control in terms of power overcoming the restrictive definition of control rights as voting rights. Second, we look at the overall distribution of control rights and cash-flow rights. Thus, our measure of divergence between control and cash-flow rights considers all blockholders and not only the largest shareholder.

The effect related to the presence of a single blockholder having complete control of the firm, measured by the dummy variable *Control*, is not statistically significant even though the magnitude of this effect varies considerably among the proposed estimators. The effect of firm's leverage on production tends to be positive and is found to be statistically significant under the fixed effects and the difference GMM estimators. With the exception of the fixed effect estimator, firm's age is found to have a negative effect on productivity even though such effect tends to be statistically insignificant.

3.6.1 Robustness checks

In unreported regressions, we repeat our analysis by using both the sum of shares collectively held by blockholders and the number of blockholders as measures of ownership concentration. While retrieving similar input elasticities, in neither of these instances we obtain evidence of a significant impact of ownership characteristics on firm's performance. All in all, our results confirm that the distribution of controlling power among relevant shareholders is a complex phenomenon which is hard to be appropriately represented by simple measures that merely rely either on the number of blockholders or on the sum of their relative shares. In particular, these simple measures cannot capture the differential impact related to the monitoring and shareholder conflict dimension associated to increases in ownership concentration.

Finally we verify one of the main assumption on which our model relies. As suggested by Levinsohn and Petrin (2003), we informally verify that our estimates are consistent with the assumption of intermediate inputs being stochastically increasing in the level of unobserved heterogeneity ω_{it}. To this end, we compute the predicted level of ω_{it} implied by our estimates[21] which reads as

$$\widehat{\omega}_{it} = \widehat{\phi}_{it}(\mathbf{z}_{it}) - \widehat{\beta}k_{it} - \widehat{\gamma}m_{it} \tag{3.60}$$

and plot it against the corresponding level of intermediate inputs and capital. Figure 3.4 shows that predicted productivity shocks are empirically consistent with our model in that they tend to increase in the usage of intermediate inputs, holding the capital level constant.

3.7 Conclusion

In this paper we shed some light on the impact of ownership concentration on firm performance using a sample of Italian listed manufacturing firms. Previous

[21] We use the estimated parameters reported in Table 3.3 but results do not differ considerably when using estimates in Tables 3.4 and 3.5.

literature on this topic has relied upon financial measures of firm performance (accounting profit ratios such as return on assets and return on equity or market value ratios such as Tobin Q), which are affected by accounting practices and investors' sentiment (Gedajlovic and Shapiro (1998), Maury and Pajuste (2005), Laeven and Levine (2008) and Konijn et al. (2011), Russino et al. (2019), to cite a few). In this paper we measure firm performance in terms of productivity.

To our knowledge, this is the first study relating ownership structure to firm productivity using structural approach techniques. Our choice of measuring firm performance in terms of productivity allows us to analyze the relationship between ownership concentration and firm value within a structural framework explicitly developed to deal with endogeneity issues. We extend the "proxy variable" approach introduced by Olley and Pakes (1996) to incorporate corporate governance variables representing the ownership structure. Our econometric setting allows to control for unobserved firm heterogeneity and for the endogeneity of both inputs and blockholders' ownership concentration.

Focusing on measures of ownership concentration that depend on the distribution of ownership among all blockholders and reflect both the distribution of control and the distribution of cash-flow rights, we show how the allocation of ownership affects managerial opportunism and the conflict of interests among blockholders.

We find that the relationship between ownership concentration and firm productivity is non-monotonic and assumes an inverted U-shaped form. We argue that our findings arise from the interaction between the monitoring dimension and the shareholder conflict dimension related to increases in ownership concentration. At low values of ownership concentration, an increase in concentration has a positive effect driven by the related improvement in the shareholders' incentives to engage in costly monitoring of managers. After a threshold level of concentration a further increase in ownership concentration will be detrimental because it exacerbates the conflict of interests between controlling and non-controlling shareholders.

Due to data constraints, we remain silent about the heterogeneity across blockholder types. Distinguishing, for instance, between institutional and family shareholders could provide a more comprehensive representation of shareholding structure. Such an extension might be easily incorporated within our model by future research. On a methodological ground, we are mainly concerned with addressing the endogeneity issue. Future research might try to explicitly account for two other relevant sources of bias, namely sample selection and measurement error.

Table 3.1: Summary statistics

	Mean	Std. Dev.	Min	Max
Economic variables:				
y	12.601	1.682	8.040	18.258
l	6.993	1.581	1.099	11.305
k	11.313	2.121	2.708	18.236
m	11.482	1.943	4.736	17.657
Ownership variables:				
SumSH Total	.669	.172	.113	1
SumSH Voting	.657	.168	.113	1
Number blockholders	3.996	2.218	1	12
H Total	.628	.253	.130	1
H Voting	.645	.257	.130	1
$H^{banzhaf}$	.869	.279	.127	1
CR	.255	.208	-.400	.713
Controls:				
leverage	.151	.128	0	.750
$\ln(age)$	3.586	.876	0	5.094

$N = 673$, Number of firms = 116. All economic variables are in logs.

Table 3.2: Sample composition

	0	1
Majority	37.59 %	62.41%
Control	18.87%	81.13%

$N = 673$, Number of firms = 116.

Table 3.3: Transition Probabilities

Majority	0	1
0	96.57%	3.43%
1	3.49%	96.51%
Control	0	1
0	93.40 %	6.60%
1	0.90%	99.10%

$N = 673$, Number of firms = 116.

Table 3.4: Parameter Estimates: $H^{banzhaf}$

	OLS	**FE**	**Diff-GMM**	**Our Model**
Constant	3.011			
	(.493)			
ρ			.816	
			(.152)	
Economic variables:				
l	.296	.132	.169	.126
	(.052)	(.054)	(.083)	(.014)
k	.165	.090	.056	.348
	(.059)	(.062)	(.061)	(.009)
m	.482	.282	.547	.428
	(.048)	(.068)	(.144)	(.010)
Ownership variables:				
$H^{banzhaf}$	.341	1.106	.830	.883
	(1.492)	(1.024)	(1.609)	(.087)
$(H^{banzhaf})^2$	-.204	-.810	-.647	-1.161
	(1.528)	(1.269)	(2.002)	(.063)
Control	.069	-.016	.098	.328
	(.391)	(.650)	(1.005)	(.140)
Controls:				
leverage	.235	.496	.928	.422
	(.381)	(.221)	(.409)	(.552)
age	-.073	.099	-.008	-.070
	(.045)	(.121)	(.030)	(.098)
Observations	673			
Number of firms	116			

Standard errors are reported in parentheses. OLS, FE and Diff-GMM standard errors are robust with clustering at the firm level. Under our model, standard errors are obtained through bootstrapping. All estimators control for year fixed effects.

Table 3.5: Parameter Estimates: $H\ Voting$

	OLS	**FE**	**Diff-GMM**	**Our Model**
Constant	3.154			
	(.434)			
ρ			.813	
			(.156)	
Economic variables:				
l	.295	.131	.161	.139
	(.051)	(.054)	(.085)	(.013)
k	.165	.090	.094	.358
	(.056)	(.057)	(.071)	(.006)
m	.478	.282	.489	.437
	(.048)	(.068)	(.138)	(.008)
Ownership variables:				
H Voting	.193	-.302	.378	1.016
	(.939)	(.671)	(1.219)	(.078)
(H Voting)^2	.110	.249	-.397	-1.044
	(.738)	(.535)	(.969)	(.066)
Control	-.027	-.009	.075	.270
	(.101)	(.114)	(.234)	(.369)
Controls:				
leverage	.275	.491	.857	.892
	(.371)	(.217)	(.389)	(1.821)
age	-.077	.095	-.006	-.022
	(.045)	(.125)	(.028)	(.147)
Observations	673			
Number of firms	116			

Standard errors are reported in parentheses. OLS, FE and Diff-GMM standard errors are robust with clustering at the firm level. Under our model, standard errors are obtained through bootstrapping. All estimators control for year fixed effects.

Table 3.6: Parameter Estimates: CR

	OLS	**FE**	**Diff-GMM**	**Our Model**
Constant	3.154			
	(.382)			
ρ			.815	
			(.155)	
Economic variables:				
l	.297	.131	.160	.130
	(.051)	(.054)	(.078)	(.014)
k	.164	.087	.084	.368
	(.060)	(.060)	(.075)	(.010)
m	.480	.283	.516	.439
	(.048)	(.067)	(.133)	(.012)
Ownership variables:				
CR	.035	-.006	.479	.890
	(.315)	(.393)	(.555)	(.095)
CR^2	-.591	.117	-.555	-1.241
	(.337)	(.475)	(.932)	(.065)
Control	.137	-.022	.082	-.117
	(.108)	(.098)	(.192)	(.215)
Controls:				
leverage	.307	.495	.885	.357
	(.378)	(.214)	(.392)	(.558)
age	-.080	.105	-.008	-.024
	(.045)	(.123)	(.028)	(.081)
Observations	673			
Number of firms	116			

Standard errors are reported in parentheses. OLS, FE and Diff-GMM standard errors are robust with clustering at the firm level. Under our model, standard errors are obtained through bootstrapping. All estimators control for year fixed effects.

Figure 3.3: Marginal effect of CR_{it} on the log of total output

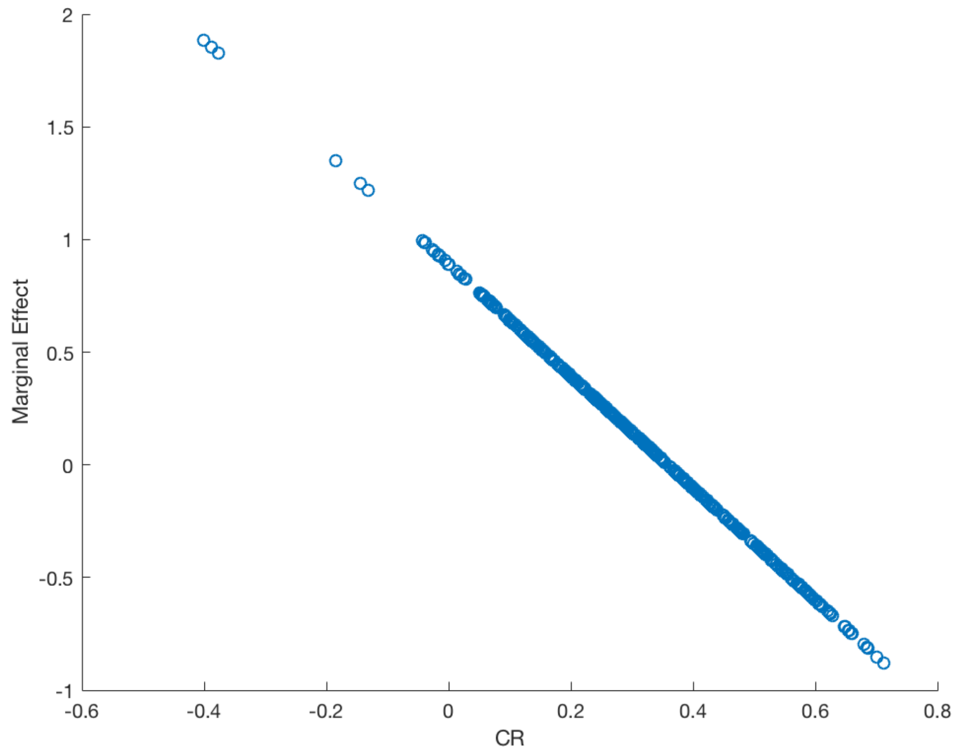
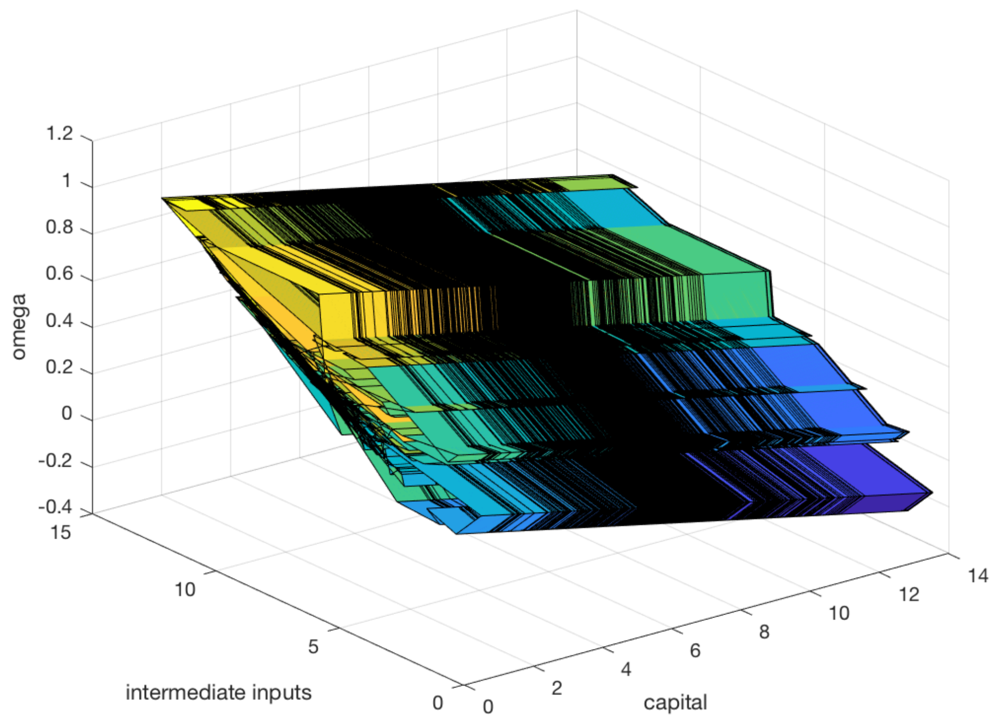


Figure 3.4: Omega as a Function of Capital and Intermediate Inputs



A

Appendix to Chapter 2

A.1 Local Identifiability

Suppose there are $J+1$ options (i.e. $j = 0, 1, \ldots, J$) so that there are $2^{J+1} = T+1$ consideration sets. For each subject i, choice probabilities

$$p_i\left(j \mid \boldsymbol{\beta}, \boldsymbol{\gamma}; \mathbf{x}_i\right) = \rho_i(\emptyset) p_i^{MNL}\left(j \mid X\right) + \sum_{\emptyset \neq C : j \in C} \rho_i(C) p_i^{MNL}(j \mid C), \quad j = 1, \ldots, J$$

can be stacked in the vector

$$\mathbf{p}_i = \begin{bmatrix} p_i\left(1 \mid \boldsymbol{\beta}, \boldsymbol{\gamma}; \mathbf{x}_i\right) \\ \vdots \\ p_i\left(j \mid \boldsymbol{\beta}, \boldsymbol{\gamma}; \mathbf{x}_i\right) \\ \vdots \\ p_i\left(J \mid \boldsymbol{\beta}, \boldsymbol{\gamma}; \mathbf{x}_i\right) \end{bmatrix} = \mathbf{P}_i \boldsymbol{\lambda}_i, \tag{A.1}$$

say, where

- $\mathbf{p}_i$ is a $J \times 1$ vector with j-th element equal to $p_i\left(j \mid \boldsymbol{\beta}, \boldsymbol{\gamma}; \mathbf{x}_i\right)$;
- $\boldsymbol{\lambda}_i$ is a $(T+1) \times 1$ vector with elements equal to $\left[\rho_i\left(\varnothing\right), \rho_i\left(C_1\right), \ldots, \rho_i\left(C_T\right)\right]'$;
- $\mathbf{P}_i$ is a $J \times (T+1)$ matrix whose (j, s) entry is given by $\mathbf{P}_i(j, s) = p_i^{MNL}(j \mid C_s)$.

Suppose we parametrize $\boldsymbol{\lambda}_i = \boldsymbol{\lambda}_i(\boldsymbol{\gamma}; \mathbf{x}_i)$ as a function of a real valued matrix $\boldsymbol{\gamma}$ of attention parameters and a $(K \times 1)$ vector $\mathbf{x}_i$ of individual regressors (including a constant term), and $\mathbf{P}_i = \mathbf{P}_i(\boldsymbol{\beta}; \mathbf{x}_i)$ as a function of a real valued matrix $\boldsymbol{\beta}$ of utility parameters and the same individual covariates $\mathbf{x}_i$.

$\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ respectively read as

$$\boldsymbol{\beta} = \begin{bmatrix} \mathbf{0} & \boldsymbol{\beta}_1 & \dots & \boldsymbol{\beta}_J \end{bmatrix} \tag{A.2}$$

and

$$\boldsymbol{\gamma} = \begin{bmatrix} \boldsymbol{\gamma}_0 & \boldsymbol{\gamma}_1 & \dots & \boldsymbol{\gamma}_J \end{bmatrix} \tag{A.3}$$

and are both of dimension $(K \times (J+1))$. Notice that the number of free utility parameters equals $(K \times J)$ because of the normalization $\boldsymbol{\beta}_0 = \mathbf{0}$. As for $\boldsymbol{\gamma}$, instead, we have $((K-1) \times J)$ free parameters since we are imposing the first element of $\boldsymbol{\gamma}_j$ equal for any $j = 0, 1, \dots, J$.

Letting $\boldsymbol{\psi} = [\boldsymbol{\beta}\ \boldsymbol{\gamma}]$, probabilities in (A.1) rewrite as

$$\mathbf{p}_i(\boldsymbol{\psi}; \mathbf{x}_i) = \mathbf{P}_i(\boldsymbol{\beta}; \mathbf{x}_i) \boldsymbol{\lambda}_i(\boldsymbol{\gamma}; \mathbf{x}_i) \tag{A.4}$$

Definition 1 A model is said to be *locally identifiable* if, for any $\boldsymbol{\psi}_0 = [\boldsymbol{\beta}_0\ \boldsymbol{\gamma}_0]$, the set of $\boldsymbol{\psi} = [\boldsymbol{\beta}\ \boldsymbol{\gamma}]$ for which $\mathbf{p}_i(\boldsymbol{\psi}; \mathbf{x}_i) = \mathbf{p}_i(\boldsymbol{\psi}_0; \mathbf{x}_i)$ satisfy $||\boldsymbol{\psi} - \boldsymbol{\psi}_0|| > \delta$ for some $\delta > 0$.

Local identifiability of the model is guaranteed if $\mathbf{p}_i(\boldsymbol{\psi}; \mathbf{x}_i)$ is injective in a neighbourhood of $\boldsymbol{\psi}$. The inverse function theorem states that this condition is satisfied when the Jacobian matrix of (A.4) is full rank.

A.1.1 Analytical derivation of the Jacobian matrix

The most basic scenario of interest where our model could be applied contemplates 3 mutually exclusive alternatives (i.e. $j = 0, 1, 2$) and a single regressor plus a constant (i.e. $K = 2$ and $\mathbf{x}_i = [1 \ x_i]'$ for each i). In this simplified framework, utility and attention parameters are respectively given by:

$$\boldsymbol{\beta} = \begin{bmatrix} \mathbf{0} & \boldsymbol{\beta}_1 & \boldsymbol{\beta}_2 \end{bmatrix} = \begin{bmatrix} 0 & \beta_1 & \beta_2 \\ 0 & \beta_{1x} & \beta_{2x} \end{bmatrix} \tag{A.5}$$

and

$$\boldsymbol{\gamma} = \begin{bmatrix} \boldsymbol{\gamma}_0 & \boldsymbol{\gamma}_1 & \boldsymbol{\gamma}_2 \end{bmatrix} = \begin{bmatrix} \gamma & \gamma & \gamma \\ \gamma_{0x} & \gamma_{1x} & \gamma_{2x} \end{bmatrix}. \tag{A.6}$$

Thus, for each subject i, the Jacobian matrix of (A.4) is given by

$$\mathbf{J}(\boldsymbol{\psi})_i = \begin{bmatrix} \frac{\partial p_i(1|\boldsymbol{\beta},\boldsymbol{\gamma};\mathbf{x}_i)}{\partial \boldsymbol{\beta}_1'} & \frac{\partial p_i(1|\boldsymbol{\beta},\boldsymbol{\gamma};\mathbf{x}_i)}{\partial \boldsymbol{\beta}_2'} & \frac{\partial p_i(1|\boldsymbol{\beta},\boldsymbol{\gamma};\mathbf{x}_i)}{\partial \boldsymbol{\gamma}_0'} & \frac{\partial p_i(1|\boldsymbol{\beta},\boldsymbol{\gamma};\mathbf{x}_i)}{\partial \boldsymbol{\gamma}_1'} & \frac{\partial p_i(1|\boldsymbol{\beta},\boldsymbol{\gamma};\mathbf{x}_i)}{\partial \boldsymbol{\gamma}_2'} \\ \frac{\partial p_i(2|\boldsymbol{\beta},\boldsymbol{\gamma};\mathbf{x}_i)}{\partial \boldsymbol{\beta}_1'} & \frac{\partial p_i(2|\boldsymbol{\beta},\boldsymbol{\gamma};\mathbf{x}_i)}{\partial \boldsymbol{\beta}_2'} & \frac{\partial p_i(2|\boldsymbol{\beta},\boldsymbol{\gamma};\mathbf{x}_i)}{\partial \boldsymbol{\gamma}_0'} & \frac{\partial p_i(2|\boldsymbol{\beta},\boldsymbol{\gamma};\mathbf{x}_i)}{\partial \boldsymbol{\gamma}_1'} & \frac{\partial p_i(2|\boldsymbol{\beta},\boldsymbol{\gamma};\mathbf{x}_i)}{\partial \boldsymbol{\gamma}_2'} \end{bmatrix}. \tag{A.7}$$

The entries of each row of (A.7) turn out to be a linear combination of logit and multinomial logit derivatives. In particular, differentiating (A.1) with respect to free utility and attention parameters, we have:

$$\frac{\partial p_i\left(j \mid \boldsymbol{\beta}, \boldsymbol{\gamma}; \mathbf{x}_i\right)}{\partial \beta_j} = p_i\left(j \mid \boldsymbol{\beta}, \boldsymbol{\gamma}; \mathbf{x}_i\right) - \Bigg(\rho_i(\emptyset)\left(p_i^{MNL}\left(j \mid X\right)\right)^2 + \sum_{\emptyset \neq C : j \in C} \rho_i(C)\left(p_i^{MNL}(j \mid C)\right)^2\Bigg), \tag{A.8}$$

$$\frac{\partial p_i\left(j \mid \boldsymbol{\beta}, \boldsymbol{\gamma}; \mathbf{x}_i\right)}{\partial \beta_{jx}} = x_i \frac{\partial p_i\left(j \mid \boldsymbol{\beta}, \boldsymbol{\gamma}; \mathbf{x}_i\right)}{\partial \beta_j}, \tag{A.9}$$

$$\frac{\partial p_i\left(j \mid \boldsymbol{\beta}, \boldsymbol{\gamma}; \mathbf{x}_i\right)}{\partial \beta_{k|k\neq j}} = -\rho_i(\emptyset) p_i^{MNL}\left(j \mid X\right) p_i^{MNL}\left(k \mid X\right) - \sum_{\emptyset \neq C : j,k \in C} \rho_i(C) p_i^{MNL}(j \mid C) p_i^{MNL}(k \mid C), \tag{A.10}$$

$$\frac{\partial p_i\left(j \mid \boldsymbol{\beta}, \boldsymbol{\gamma}; \mathbf{x}_i\right)}{\partial \beta_{kx|k\neq j}} = x_i \frac{\partial p_i\left(j \mid \boldsymbol{\beta}, \boldsymbol{\gamma}; \mathbf{x}_i\right)}{\partial \beta_{k|k\neq j}}, \tag{A.11}$$

$$\frac{\partial p_i\left(j \mid \boldsymbol{\beta}, \boldsymbol{\gamma}; \mathbf{x}_i\right)}{\partial \gamma} = -\left(\sum_{l=0}^{J} \Lambda(\boldsymbol{\gamma_l}'\boldsymbol{x_i})\right) \rho_i(\emptyset) p_i^{MNL}\left(j \mid X\right) + \sum_{\emptyset \neq C : j \in C} \left(\sum_{m \in C}(1 - \Lambda(\boldsymbol{\gamma_m}'\boldsymbol{x_i})) - \sum_{n \notin C} \Lambda(\boldsymbol{\gamma_n}'\boldsymbol{x_i})\right) \rho_i(C) p_i^{MNL}(j \mid C), \tag{A.12}$$

$$\frac{\partial p_i\left(j \mid \boldsymbol{\beta}, \boldsymbol{\gamma}; \mathbf{x}_i\right)}{\partial \gamma_{jx}} = x_i \left[-\Lambda(\boldsymbol{\gamma_j}'\boldsymbol{x_i})\rho_i(\emptyset)p_i^{MNL}\left(j \mid X\right) + \right.$$
$$\left. \sum_{\emptyset \neq C: j \in C} (1 - \Lambda(\boldsymbol{\gamma_j}'\boldsymbol{x_i}))\rho_i(C)p_i^{MNL}(j \mid C) \right], \tag{A.13}$$

$$\frac{\partial p_i\left(j \mid \boldsymbol{\beta}, \boldsymbol{\gamma}; \mathbf{x}_i\right)}{\partial \gamma_{kx|k\neq j}} = x_i \left[-\Lambda(\boldsymbol{\gamma_k}'\boldsymbol{x_i}))\rho_i(\emptyset)p_i^{MNL}\left(j \mid X\right) + \right.$$
$$\left. \sum_{\emptyset \neq C: j \in C} (\mathbf{1}(k \in C) - \Lambda(\boldsymbol{\gamma_k}'\boldsymbol{x_i}))\rho_i(C)p_i^{MNL}(j \mid C) \right]. \tag{A.14}$$

The formulae above immediately extend to the generic case where there are more than 3 alternatives and the number of explanatory variables exceeds 2. For each of the 5729 subjects in our sample, the individual Jacobian in (A.7) is given by a matrix of dimension

$$J \times (\underbrace{KJ}_{\text{\# of free utility parameters}} + \underbrace{(K-1)(J+1)+1}_{\text{\# of free attention parameters}}) \tag{A.15}$$

with $J = 5$ and $K = 6$, respectively.

To check local identifiability, we first concatenate vertically all the individual Jacobian matrices and obtain

$$\mathbf{J}(\boldsymbol{\psi}) = \begin{bmatrix} \mathbf{J}(\boldsymbol{\psi})_1 \\ \vdots \\ \mathbf{J}(\boldsymbol{\psi})_{5729} \end{bmatrix}. \tag{A.16}$$

Then, we evaluate (A.16) at the maximum likelihood estimates $\widehat{\boldsymbol{\psi}} = [\widehat{\boldsymbol{\beta}}\ \widehat{\boldsymbol{\gamma}}]$ reported in Table 1.1 and we find it to be full-column rank. Moreover, its condition number[1] is around 322 and this allows us to conclude, with considerable confidence, that the model is locally identified at least with respect to the explanatory variables in our dataset and estimated parameters. While still using the regressors in our data, we conduct an additional check by evaluating (A.16) at parameter values which are randomly sampled from a multivariate Normal distribution with variance equal to I, $2I$, $3I$ and $4I$, respectively. For each distribution, we run $1,000$ simulations. The Jacobian is never found to be rank deficient and condition numbers[2] are always below 10^7 (see, Figure A.1).

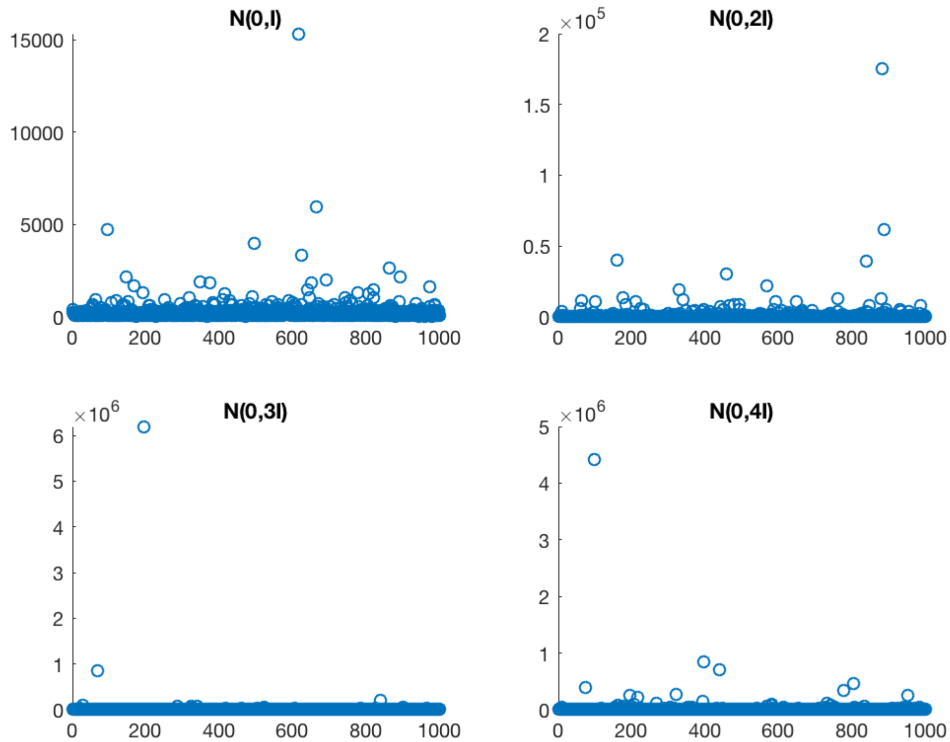
A.1.2 Further simulations

Adopting the simplest possible specification described above, we first run $20,000$ simulations in which we assume 4 individuals choosing among 3 available options and let both utility and attention be driven by a single independent variable (plus a constant term). Parameters in (A.5)-(A.6) along with the individual covariate

[1] That is, the ratio between the maximum and the minimum singular values of the Jacobian.

[2] If on a sufficiently large set of parameter points, the condition number is never higher than 10^{10}, we can be confident that the model is locally identified with probability close to one (see, Forcina (2008)).

Figure A.1: Condition Numbers: Actual Data



x_i are all drawn from a standard normal distribution.

The 8 parameters in (A.5)-(A.6) determine 2 "free" probabilities for each individual given that the probability of choosing alternative 0 is residually determined. Thus, for each of the 4 simulated subjects, we apply the formulae in (A.8)-(A.14) and compute individual Jacobian matrices $\mathbf{J}(\boldsymbol{\psi})_i$.

By vertically concatenating individual Jacobian matrices, at each simulation, we obtain a square 8×8 Jacobian of the form

$$\mathbf{J}(\boldsymbol{\psi}) = \begin{bmatrix} \mathbf{J}(\boldsymbol{\psi})_1 \\ \mathbf{J}(\boldsymbol{\psi})_2 \\ \mathbf{J}(\boldsymbol{\psi})_3 \\ \mathbf{J}(\boldsymbol{\psi})_4 \end{bmatrix}. \tag{A.17}$$

Constraining the number of individuals allows us to identify rank deficiency which would signal local non identifiability of the model. In all the simulations the Jacobian is found to be full rank. Furthermore, the condition number exceeds 10^5 in just 6 instances.

Then, in order to assess the usefulness of imposing the first element of $\boldsymbol{\gamma_j}$ equal across j we run $20,000$ further simulations of two different models. The first model corresponds to the one in (2.8) while in the second one we allow the attention parameters of the constant term to vary across alternatives. For the first specification parameters are the same as in (A.5)-(A.6) while for the second one they are given by:

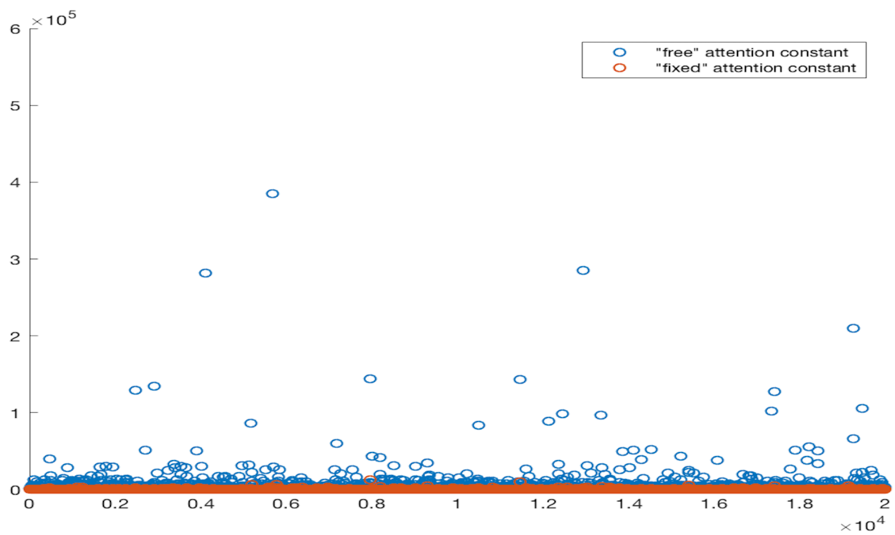
$$\boldsymbol{\beta} = \begin{bmatrix} 0 & \beta_1 & \beta_2 \\ 0 & \beta_{1x} & \beta_{2x} \end{bmatrix} \tag{A.18}$$

$$\boldsymbol{\gamma} = \begin{bmatrix} \gamma_0 & \gamma_1 & \gamma_2 \\ \gamma_{0x} & \gamma_{1x} & \gamma_{2x} \end{bmatrix} \tag{A.19}$$

Under this unrestricted specification, we assume 5 individuals choosing among 3 alternatives so as to get a 10×10 Jacobian matrix at each simulation.

Under both models the Jacobian is always found to be full rank. However, the condition number of the Jacobian is higher than 10^4 in just 7 instances under the first specification while this threshold is exceeded in 2745 cases under the second one (see, Figure A.2).

Figure A.2: Condition Numbers: 3 Alternatives & 2 Regressors

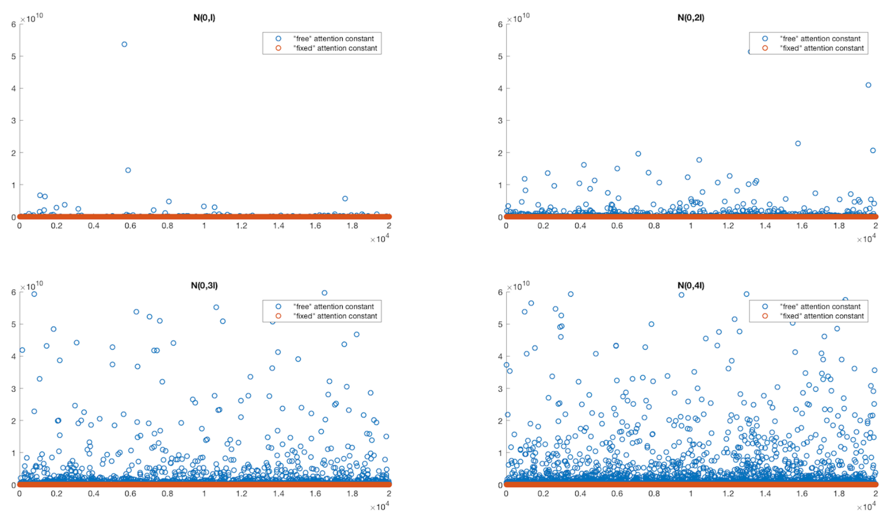


To provide further support to our identification result, we run 4 sets of 20,000 further simulations in which data is generated so as to mimic the SIPP dataset.

In particular we assume 6 options and 5 regressors[3] (including the constant term). The size of the parameter vector is equal to 50 in the case where the attention constant is assumed equal for all alternatives while is equal to 55 if this constant is let vary across options. In the same fashion as before, we constraint the number of simulated individuals to be equal to 10 in the restricted specification while being equal to 11 in the unrestricted one.

The four subplots of Figure A.3 scatter the condition numbers of the Jacobian computed as in (A.17) at each simulation when parameters are randomly drawn from a Normal distribution with variance equal to I, $2I$, $3I$ and $4I$, respectively. In

Figure A.3: Condition Numbers: 6 Alternatives & 5 Regressors



the case where a common average attention is imposed across alternatives (namely, our model), the condition number never exceeds 10^6 while, in the unrestricted

[3]Regressors are drawn from a standard normal.

specification, this threshold is exceeded in 663 instances when parameters are drawn from $N(0, I)$ and in almost half of the simulations (9,334 over 20,000 simulations) when parameters come from $N(0, 4I)$.

All in all, we can conclude that imposing a common "average attention" considerably facilitates parameters identification.

Bibliography

Abaluck, J. and Adams, A. (2017). What Do Consumers Consider Before They Choose? Identification From Asymmetric Demand Responses. Technical report, NBER.

Ackerberg, D. (2016). Timing assumptions and efficiency: empirical evidence in the production function context. Unpublished.

Ackerberg, D. A., Caves, K., and Frazer, G. (2015). Identification properties of recent production function estimators. *Econometrica*, 83(6):2411–2451.

Allman, E. S., Matias, C., and Rhodes, J. A. (2009). Identifiability of parameters in latent structure models with many observed variables. *Ann. Statist.*, 37(6A):3099–3132.

Anderson, T. W. and Hsiao, C. (1982). Formulation and estimation of dynamic models using panel data. *Journal of Econometrics*, 18(1):47–82.

Arellano, M. and Bond, S. (1991). Some tests of specification for panel data: Monte carlo evidence and an application to employment equations. *Review of Economic Studies*, 58(2):277–297.

Attanasio, O., Pacelli, L., and Dos Reis, I. R. (2000). Aggregate implications of plant level investment behavior: evidence from the uk. Working paper, University College London, United Kingdom.

Attig, N., Ghoul, S. E., and Guedhami, O. (2009). Do multiple large shareholders play a corporate governance role? evidence from east asia. *Journal of Financial Research*, 32(4):395–422.

Aumann, R. J. (2005). Consciousness. Discussion Paper Series dp391, The Federmann Center for the Study of Rationality, The Hebrew University, Jerusalem.

Banzhaf, J. F. (1965). Weighted voting doesn't work: A mathematical analysis. *Rutgers Law Review*, 19(2):317–343.

Barseghyan, L., Coughlin, M., Molinari, F., and Teitelbaum, J. C. (2019). Heterogeneous Choice Sets and Preferences. Papers 1907.02337, arXiv.org.

Bartolucci, F., Farcomeni, A., and Pennoni, F. (2012). *Latent Markov models for longitudinal data*. Chapman and Hall/CRC press.

Basar, G. and Bhat, C. (2004). A parameterized consideration set model for airport choice: an application to the San Francisco Bay Area. *Transportation Research Part B: Methodological*, 38(10):889–904.

Baum, L. E., Petrie, T., Soules, G., and Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. *The Annals of Mathematical Statistics*, 41(1):164–171.

Ben-Akiva, M. and Boccara, B. (1995). Discrete choice models with latent choice sets. *International Journal of Research in Marketing*, 12(1):9 – 24.

Benartzi, S. (2001). Excessive Extrapolation and the Allocation of 401 (k) Accounts to Company Stock. *Journal of Finance*, 56(5):1747–1764.

Berle, A. A. and Means, G. (1776). *The Modern Corporation and Private Property*. MacMillan, New York.

Bianchi, M. and Bianco, M. (2008). *L'evoluzione della corporate governance in Italia: Meno piramidi più coalizioni?* Il Mulino, Bologna.

Bloom, N., Mahajan, A., McKenzie, D., and Roberts, J. (2010). Why do firms in developing countries have low productivity? *American Economic Review*, 100(2):619–23.

Blundell, R. and Bond, S. (2000). Gmm estimation with persistent panel data: an application to production functions. *Econometric Reviews*, 19(3):321–340.

Bodie, Z., Merton, R. C., and Samuelson, W. F. (1992). Labor Supply Flexibility and Portfolio Choice in a Life-Cycle Model. *Journal of Economic Dynamics and Control*, 16:427–449.

Brown, J. R., Liang, N., and Weisbenner, S. (2006). 401(k) matching contributions in company stock: Costs and benefits for firms and workers. *Journal of Public Economics*, 90(6-7):1315–1346.

Caplin, A., Dean, M., and Leahy, J. (2018). Rational Inattention, Optimal Consideration Sets, and Stochastic Choice. *The Review of Economic Studies*, 86(3):1061–1094.

Caron, R. and Traynor, T. (2005). The zero set of a polynomial.

Chamberlain, G. (1982). Multivariate regression models for panel data. *Journal of Econometrics*, 18(1):5–46.

Choi, J. J., Laibson, D., and Madrian, B. C. (2005). Are Empowerment and Education Enough? Underdiversification in 401(k) Plans. *Brookings Papers on Economic Activity*, 2:151–213.

Choi, J. J., Laibson, D., Madrian, B. C., and Metrick, A. (2002). Defined Contribution Pensions: Plan Rules, Participant Choices, and the Path of Least Resistance. *Tax Policy and the Economy*, 16(1):67–113.

Dardanoni, V., Manzini, P., and Mariotti, M. (2017). A General Model of Consideration Sets: Theory and Applications. Unpublished.

Dardanoni, V., Manzini, P., Mariotti, M., and Tyson, C. J. (2019). Inferring Cognitive Heterogeneity from Aggregate Choices. Submitted.

De Miguel, A., Pindado, J., and de la Torre, C. (2004). Ownership structure and firm value: new evidence from spain. *Strategic Management Journal*, 25(12):1199–1207.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22.

Demsetz, H. and Villalonga, B. (2001). Ownership structure and corporate performance. *Journal of corporate finance*, 7(3):209–233.

Doms, M. and Dunne, T. (1998). Capital adjustment patterns in manufacturing plants. *Review of Economic Dynamics*, 1(2):409–429.

Draganska, M. and Klapper, D. (2011). Choice Set Heterogeneity and the Role of Advertising: An Analysis with Micro and Macro Data. *Journal of Marketing Research*, 48(4):653–669.

Dubey, P. and Shapley, L. S. (1979). Mathematical Properties of the Banzhaf Power Index. *Mathematics of Operations Research*, 4(2):99–131.

Edwards, J. S. and Weichenrieder, A. J. (2009). Control rights, pyramids, and the measurement of ownership concentration. *Journal of Economic Behavior & Organization*, 72(1):489–508.

Forcina, A. (2008). Identifiability of extended latent class models with individual covariates. *Computational Statistics and Data Analysis*, 52(12):5263–5268.

Foster, L., Haltiwanger, J. C., and Krizan, C. J. (2001). *Aggregate Productivity Growth: Lessons from Microeconomic Evidence*, pages 303–372. University of Chicago Press.

Gal, P. N. (2013). Measuring total factor productivity at the firm level using oecd-orbis. (1049).

Gaynor, M., Propper, C., and Seiler, S. (2016). Free to Choose? Reform, Choice, and Consideration Sets in the English National Health Service. *American Economic Review*, 106(11):3521–3557.

Gedajlovic, E. R. and Shapiro, D. M. (1998). Management and ownership effects: Evidence from five countries. *Strategic Management Journal*, 19(6):533–553.

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

Gourville, J. T. and Soman, D. (2005). Overchoice and assortment type: When and why variety backfires. *Marketing Science*, 24(3):382–395.

Grossman, S. J. and Hart, O. D. (1980). Takeover bids, the free-rider problem, and the theory of the corporation. *The Bell Journal of Economics*, 11(1):42–64.

Hall, M. and Tideman, N. (1967). Measures of concentration. *Journal of the American Statistical Association*, 62(317):162–168.

Hauser, J. R. and Wernerfelt, B. (1990). An evaluation cost model of consideration sets. *Journal of Consumer Research*, 16(4):393–408.

Heiss, F., Leive, A., McFadden, D., and Winter, J. (2013). Plan selection in medicare part d: Evidence from administrative data. *Journal of Health Economics*, 32(6):1325–1344.

Honka, E. and Chintagunta, P. (2017). Simultaneous or sequential? search strategies in the u.s. auto insurance industry. *Marketing Science*, 36(1):21–42.

Honka, E. and Chintagunta, P. K. (2016). Simultaneous or Sequential? Search Strategies in the U.S. Auto Insurance Industry. *Marketing Science*, pages 1–22.

Hortaçsu, A., Madanizadeh, S. A., and Puller, S. L. (2017). Power to choose? an analysis of consumer inertia in the residential electricity market. *American Economic Journal: Economic Policy*, 9(4):192–226.

Hulten, C. R. (2001). Total factor productivity: A short biography. In *New Developments in Productivity Analysis*, pages 1–54. National Bureau of Economic Research, Inc.

Jensen, M. C. and Meckling, W. H. (1976). Theory of the firm: Managerial behavior, agency costs and ownership structure. *Journal of Financial Economics*, 3(4):305 – 360.

Konijn, S. J., Kräussl, R., and Lucas, A. (2011). Blockholder dispersion and firm value. *Journal of Corporate Finance*, 17(5):1330–1339.

Kruskal, J. B. (1977). Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Linear Algebra and its Applications*, 18(2):95 – 138.

Laeven, L. and Levine, R. (2008). Complex ownership structures and corporate valuations. *Review of Financial Studies*, 21(2):579–604.

Leech, D. (2002). An empirical comparison of the performance of classical power indices. *Political Studies*, 50(1):1–22.

Levinsohn, J. and Petrin, A. (2003). Estimating production functions using inputs to control for unobservables. *Review of Economic Studies*, 70(2):317–341.

Lewbel, A. (2018). The Identification Zoo -Meanings of Identification in Econometrics. *Journal of Economic Literature (Forthcoming)*.

Luce, R. (1959). *Individual choice behavior: a theoretical analysis*. Wiley.

Maddala, G. (1979). A note on the form of the production function and productivity. In *Measurement and Interpretation of Productivity*. National Academy of Sciences, Washington, DC.

Manski, C. F. (1977). The structure of random utility models. *Theory and Decision*, 8(3):229–254.

Manzini, P. and Mariotti, M. (2007). Sequentially rationalizable choice. *The American Economic Review*, 97(5):1824–1839.

Manzini, P. and Mariotti, M. (2014). Stochastic choice and consideration sets. *Econometrica*, 82(3):1153–1176.

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

Maury, B. and Pajuste, A. (2005). Multiple large shareholders and firm value. *Journal of Banking Finance*, 29(7):1813 – 1834.

McFadden, D. (1973). *Conditional Logit Analysis of Qualitative Choice Behavior*. BART impact studies final report series: Traveler behavior studies. Institute of Urban and Regional Development, University of California.

McLachlan, G. and Krishnan, T. (2008). *The EM algorithm and extensions*. Wiley series in probability and statistics. Wiley-Interscience.

Milnor, J. W. and Shapley, L. (1978). Values of large games ii: Oceanic games. *Mathematics of Operations Research*, 3(4):290–307.

Modigliani, F. and Miller, M. H. (1958). The cost of capital, corporation finance and the theory of investment. *The American Economic Review*, 48(3):261–297.

Morck, R., Shleifer, A., and Vishny, R. W. (1988). Management ownership and market valuation: An empirical analysis. *Journal of financial economics*, 20:293–315.

Olley, G. S. and Pakes, A. (1996). The dynamics of productivity in the telecommunications equipment industry. *Econometrica*, 64(6):1263–97.

Overland, C., Mavruk, T., and Sjögren, S. (2012). Keeping it real or keeping it simple? ownership concentration measures compared. In *The 4th workshop of the Nordic Corporate Governance Network, Reykjavik University June 15, 2012*.

Papke, L. E. and Poterba, J. M. (1995). Survey evidence on employer match rates and employee saving behavior in 401(k) plans. *Economics Letters*, 49(3):313–317.

Rhodes, J. A. (2010). A concise proof of kruskal's theorem on tensor decomposition. *Linear Algebra and its Applications*, 432(7):1818 – 1824.

Roberts, M. R. and Whited, T. M. (2013). Endogeneity in empirical corporate finance1. In *Handbook of the Economics of Finance*, volume 2, pages 493–572. Elsevier.

Robinson, P. M. (1988). Root-n-consistent semiparametric regression. *Econometrica*, 56(4):931–954.

Rubinstein, A. and Salant, Y. (2006). A model of choice from lists. *Theoretical Economics*, 1(1):3–17.

Russino, A., Picone, P. M., and Dagnino, G. B. (2019). Unveiling the role of multiple blockholders: Evidence from closely held firms. *Corporate Governance: An International Review*.

Samuelson, P. A. (1938). A note on the pure theory of consumer's behaviour. *Economica*, 5(17):61–71.

Shleifer, A. and Vishny, R. W. (1986). Large shareholders and corporate control. *Journal of Political Economy*, 94(3):461–488.

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

Sidiropoulos, N. D. and Bro, R. (2000). On the uniqueness of multilinear decomposition of n-way arrays. *Journal of Chemometrics*, 14(3):229–239.

Skrondal, A. and Rabe-Hesketh, S. (2004). *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models*. Chapman & Hall/CRC Interdisciplinary Statistics. CRC Press.

Smith, A. (1776). *An Inquiry into the Nature and Causes of the Wealth of Nations*. McMaster University Archive for the History of Economic Thought.

Sunden, A. E. and Surette, B. J. (1998). Gender differences in the allocation of assets in retirement savings plans. *American Economic Review*, 88(2):207–11.

Syverson, C. (2011). What determines productivity? *Journal of Economic literature*, 49(2):326–65.

Thaler, R. H. and Benartzi, S. (2001). Naive diversification strategies in defined contribution saving plans. *American Economic Review*, 91(1):79–98.

Tian, G. Y. and Twite, G. (2011). Corporate governance, external market discipline and firm productivity. *Journal of Corporate Finance*, 17(3):403–417.

Treisman, A. M. and Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, 12(1):97 – 136.

VanDerhei, J., Holden, S., and Alonso, L. (2010). 401 (K) Plan Asset Allocation, Account Balances, and Loan Activity in 2009. *EBRI Issue Brief*, 401(350):1–67.

Watanabe, M. and Yamaguchi, K. (2003). *The EM Algorithm and Related Statistical Models*. Statistics: A Series of Textbooks and Monographs. CRC Press.

Wooldridge, J. M. (2009). On estimating firm-level production functions using proxy variables to control for unobservables. *Economics Letters*, 104(3):112–114.

Wright, P. and Barbour, F. (1977). *Phased Decision Strategies: Sequels to an Initial Screening*. Research paper. Graduate School of Business, Stanford University.