Inferring Cognitive Heterogeneity from Aggregate Choices*

Valentino Dardanoni
Università degli Studi di Palermo

Paola Manzini
University of Sussex and IZA

Marco Mariotti
Queen Mary University of London

Christopher J. Tyson
Queen Mary University of London

May 25, 2018

Abstract

One potentially important drawback of existing theories of limited attention is that they typically assume a rich dataset of choices from many menus. We study the problem of identifying the distribution of cognitive characteristics in a population of agents when only aggregate choice behavior from a *single* menu is observable. We show how both "consideration probability" and "consideration capacity" distributions can be substantially identified by aggregate choice shares. We also suggest how to embed the attention models in an econometric specification of the inference problem. Finally, we sucessfully use our results to recover the true parameters in Monte Carlo simulations of both models.

Keywords: attention, bounded rationality, consideration set, stochastic choice.

J.E.L. codes: D01, D12, D91.

*We would like to thank Jason Abaluck, Abi Adams, Levon Barseghyan, Reyer Gerlagh, Alessandro Iaria, Rod McCrorie, Yusufcan Masatlioglu, Irina Merkurieva, Francesca Molinari and Ted O'Donoghue for valuable comments and suggestions. We also thank seminar audiences at D-TEA, Edinburgh, Leuven, Oxford, St Andrews, Sussex, Tilburg, ULB (ECARES), Vienna, Warwick (CRETA), the Barcelona GSE Summer Forum, the Italian Congress of Econometrics and Empirical Economics, and the Conference on Identification and Inference in Limited Attention Models at Cornell University. Manzini and Mariotti acknowledge financial support from the ESRC through grant ES/J012513/1.

1 Introduction

1.1 Motivation

Classical revealed preference analysis has yielded a fine-grained understanding of the relationship between unobserved tastes and observed choices, and of how to infer the former from the latter. More recently, theoretical work on bounded rationality has extended this methodology to incorporate a range of cognitive factors that may affect decision making.[1] One drawback of such theories is that they typically assume a rich dataset—comprising a single individual's choices from many different overlapping menus—that can be used to identify the latent components of the cognitive model of interest. For instance, Masatlioglu et al. [28] and Cattaneo et al. [10] require data for all possible menus drawn from a universal set of alternatives; Manzini and Mariotti [26] impose a stringent ("richness") assumption on their dataset; and Caplin and Dean [7] postulate the observability of state-dependent stochastic choice data.[2]

Identification results derived using such assumptions on the choice domain are often formally elegant, and can be particularly useful for designing and interpreting experiments (as in Caplin and Dean [7]). They are less obviously relevant to field data, however, especially when the category of decision problem arises rarely (e.g., choice of hospital provider for elective surgery) or the menu is slow to change (e.g., choice of daily newspaper). Indeed, in settings with these characteristics many existing results on boundedly rational choice may appear to be unrealistically data hungry. Even where individual choices are observable, there may well be insufficient variation in the set of available options to infer the model components of interest. For this reason, it is desirable to develop alternative approaches to identification that establish more direct

[1]This literature examines cognitive factors such as computational constraints, norms and heuristics, reference points and other framing effects, and various conceptions of attention. Contributions include those of Apesteguia and Ballester [2], Baigent and Gaertner [3], Caplin and Dean [7], Caplin et al. [8], Cattaneo et al. [10], Cherepanov et al. [11], de Oliveira et al. [15], Echenique et al. [16], Manzini and Mariotti [25, 26], Masatlioglu and Nakajima [27], Masatlioglu et al. [28], Ok et al. [33], Salant and Rubinstein [37], and Tyson [42, 43], among numerous others.

[2]Even stronger assumptions about data availability are commonplace in the theory of choice under uncertainty, where the decision maker is typically imagined to express preferences over a highly structured mathematical space specially designed to facilitate identification.

links between theory and what is feasible empirically.

In this paper we focus on models of limited attention, where agents consider only a subset of the available alternatives known as the "consideration set."[3] To address the problem of data voracity noted above, we propose a novel approach that takes as primitive *a single, fixed menu* faced by a population of decision makers and for which we observe (only) the *aggregate choice shares* of the different options.[4] Individuals may of course choose differently from this menu because of taste heterogeneity (that is, variation in preferences). But even when they share the same tastes agents may choose differently due to "cognitive heterogeneity," by which we mean variation in non-preference characteristics such as attention that are relevant to behavior. Our aim is to link variation of the latter sort to the observed choice shares, developing identification results that complement existing techniques for inferring taste heterogeneity from choice data under the (usually implicit) assumption of full attention.

We emphasize that our main interest is in the extent to which the distribution of cognitive characteristics is identified by a given model of choice *per se*; that is, prior to any ancillary econometric specification that may include covariates for the individuals or the alternatives. In this sense, our primitives and objectives are typical of those found elsewhere in abstract choice theory.[5] As demonstrated below, it is necessary to adopt a full-blown econometric model only when cognitive heterogeneity is combined with taste heterogeneity and both forms of variation are unobserved. The problem of unobserved taste heterogeneity can be dealt with in a standard way, namely through parametric logit assumptions which are routinely used to identify stochastic preferences in models of full attention.

[3]This usage follows the marketing literature; see, e.g., Roberts and Lattin [35, 36] and Shocker et al. [38].

[4]Alternatively, the framework can be used to model a single individual choosing repeatedly from the same menu in different attentional states. Here the variation could arise, for example, from a retailer periodically rearranging the display of products in a deliberate attempt to manipulate their customers' consideration sets.

[5]This distinguishes our approach from contributions such as that of Abaluck and Adams [1], where identification arises from the elasticities of choice shares with respect to observable characteristics of the alternatives.

1.2 Cognitive models

In our general framework, each agent has a *cognitive type* parameter $\theta \in \Theta$ that is distributed in the population according to F. Under the assumption of homogeneous tastes, an individual of type θ will choose alternative x with probability $p_\theta(x)$ and the corresponding aggregate choice share will be $p(x) = \int_\Theta p_\theta(x)\, dF$. When the cognitive type is used to capture some form of bounded rationality, neither the individual nor the aggregate choice distribution will assign all weight to the best available option (according to the common preference order of the agents). Indeed, it is the fact that suboptimal alternatives will sometimes be chosen that will enable us to infer features of the cognitive distribution F from the observed aggregate shares.[6]

More specifically, we shall study bounded rationality in the form of limited attention, and so the cognitive parameter θ will affect the formation of the consideration set.[7] We examine in detail two different models of how this occurs, referred to as the *ρ-model* and the *γ-model*. The first is a variant of the structure in Manzini and Mariotti [26], with parameter $\rho \in [0, 1]$ controlling the likelihood that each option is considered and interpreted as the agent's general awareness of the decision-making environment. In contrast, the second model has parameter $\gamma \in \{0, 1, 2, \ldots\}$ controlling the cardinality of the consideration set and interpreted as a limit on the number of alternatives that the agent can actively investigate at any one time.[8] Both models assume that preferences are maximized over the consideration set; both include full rationality as a special case (respectively, setting $\rho = 1$ or $\gamma \to \infty$); and both specify a default consequence in case the consideration set is empty.

[6]Note that our framework has similarities to mixed models in the discrete choice literature, where θ would be a taste parameter such as the agent's unobserved marginal utility of some observed characteristic. (See Train [41] and McFadden [29].) However, since we shall use θ to control cognition instead of tastes, our setting calls for different functional form assumptions. In particular, p_θ will not have a logit specification (see Luce [22]), as would typically be assumed in relation to tastes.

[7]While we view the consideration set as a manifestation of bounded rationality, it is worth noting that other interpretations are possible. Indeed, alternatives may fail to be considered due to habit formation, search costs, or other forms of rational inattention (see, e.g., Caplin and Dean [7] and Sims [39]).

[8]Variants of this model are used by Barseghyan et al. [4] to study discrete choice with heterogeneous consideration sets and by de Clippel et al. [14] to study price competition in a setting where consumers exhibit limited attention.

1.3 Preview of results

We begin with the case of homogeneous tastes, where under weak conditions the common preference order can be inferred from the aggregate choice shares. We observe first that when this order is unrestricted, neither of our models in general allows full identification of the cognitive distribution. However, if we impose the mild assumption that all preferences are strict, then for several natural parameterizations of F both models are seen to permit full identification using a small number of observed shares. For instance, if the consideration probability ρ is uniformly distributed on an interval, then the endpoints of this interval can be recovered from the shares of the two most preferred alternatives (see Example 2).

We proceed to show that even in the absence of parametric assumptions, the cognitive distribution can for practical purposes be fully recovered in either model provided the menu of alternatives is sufficiently large. More precisely, under the γ-model the aggregate choice shares identify the probabilities of all consideration capacities less than the cardinality n of the menu (see Proposition 2). Likewise, under the ρ-model the aggregate shares identify the first n raw moments of F (see Proposition 3), which—with the aid of maximum entropy methods and results from sparsity theory—can be used to reconstruct or closely approximate the distribution itself (see Propositions 4–5). In each setting, identification follows from the system of equations that define the choice shares being recursive and linear in the relevant quantities (namely, the capacity probabilities or raw moments), so that closed-form expressions for these quantities can be obtained by inverting a triangular or anti-triangular matrix.

Turning to the case of heterogeneous tastes, we note that our identification results continue to hold generically if the distribution of preferences is known. Relaxing the latter assumption, we then develop an econometric specification of our models in which a vector of characteristics of the alternatives is observable in addition to their aggregate choice shares. In this context, we establish that both cognitive and taste parameters are (generically) identified by the levels of the choice shares in combination with their sensitivity to changes in the characteristics.

Finally, we conduct Monte Carlo simulations of both cognitive models and show

how our theoretical findings can be used to accurately recover the true parameters.

1.4 Outline

The remainder of the paper is structured as follows. Section 2 describes our general framework and the two specific models of consideration-set formation. Section 3 develops our theoretical results for homogeneous preferences, and Section 4 incorporates taste heterogeneity. Section 5 presents the Monte Carlo simulations, and Section 6 concludes.

2 Two models of consideration set formation

2.1 General framework

Let X denote the universal set of alternatives. A *menu* is any nonempty $A \subseteq X$, with which is associated a default outcome d_A (not in A). When faced with the menu A, an agent either chooses exactly one of the available alternatives or chooses none and accepts d_A. For example, we could have that:

(i) the menu contains retailers selling a particular product, and the default is not to buy;

(ii) the menu contains banks offering fixed-term deposits, and the default is to hold cash; or

(iii) the menu contains risky lotteries, and the default is a risk-free payment.

Initially, when deriving our main theoretical results (in Sections 2–3), we shall assume that all agents share the same preference order $\succsim$ over the alternatives. Equivalently, this can be thought of as using the average utilities of alternatives in the population, ignoring variation. In this sense our approach is complementary to that of the classical stochastic-choice literature in economics, where preferences are allowed to vary but cognitive capabilities are implicitly assumed to be uniform. Observe that homogeneous tastes are plausible in examples (i) and (ii) above, where preferences will be determined largely by price and interest rate comparisons, as well as in example (iii) if all agents are approximately risk neutral over the relevant stakes.

When imposing homogeneous tastes, we number the alternatives consistently with the preference order (i.e., so that more preferred options have lower indices) and arbitrarily within each indifference class. We thus write k_A for the kth best option on menu A, and the full menu appears as $A = \{1_A, 2_A, \ldots, |A|_A\}$. The homogeneous-tastes assumption is relaxed beginning in Section 4.

We model cognitive heterogeneity by assigning each agent a cognitive type $\theta \in \Theta \subset \Re$, drawn independently across agents from the distribution F. We write $p_\theta(k_A)$ for the probability that type θ chooses alternative k_A from menu A, and $p(k_A) = \int_\Theta p_\theta(k_A)\,\mathrm{d}F$ for the overall share in the population. Similarly, we write $p_\theta(d_A)$ for the probability that type θ accepts the default consequence, and $p(d_A) = \int_\Theta p_\theta(d_A)\,\mathrm{d}F$ for the population share. For each $\theta \in \Theta$ we have $\sum_{k=1}^{|A|} p_\theta(k_A) + p_\theta(d_A) = 1$, and likewise in aggregate $\sum_{k=1}^{|A|} p(k_A) + p(d_A) = 1$. When we wish to emphasize the role of the type distribution in determining the choice probabilities, we write $p(k_A; F)$ and $p(d_A; F)$.

The basic scenario of interest involves the members of a large population choosing from a fixed menu M, where $|M| = n \geq 2$. The analyst observes the aggregate choice shares, but knows neither the preferences nor the distribution F of cognitive types. In this context we shall generally suppress dependence on M, writing $p_\theta(k)$ and $p_\theta(d)$ for the type-specific frequencies and $p(k)$ and $p(d)$ for the population shares. Our goal is to deduce information about the type distribution from the shares $\langle p(1), p(2), \ldots, p(n), p(d)\rangle$, and to use this knowledge to predict aggregate choice behavior from menus other than M.

We proceed now to specialize this framework to two more concrete models illustrating different ways that an agent's attention to the alternatives may be limited. Among the options that are considered (i.e., that attract attention), each agent will choose the best according to the shared preference order. If multiple options are equally best among those considered, then they will be chosen with equal probabilities. However, since the alternatives considered may be a strict subset of those actually available, the attention deficits captured in the two specialized models may result in sub-optimal decision making.

2.2 Consideration probability: The ρ-model

Let $\rho \in [0,1] = \Theta$ denote the probability that the agent considers each alternative on the menu, with consideration independent across agents and alternatives.[9] It follows that alternative k will be chosen with positive probability if and only if the agent both (i) notices k and (ii) fails to notice each alternative that is strictly preferred to k. Moreover, by assumption, any two alternatives for which (i) and (ii) both hold will be chosen with equal probabilities. The default consequence will arise if no alternatives at all are noticed.

To write the choice probabilities explicitly, for each k denote by $\omega_k \leq k$ the smallest index ω such that $k \sim \omega$ and by $\omega^k \geq k$ the largest such index. Conditional on the cognitive type ρ, the probability that alternative k is chosen can then be expressed as

$$\begin{aligned} p_\rho(k) &= \text{Prob}[k \text{ chosen}|\text{any } i \sim k \text{ chosen}] \times \text{Prob}[\text{any } i \sim k \text{ chosen}] \\ &= \frac{1}{\omega^k - \omega_k + 1} \times \left[[1-\rho]^{\omega_k - 1} - [1-\rho]^{\omega^k} \right], \end{aligned} \tag{1}$$

and the default consequence arises with probability $p_\rho(d) = [1-\rho]^n$. Here the first term in Equation 1 is the reciprocal of the cardinality of the indifference class containing k, and the second term is the difference in probabilities between the event that no alternative better than k is considered and the event that no alternative at least as good as k is considered (the latter event being a subset of the former).

Aggregate choice shares are obtained by integrating over the type space, yielding

$$p(k) = \frac{1}{\omega^k - \omega_k + 1} \int_0^1 \left[[1-\rho]^{\omega_k - 1} - [1-\rho]^{\omega^k} \right] \mathrm{d}F, \tag{2}$$

$$p(d) = \int_0^1 [1-\rho]^n \mathrm{d}F. \tag{3}$$

If the indifference class containing alternative k is a singleton, then $\omega_k = \omega^k = k$ and

[9] Variants of this model have been studied by Manzini and Mariotti [26] and Brady and Rehbeck [6].

Equation 2 simplifies to

$$p(k) = \int_0^1 \rho[1-\rho]^{k-1} \mathrm{d}F. \tag{4}$$

This formula computes the probability of noticing k while failing to notice any of the $k-1$ superior alternatives.

2.3 Consideration capacity: The γ-model

Let $\gamma \in \{0, 1, 2, \ldots\} = \Theta$ denote the number of alternatives that the agent is able to consider; that is, the "consideration capacity." When $\gamma < n$ we assume that the agent is equally likely to consider each $\Gamma \subset M$ with $|\Gamma| = \gamma$, and in this case there are $\binom{n}{\gamma}$ candidate consideration sets. When $\gamma \geq n$ we know with certainty that the entire menu M will be considered.

As before, we denote by ω_k and ω^k the smallest and largest indices of alternatives in the indifference class containing k. For $\gamma \leq n$ the probability that no alternative better than k is considered is then $\binom{n-\omega_k+1}{\gamma}/\binom{n}{\gamma}$; the probability that no alternative at least as good as k is considered is $\binom{n-\omega^k}{\gamma}/\binom{n}{\gamma}$; and the probability of choosing something indifferent to k is the difference between these two ratios. It follows that the type-conditional choice frequencies are

$$\begin{aligned} p_\gamma(k) &= \text{Prob}[k \text{ chosen}|\text{any } i \sim k \text{ chosen}] \times \text{Prob}[\text{any } i \sim k \text{ chosen}] \\ &= \frac{1}{\omega^k - \omega_k + 1} \times \frac{\binom{n-\omega_k+1}{\gamma} - \binom{n-\omega^k}{\gamma}}{\binom{n}{\gamma}}, \end{aligned} \tag{5}$$

$$p_\gamma(d) = \begin{cases} 0 & \text{if } \gamma > 0, \\ 1 & \text{if } \gamma = 0. \end{cases} \tag{6}$$

Defining the probability masses

$$\pi(0) = F(0), \tag{7}$$

$$\forall \gamma \in \{1, 2, \ldots, n-1\}, \quad \pi(\gamma) = F(\gamma) - F(\gamma - 1), \tag{8}$$

$$\pi(n) = 1 - F(n-1); \tag{9}$$

the corresponding aggregate choice shares are

$$p(k) = \frac{1}{\omega^k - \omega_k + 1} \sum_{\gamma=1}^{n-\omega_k+1} \frac{\binom{n-\omega_k+1}{\gamma} - \binom{n-\omega^k}{\gamma}}{\binom{n}{\gamma}} \pi(\gamma), \tag{10}$$

$$p(d) = \pi(0). \tag{11}$$

If the indifference class containing alternative k is a singleton, then Equation 10 becomes

$$p\left(k\right) = \sum_{\gamma=1}^{n-k+1} \frac{\binom{n-k+1}{\gamma} - \binom{n-k}{\gamma}}{\binom{n}{\gamma}} \pi(\gamma) = \sum_{\gamma=1}^{n-k+1} \frac{\binom{n-k}{\gamma-1}}{\binom{n}{\gamma}} \pi(\gamma). \tag{12}$$

The latter formula computes the probability of realizing a consideration set made up of k and other noticed alternatives drawn from the $n - k$ options that are inferior to k.

2.4 Relationship between the two models

In the ρ-model the same consideration probability applies independently to each option, and so all subsets of the menu of a given size are equally likely to be the consideration set. It follows that the ρ-model is a special case of the γ-model, where the consideration set contains exactly $\gamma \leq n$ alternatives with probability

$$\pi(\gamma) = \int_0^1 \binom{n}{\gamma} \rho^\gamma [1-\rho]^{n-\gamma} \mathrm{d}F. \tag{13}$$

Specializing Equations 10–11 to this case, we obtain

$$\begin{aligned}
p(k) &= \frac{1}{\omega^k - \omega_k + 1} \sum_{\gamma=1}^{n-\omega_k+1} \frac{\binom{n-\omega_k+1}{\gamma} - \binom{n-\omega^k}{\gamma}}{\binom{n}{\gamma}} \int_0^1 \tbinom{n}{\gamma} \rho^\gamma [1-\rho]^{n-\gamma} \mathrm{d}F \\
&= \frac{1}{\omega^k - \omega_k + 1} \int_0^1 \left[\sum_{\gamma=0}^{n-\omega_k+1} \tbinom{n-\omega_k+1}{\gamma} \rho^\gamma [1-\rho]^{n-\gamma} - \sum_{\gamma'=0}^{n-\omega^k} \tbinom{n-\omega^k}{\gamma'} \rho^{\gamma'} [1-\rho]^{n-\gamma'} \right] \mathrm{d}F \\
&= \frac{1}{\omega^k - \omega_k + 1} \int_0^1 \left[[1-\rho]^{\omega_k - 1} - [1-\rho]^{\omega^k} \right] \mathrm{d}F
\end{aligned} \tag{14}$$

and $p(d) = \pi(0) = \int_0^1 [1-\rho]^n \mathrm{d}F$, reproducing Equations 2–3.

3 Inference from aggregate choices

3.1 Preference identification

Under very weak conditions, the agents' common preference order $\succsim$ over the alternatives is fully revealed by the observed choice shares. More precisely, we have the following result.

Proposition 1. (i) *In the γ-model, if $\sum_{\gamma=2}^{n-\omega_n+2} \pi(\gamma) > 0$ then $p(j) \geq p(k) \iff j \succsim k$.* (ii) *In the ρ-model, if the support of F intersects $(0,1)$ then $p(j) \geq (k) \iff j \succsim k$.*

Indeed, for any alternatives j and k such that $\omega^j < \omega_k$ we have in the γ-model that

$$p_\gamma(j) \geq \frac{\binom{n-\omega^j}{\gamma-1}}{\binom{n}{\gamma}} \geq \frac{\binom{n-\omega_k}{\gamma-1}}{\binom{n}{\gamma}} \geq p_\gamma(k) \tag{15}$$

for all $\gamma \in \{1, 2, \ldots, n\}$, with the second inequality in the chain strict for $2 \leq \gamma \leq n - \omega^j + 1$. In Equation 15, the first inequality in the chain reflects the fact that j will be chosen if the consideration set consists of j and $\gamma - 1$ alternatives that are strictly worse. Similarly, the third inequality reflects the fact that k will be chosen only if the consideration set consists of k and $\gamma - 1$ alternatives that are no better. It follows that condition (i) in Proposition 1 is sufficient for $p_\gamma(j) \geq p_\gamma(k) \iff j \succsim k$. That is to say, if positive probability is assigned to any capacity that is sufficiently small (but larger than one), then a larger choice share for an alternative is equivalent to a higher position in the preference order.

As noted above in Section 2.4, the ρ-model is a special case of the γ-model with capacity distribution given by Equation 13. In this context condition (ii) in Proposition 1 implies that $\pi(2) = \frac{n[n-1]}{2} \int_0^1 \rho^2 [1-\rho]^{n-2} \mathrm{d}F > 0$, whereupon condition (i) also holds.

In summary, with homogeneous tastes the preference order is for practical purposes fully revealed by the order of the choice shares in both the ρ- and γ-models. A mild assumption on F is needed to guarantee complete revelation, but we shall in any event be imposing stronger conditions on this distribution in the analysis of cognitive identification to follow.

3.2 Cognitive identification in general: The problem of indifferences

While the preference order is fully identified in both models under weak conditions, in general it may not be possible to retrieve the distribution of cognitive parameters from the observed choice probabilities. We demonstrate this with an example in the setting of the γ-model, though the point holds also for the ρ-model.

Example 1. [*γ-model; identification failure*] Let $n = 4$ and $1 \succ 2 \sim 3 \succ 4$. Under the γ-model, Equations 10–11 yield

$$p(1) = \frac{1}{4}\pi(1) + \frac{1}{2}\pi(2) + \frac{3}{4}\pi(3) + \pi(4), \tag{16}$$

$$p(2) = p(3) = \frac{1}{4}\pi(1) + \frac{1}{4}\pi(2) + \frac{1}{8}\pi(3), \tag{17}$$

$$p(4) = \frac{1}{4}\pi(1), \tag{18}$$

$$p(d) = \pi(0). \tag{19}$$

Using Equations 18–19, we can recover the probability masses $\pi(0) = p(d)$ and $\pi(1) = 4p(4)$. However, the remaining equations fail to identify the masses $\pi(2)$, $\pi(3)$, and $\pi(4)$, due to the redundancy of the expressions for the choice shares $p(2)$ and $p(3)$. □

To understand the preceding example, recall that we observe (only) the aggregate choice shares $\langle p(1), p(2), \ldots, p(n), p(d)\rangle$, constituting a dataset with n degrees of freedom. Since with homogeneous tastes no degrees of freedom are needed to identify the preferences, we expect the data to reveal at most n cardinal features of the cognitive distribution. One obstacle to maximal identification in this sense is the presence (as in the example) of indifferences in the preference order, which reduce the effective number of degrees of freedom in the dataset.

To avoid problems with identification arising solely from indifferences, we shall focus on the case of strict (i.e., strict total order) preferences for the remainder of Section 3. With this restriction we shall see that maximal identification is generally possible in both the ρ-model and the γ-model. Parameterized distributions with n or fewer parameters can be fully revealed by the choice shares (see Section 3.3), and in non-parametric settings we can identify up to n probability masses (see Section 3.4.2) or moments of F (see Section 3.4.3).

Observe that in applications of the cognitive heterogeneity framework indifferences may arise with probability zero, in which case the analysis with strict preferences is sufficient. This will be true, for instance, when tastes are generated by a random utility model with atomless error terms, as in the simulation exercises described below in Section 5.

3.3 Cognitive identification with strict preferences: Parametric analysis

Assume that all preferences are strict, which is to say that $1 \succ 2 \succ \cdots \succ n$. In this case each alternative k makes up its own indifference class, and so Equations 4 and 12 apply for the ρ- and γ-models, respectively. We proceed now to consider several natural functional forms for the cognitive type distribution, aiming to show that the cognitive parameters can be revealed in a straightforward fashion from a relatively small number of appropriately selected choice-share observations. In addition to increasing our familiarity with the models under investigation, the examples below highlight the non-obvious ways that aggregate choices can convey information about the cognitive distribution.

In our first two examples, the type distribution can be retrieved from the same number of choice share observations as there are cognitive parameters.

Example 2. [*ρ-model; uniform distribution*] For $0 \leq \rho_{\min} < \rho_{\max} \leq 1$, let the consideration probability ρ be distributed uniformly on $[\rho_{\min}, \rho_{\max}]$. Then $F(\rho) = \frac{\rho - \rho_{\min}}{\rho_{\max} - \rho_{\min}}$ and the first two choice shares from Equation 4 are

$$p(1) = \frac{\rho_{\max} + \rho_{\min}}{2}, \tag{20}$$

$$p(2) = \frac{\rho_{\max} + \rho_{\min}}{2} - \frac{\rho_{\max}^2 + \rho_{\max}\rho_{\min} + \rho_{\min}^2}{3}. \tag{21}$$

The transformation in Equations 20–21 can be inverted to yield

$$\rho_{\max} = p\,(1) + \sqrt{3[p(1) - p(1)^2 - p(2)]}, \tag{22}$$

$$\rho_{\min} = p\,(1) - \sqrt{3[p(1) - p(1)^2 - p(2)]}; \tag{23}$$

expressions for the (unobserved) parameters in terms of the (observed) choice shares. □

Example 3. [*γ-model; Poisson distribution*] For $\mu \geq 0$, let the consideration capacity γ have the Poisson distribution $\pi(\gamma) = \frac{\mu^\gamma}{\gamma!}e^{-\mu}$ (for $0 \leq \gamma < n$). Then Equation 11 yields the default share $p\,(d) = e^{-\mu}$, and we obtain $\mu = -\log p(d)$. □

In our next example identification of the type distribution is more challenging, and involves recovering a pair of parameters from the choice shares of the two worst alternatives, the default share, and the size of the menu.

Example 4. [*γ-model; Pascal distribution*] For $r \in \{1, 2, 3, \ldots\}$ and $q \in (0, 1)$, let γ have the Pascal (or "negative binomial") distribution $\pi(\gamma) = \binom{\gamma+r-1}{\gamma}[1-q]^r q^\gamma$ (for $0 \leq \gamma < n$). Here Equations 11–12 become

$$p(k) = [1-q]^r \sum_{\gamma=1}^{n-k+1} \frac{\binom{n-k}{\gamma-1}}{\binom{n}{\gamma}} \binom{\gamma+r-1}{\gamma} q^\gamma, \tag{24}$$

$$p(d) = [1-q]^r. \tag{25}$$

We can then compute the choice-share ratios

$$\frac{p(n)}{p(d)} = \frac{qr}{n}, \tag{26}$$

$$\frac{p(n-1)}{p(n)} = 1 + \frac{q[r+1]}{n-1}; \tag{27}$$

allowing us to express the parameters as

$$q = [n-1]\left[\frac{p(n-1)}{p(n)} - 1\right] - \frac{np(n)}{p(d)}, \tag{28}$$

$$r = \frac{np(n)^2}{p(d)[n-1]\,[p(n-1) - p(n)] - np(n)^2}. \;\square \tag{29}$$

The final example introduces a consideration-probability distribution that may be useful in empirical applications of the ρ-model.[10] In this instance we obtain closed-form expressions for the cognitive parameters in two special cases, though not with complete generality.

Example 5. [*ρ-model; Kumaraswamy distribution*] For $a, b > 0$, let ρ have the Kumaraswamy [20] distribution $F(\rho) = 1 - [1 - \rho^a]^b$. If $b = 1$, then $F(\rho) = \rho^a$ and Equations 3–4 appear as

$$p(k) = a \int_0^1 \rho^a [1-\rho]^{k-1} \mathrm{d}\rho = aB(a+1, k), \tag{30}$$

$$p(d) = a \int_0^1 \rho^{a-1} [1-\rho]^n \mathrm{d}\rho = aB(a, n+1); \tag{31}$$

where B is the beta function.[11] The choice share for the best alternative is then $p(1) = \frac{a}{a+1}$, and we can conclude that $a = \frac{p(1)}{1-p(1)}$.

Alternatively, if $a = 1$ then $F(\rho) = 1 - [1-\rho]^{\,b}$ and the choice shares are

$$p(k) = b \int_0^1 \rho [1-\rho]^{\,k+b-2} \mathrm{d}\rho = bB(2, k+b-1) = \frac{b}{[k+b][k+b-1]}, \tag{32}$$

$$p(d) = b \int_0^1 [1-\rho]^{\,n+b-1} \mathrm{d}\rho = bB(1, n+b) = \frac{b}{n+b}. \tag{33}$$

From the default share we then obtain $b = \frac{np(d)}{1-p(d)}$.

In general Equation 4 takes the form

$$p(k) = ab \int_0^1 \rho^a [1-\rho]^{\,k-1} [1-\rho^a]^{b-1} \mathrm{d}\rho, \tag{34}$$

[10] An exercise of this sort using over-the-counter painkiller sales data is available from the authors upon request.

[11] Recall that for $y, z > 0$ the beta function is defined by $B(y,z) = \int_0^1 t^{y-1}[1-t]^{z-1}\mathrm{d}t$.

allowing the first two raw moments of F to be written as

$$m_1 = ab \int_0^1 \rho^a \left[1 - \rho^a\right]^{b-1} \mathrm{d}\rho = p(1), \tag{35}$$

$$m_2 = ab \int_0^1 \rho^{a+1} \left[1 - \rho^a\right]^{b-1} \mathrm{d}\rho = ab \int_0^1 [1 - [1 - \rho]]\rho^a \left[1 - \rho^a\right]^{b-1} \mathrm{d}\rho = p(1) - p(2). \tag{36}$$

This suggests that the difficulty of expressing the parameters in terms of the choice shares is primarily due to the difficulty of inverting the map $\langle a, b\rangle \mapsto \langle m_1, m_2\rangle$ for this functional form, rather than to any feature of the ρ-model itself. □

The observation that moments m_j of the Kumaraswamy distribution can be expressed as weighted sums of choice shares extends to values of $j > 2$, and is in fact a general feature of the consideration-probability model. This property is exploited in the nonparametric analysis of the ρ-model in Sections 3.4–3.5 below.

3.4 Cognitive identification with strict preferences: Nonparametric analysis

3.4.1 The nonparametric inference problem

The examples in Section 3.3 have demonstrated a variety of ways that information about the type distribution F can be encoded in the choice shares, depending on the cognitive model and the specific parameterization employed. With this introduction, we turn now to the general structure of the inference problem. We shall see that (with strict preferences) identification of the type distribution remains tractable in both models even without parametric assumptions. This is because the observed choice shares are linear functions of the probability masses $\pi(\gamma)$ in the γ-model and of the moments m_j of F in the ρ-model. Moreover, each system of equations has a simple triangular structure that enables it to be solved recursively, using one additional choice share at each step.

These features of the inference problem imply that under either of our cognitive models, the information encoded in the aggregate choice shares can be revealed by inverting a triangular $n \times n$ matrix. In contrast to Example 1, where maximal identification failed due to indifferences in the preference order, with strict preferences we can always fully exploit the available data to reveal n cardinal features of the type distribution. In the γ-model adding an alternative to the menu will yield an extra probability mass, while in the ρ-model it will yield an extra raw moment of F. In the latter case we can then use well-established tools (both maximum entropy methods and results from sparsity theory) to show that knowledge of the moments will allow us to approximate the distribution itself with increasing accuracy as the number of alternatives grows (see Section 3.5).

3.4.2 The γ-model: Recovering n probability masses

Without parametric assumptions on F, the aggregate choice shares in the γ-model are given by Equation 12. These relations can be written together in matrix form as

$$\underbrace{\begin{bmatrix} p(1) \\ \vdots \\ p(k) \\ \vdots \\ p(n) \end{bmatrix}}_{\mathbf{p}} = \underbrace{\begin{bmatrix} \frac{1}{n} & \frac{2}{n} & \cdots & \frac{\gamma}{n} & \cdots & 1 \\ \vdots & \vdots & & \vdots & & \\ \frac{1}{n} & \frac{2[n-k]}{n[n-1]} & \cdots & \frac{\binom{n-k}{\gamma-1}}{\binom{n}{\gamma}} & & 0 \\ \vdots & & & & & \vdots \\ \frac{1}{n} & 0 & \cdots & 0 & \cdots & 0 \end{bmatrix}}_{\mathbf{C}} \underbrace{\begin{bmatrix} \pi(1) \\ \pi(2) \\ \vdots \\ \pi(\gamma) \\ \vdots \\ \pi(n) \end{bmatrix}}_{\boldsymbol{\pi}}. \tag{37}$$

The upper anti-triangular matrix $\mathbf{C}$ has a lower anti-triangular inverse, allowing us to write $\boldsymbol{\pi} = \mathbf{C}^{-1}\mathbf{p}$. Indeed, we can compute the components of the vector $\boldsymbol{\pi}$ explicitly as

$$\pi(\gamma) = \binom{n}{\gamma} \sum_{k=n-\gamma+1}^{n} [-1]^{[\gamma-1]-[n-k]} \binom{\gamma-1}{n-k} p(k), \tag{38}$$

and of course $\pi(0) = p(d) = 1 - \sum_{k=1}^{n} p(k)$. Note that since by definition $\pi(n) = 1 - F(n-1)$, it is in fact the probabilities of the consideration capacities $\gamma = 0, 1, \ldots, n-1$ that are revealed; and $\gamma = n$ cannot be disambiguated from higher values. Indeed, all capacities greater than or equal to the number of alternatives will always be behaviorally indistinguishable. We summarize our conclusions for the consideration-capacity model as follows.

Proposition 2. *In the γ-model with strict preferences, the probability masses $\langle \pi(\gamma) \rangle_{\gamma=0}^{n-1}$ are uniquely determined by the aggregate choice shares $\langle p(k) \rangle_{k=1}^{n}$.*

3.4.3 The ρ-model: Recovering n raw moments

The aggregate choice shares in the ρ-model are given by Equation 4. Writing $m_j = \int_0^1 \rho^j \mathrm{d}F$ for the jth raw moment of the type distribution, we can expand the binomial in the choice-share formula to yield

$$p(k) = \int_0^1 \rho \left[\sum_{j=0}^{k-1} \binom{k-1}{j} [-\rho]^j \right] \mathrm{d}F = \sum_{j=1}^{k} (-1)^{j-1} \binom{k-1}{j-1} m_j. \tag{39}$$

Similarly, Equation 3 becomes

$$p(d) = \int_0^1 \left[\sum_{j=0}^{n} \binom{n}{j} [-\rho]^j \right] \mathrm{d}F = 1 + \sum_{j=1}^{n} [-1]^j \binom{n}{j} m_j. \tag{40}$$

The relations in Equation 39 can be written together in matrix form as

$$\underbrace{\begin{bmatrix} p(1) \\ \vdots \\ p(k) \\ \vdots \\ p(n) \end{bmatrix}}_{\mathbf{p}} = \underbrace{\begin{bmatrix} 1 & 0 & \cdots & 0 & \cdots & 0 \\ \vdots & & & & & \vdots \\ 1 & [-1][k-1] & \cdots & [-1]^{j-1}\binom{k-1}{j-1} & & 0 \\ \vdots & \vdots & & \vdots & & \\ 1 & [-1][n-1] & \cdots & [-1]^{j-1}\binom{n-1}{j-1} & \cdots & [-1]^{n-1} \end{bmatrix}}_{\mathbf{R}} \underbrace{\begin{bmatrix} m_1 \\ m_2 \\ \vdots \\ m_j \\ \vdots \\ m_n \end{bmatrix}}_{\mathbf{m}}. \tag{41}$$

The lower triangular matrix $\mathbf{R}$ is involutory (i.e., equal to its own inverse), allowing us to write $\mathbf{m} = \mathbf{R}\mathbf{p}$. Equivalently, the components of the vector $\mathbf{m}$ are given by

$$m_j = \sum_{k=1}^{j} [-1]^{k-1} \binom{j-1}{k-1} p(k). \tag{42}$$

We summarize our conclusions for the consideration-probability model as follows.

Proposition 3. *In the ρ-model with strict preferences, the raw moments $\langle m_j \rangle_{j=1}^{n}$ are uniquely determined by the aggregate choice shares $\langle p(k) \rangle_{k=1}^{n}$.*

As an aside, note that the binomial in Equation 13 can be expanded to yield

$$\begin{aligned}
\pi(\gamma) &= \int_0^1 \binom{n}{\gamma} \rho^{\gamma} [1-\rho]^{n-\gamma} \mathrm{d}F \\
&= \binom{n}{\gamma} \int_0^1 \rho^{\gamma} \left[\sum_{i=0}^{n-\gamma} \binom{n-\gamma}{i} [-\rho]^i \right] \mathrm{d}F \\
&= \binom{n}{\gamma} \sum_{i=0}^{n-\gamma} \binom{n-\gamma}{i} [-1]^i m_{\gamma+i} \\
&= \binom{n}{\gamma} \sum_{j=\gamma}^{n} \binom{n-\gamma}{j-\gamma} [-1]^{j-\gamma} m_j. \qquad (43)
\end{aligned}$$

In matrix form, the relations in Equation 43 appear as

$$\underbrace{\begin{bmatrix} \pi(1) \\ \vdots \\ \pi(\gamma) \\ \vdots \\ \pi(n) \end{bmatrix}}_{\boldsymbol{\pi}} = \underbrace{\begin{bmatrix} n & -n[n-1] & \cdots & n\binom{n-1}{j-1}[-1]^{j-1} & \cdots & n[-1]^{n-1} \\ & & & \vdots & & \vdots \\ 0 & 0 & & \binom{n}{\gamma}\binom{n-\gamma}{j-\gamma}[-1]^{j-\gamma} & \cdots & \binom{n}{\gamma}[-1]^{n-\gamma} \\ \vdots & \vdots & & & & \vdots \\ 0 & 0 & \cdots & 0 & & 1 \end{bmatrix}}_{\mathbf{Q}} \underbrace{\begin{bmatrix} m_1 \\ m_2 \\ \vdots \\ m_j \\ \vdots \\ m_n \end{bmatrix}}_{\mathbf{m}}. \qquad (44)$$

Combining Equations 37 and 44, we then have $\mathbf{p} = \mathbf{C}\boldsymbol{\pi} = \mathbf{C}[\mathbf{Qm}] = [\mathbf{CQ}]\mathbf{m}$. This is equivalent to the direct calculation of $\mathbf{p}$ in Equation 41, since it can be verified that $\mathbf{CQ} = \mathbf{R}$.

3.5 Beyond moments in the ρ-model

3.5.1 From moments to distributions

Throughout Section 3.5 we shall treat as known a finite number of moments of the type distribution F, appealing to Proposition 3 for justification. We proceed to outline two different strategies for ensuring that this moment information adequately captures the distribution itself. The first strategy relies upon discreteness of the distribution and guarantees a unique characterization of F, while the second relies upon differentiability and guarantees convergence to F in the limit (with respect to n).

3.5.2 Discrete distributions

Assume first that F is a discrete distribution, with ρ taking on values $\langle \rho_1, \rho_2, \ldots, \rho_L \rangle$. The number L of cognitive types is known, though the values themselves may be unknown. We assume, however, that the values must be located on a finite grid of admissible points in $[0, 1]$, which can be as fine as desired.

The realized values of ρ have probabilities $\langle \xi(\rho_1), \xi(\rho_2), \ldots, \xi(\rho_L) \rangle$, strictly positive and summing to one, so that the jth moment of F appears as

$$m_j = \sum_{\ell=1}^{L} \xi(\rho_\ell)\, \rho_\ell^j. \tag{45}$$

Since the first n moments are known, Equation 45 provides a system of n equalities in $2L$ unknowns; namely, the values ρ_ℓ and the associated probabilities $\xi(\rho_\ell)$. This system can be solved for n sufficiently large, but it is not obvious that the solution will be unique.

Assume now that the grid of admissible values of ρ is $\left\langle 0, \frac{1}{N}, \frac{2}{N}, \ldots, 1 \right\rangle$, with the fineness parameter N large relative to L.[12] In this case F is a discrete distribution defined entirely by the probability masses $\langle \xi(\frac{\ell}{N}) \rangle_{\ell=0}^{N}$, of which exactly $L \ll N$ are nonzero. Recovering the distribution then amounts to finding a solution of the system

[12] We use an evenly spaced grid of admissible values for notational simplicity, but this is not essential for our conclusions.

$$\underbrace{\begin{bmatrix} 1 \\ m_1 \\ \vdots \\ m_j \\ \vdots \\ m_n \end{bmatrix}}_{\bar{\mathbf{m}}} = \underbrace{\begin{bmatrix} 1 & 1 & \cdots & 1 & \cdots & 1 \\ 0 & \frac{1}{N} & \cdots & \frac{\ell}{N} & \cdots & 1 \\ \vdots & \vdots & & \vdots & & \vdots \\ 0 & [\frac{1}{N}]^j & \cdots & [\frac{\ell}{N}]^j & \cdots & 1 \\ \vdots & \vdots & & \vdots & & \vdots \\ 0 & [\frac{1}{N}]^n & \cdots & [\frac{\ell}{N}]^n & \cdots & 1 \end{bmatrix}}_{\mathbf{V}} \underbrace{\begin{bmatrix} \xi(0) \\ \xi\left(\frac{1}{N}\right) \\ \vdots \\ \xi\left(\frac{\ell}{N}\right) \\ \vdots \\ \xi(1) \end{bmatrix}}_{\boldsymbol{\xi}}, \tag{46}$$

with all components of the solution vector $\boldsymbol{\xi}$ weakly positive and exactly L components strictly positive. Here $\mathbf{V}$ is a Vandermonde matrix with many more columns (grid points) than rows (known moments), implying an underdetermined system.[13] But the number L of grid points actually used could in principle be larger or smaller than n.

A result of Cohen and Yeredor [12, Theorem 1] applies to precisely this situation, stating that Equation 46 has a unique solution whenever $n \geq 2L$. We thus conclude the following.

Proposition 4. *In the ρ-model with strict preferences, if F is a discrete distribution over L admissible types, with $n \geq 2L$, then it is uniquely determined by the aggregate choice shares $\langle p(k) \rangle_{k=1}^{n}$.*

That is to say, for practical purposes any discrete distribution for ρ can be fully recovered from aggregate choice data provided the number of alternatives is large compared to the number of cognitive types.

3.5.3 Differentiable distributions

Now assume that the type distribution F possesses a probability density f. In this case we will not be able to fully recover the distribution from the first n moments. Instead, we wish to ensure that the known moments yield a reliable approximation of the true distribution.

Our analysis relies on standard techniques from the "Hausdorff moment problem" for distributions on a closed interval. Adopting a maximum entropy approach, define

[13] See, e.g., Macon and Spitzbart [24] for the definition and properties of Vandermonde matrices.

the nth approximating density f_n as the solution to the optimization problem

$$\max_{f_n} \int_0^1 \left[-\log f_n(\rho)\right] f_n(\rho)\,\mathrm{d}\rho \tag{47}$$

subject to the jth-moment constraint

$$\int_0^1 \rho^j f_n(\rho)\,\mathrm{d}\rho = m_j \tag{48}$$

for each $j = 0, 1, \ldots, n$. Mead and Papanicolaou [31, Theorem 2] establish that a solution to this problem exists and is unique.[14] Moreover, for each continuous map $\psi : [0,1] \to \Re$ we have

$$\lim_{n\to\infty} \int_0^1 \psi(\rho)\, f_n(\rho)\,\mathrm{d}\rho = \int_0^1 \psi(\rho)\, f(\rho)\mathrm{d}\rho. \tag{49}$$

Write F_n for the distribution function associated with the approximating density f_n.

Observe now that for any menu A and each $k \le \min\{n, |A|\}$, we have

$$p(k_A; F_n) = p(k_M; F_n) = p(k_M; F) = p(k_A; F). \tag{50}$$

Here the first and third equalities follow from the fact that in the ρ-model an alternative's choice share depends only on its position on the menu according to the preference order. Moreover, the shares of the n best alternatives are determined by the first n moments (see Equation 41), which coincide for F and F_n (see Equation 48). This yields the second equality above, and we can summarize our findings as follows.

Proposition 5. *In the ρ-model with strict preferences, if F is differentiable then there exists a sequence $\langle F_n \rangle_{n=1}^{\infty}$ of distributions such that:* (i) *each F_n is defined by $\langle m_j \rangle_{j=1}^{n}$;* (ii) *$F_n$ converges weakly to F; and* (iii) *for each menu A and $k \le \min\{n, |A|\}$ we have $p(k_A; F_n) = p(k_A; F)$.*

Equation 48 ensures that each approximation F_n is observationally indistinguishable from the true F in the sense that the two distributions generate the same first n

[14] Indeed, the solution takes the form $f_n(\rho) = \exp[-\sum_{j=0}^{n} \lambda_j \rho^j]$, where the quantities $\langle \lambda_j \rangle_{j=0}^{n}$ are the Lagrange multipliers on the constraints in Equation 48.

moments, and hence the same aggregate choice shares over the observed menu M. Proposition 5 reinforces this conclusion by guaranteeing that the cognitive heterogeneity in the population is accurately reflected in two additional ways: First, as the size of the observed menu increases, the resulting approximations approach the true distribution in the sense of weak convergence. And second, for any menu size n the approximation F_n matches the true F not just over M, but also over the n best alternatives on any other menu A about which we may wish to make predictions.

4 Preference heterogeneity

4.1 Known taste distribution

Section 3 studied the theoretical identification properties of our two models of consideration set formation under the assumption that preferences are unobserved but homogeneous. In this section, we aim to show that our analysis can be extended to allow for preference heterogeneity, provided the taste distribution is known and statistically independent of the cognitive distribution.[15] We then proceed in Section 4.2 to consider how our models can be used when the taste distribution is itself unknown.

To incorporate preference heterogeneity into the present framework, order the alternatives arbitrarily as $M = \{1, 2, \ldots, n\}$ and write $\varphi : M \to \{1, 2, \ldots, n\}$ for the map that associates each option with its preference rank.[16] We enumerate the possible rankings as $\langle \varphi_h \rangle_{h=1}^{n!}$, write τ_h for the probability of ranking φ_h, and denote by $\mathbf{P}(h)$ the $n \times n$ permutation matrix corresponding to ranking φ_h.[17] In the context of the γ-model,

[15] The distribution of taste parameters—such as discount factors or risk-aversion coefficients—may be treated as known for our purposes if these characteristics can be elicited from agents separately, in a setting (e.g., a laboratory experiment) where limited attention is thought to be irrelevant or controllable to an acceptable degree.

[16] Note that this formulation imposes the assumption of strict preferences maintained throughout Sections 3.3–3.5.

[17] More explicitly, the permutation matrix $\mathbf{P}(h)$ translates the kth row of any $n \times n$ target matrix $\mathbf{A}$ into the $\varphi_h(k)$th row of the product $\mathbf{P}(h)\mathbf{A}$.

Equation 37 then becomes

$$\mathbf{p} = \sum_{h=1}^{n!} \tau_h [\mathbf{P}(h)\mathbf{C}]\boldsymbol{\pi} = \left[\sum_{h=1}^{n!} \tau_h \mathbf{P}(h)\mathbf{C} \right] \boldsymbol{\pi}. \tag{51}$$

And similarly, in the context of the ρ-model, Equation 41 becomes

$$\mathbf{p} = \sum_{h=1}^{n!} \tau_h [\mathbf{P}(h)\mathbf{R}]\mathbf{m} = \left[\sum_{h=1}^{n!} \tau_h \mathbf{P}(h)\mathbf{R} \right] \mathbf{m}. \tag{52}$$

We conclude that it is still possible to use the choice shares to compute the vectors $\boldsymbol{\pi}$ of probability masses and $\mathbf{m}$ of raw moments, as long as the corresponding matrices $\sum_{h=1}^{n!} \tau_h \mathbf{P}(h)\mathbf{C}$ and $\sum_{h=1}^{n!} \tau_h \mathbf{P}(h)\mathbf{R}$ are nonsingular.

The following example illustrates the handling of known preference heterogeneity for the case of the γ-model.

Example 6. [*γ-model; exploded logit*] Let $n = 3$, define $u : M \to \Re$ by $u(k) = \log k$, and suppose that the distribution of tastes is determined by an exploded logit based on u (see, e.g., Luce and Suppes [23]). For instance, the probability assigned to the ranking φ_2 given by $2 \succ 3 \succ 1$ is

$$\tau_2 = \frac{e^{u(2)}}{e^{u(1)} + e^{u(2)} + e^{u(3)}} \times \frac{e^{u(3)}}{e^{u(1)} + e^{u(3)}} \times \frac{e^{u(1)}}{e^{u(1)}} = \frac{2}{1+2+3} \times \frac{3}{1+3} \times \frac{1}{1} = \frac{1}{4}. \tag{53}$$

Under the γ-model we then have

$$\begin{aligned}
\sum_{h=1}^{6} \tau_h \mathbf{P}(h)\mathbf{C} = {} & \frac{1}{3} \underbrace{\begin{bmatrix} \frac{1}{3} & 0 & 0 \\ \frac{1}{3} & \frac{1}{3} & 0 \\ \frac{1}{3} & \frac{2}{3} & 1 \end{bmatrix}}_{\varphi_1:\, 3\succ 2\succ 1} + \frac{1}{4} \underbrace{\begin{bmatrix} \frac{1}{3} & 0 & 0 \\ \frac{1}{3} & \frac{2}{3} & 1 \\ \frac{1}{3} & \frac{1}{3} & 0 \end{bmatrix}}_{\varphi_2:\, 2\succ 3\succ 1} + \frac{1}{6} \underbrace{\begin{bmatrix} \frac{1}{3} & \frac{1}{3} & 0 \\ \frac{1}{3} & 0 & 0 \\ \frac{1}{3} & \frac{2}{3} & 1 \end{bmatrix}}_{\varphi_3:\, 3\succ 1\succ 2} \cdots \\
& \cdots + \frac{1}{10} \underbrace{\begin{bmatrix} \frac{1}{3} & \frac{2}{3} & 1 \\ \frac{1}{3} & 0 & 0 \\ \frac{1}{3} & \frac{1}{3} & 0 \end{bmatrix}}_{\varphi_4:\, 1\succ 3\succ 2} + \frac{1}{12} \underbrace{\begin{bmatrix} \frac{1}{3} & \frac{1}{3} & 0 \\ \frac{1}{3} & \frac{2}{3} & 1 \\ \frac{1}{3} & 0 & 0 \end{bmatrix}}_{\varphi_5:\, 2\succ 1\succ 3} + \frac{1}{15} \underbrace{\begin{bmatrix} \frac{1}{3} & \frac{2}{3} & 1 \\ \frac{1}{3} & \frac{1}{3} & 0 \\ \frac{1}{3} & 0 & 0 \end{bmatrix}}_{\varphi_6:\, 1\succ 2\succ 3} = \begin{bmatrix} \frac{1}{3} & \frac{7}{36} & \frac{1}{6} \\ \frac{1}{3} & \frac{16}{45} & \frac{1}{3} \\ \frac{1}{3} & \frac{9}{20} & \frac{1}{2} \end{bmatrix},
\end{aligned} \tag{54}$$

which is a nonsingular matrix (with determinant $\frac{1}{270}$). □

To recover the probability masses in the γ-model, we need the matrix $[\sum_{h=1}^{n!} \tau_h \mathbf{P}(h)]\mathbf{C}$ to be nonsingular. Since $\mathbf{C}$ has full rank, this amounts to $\sum_{h=1}^{n!} \tau_h \mathbf{P}(h)$ being nonsingular. The latter is a convex combination of permutation matrices, or equivalently a bistochastic matrix.[18] Clearly this matrix is not invertible for every distribution $\tau = \langle \tau_h \rangle_{h=1}^{n!}$; for instance, if each $\tau_h = \frac{1}{n!}$ then each entry of $\sum_{h=1}^{n!} \tau_h \mathbf{P}(h)$ is $\frac{1}{n}$ and the matrix is singular. However, this situation is nongeneric.

Proposition 6. *In the γ-model with strict preferences and observed preference heterogeneity, for almost all distributions τ of preferences the probability masses $\langle \pi(\gamma) \rangle_{\gamma=0}^{n-1}$ are uniquely determined by the aggregate choice shares $\langle p(k) \rangle_{k=1}^{n}$.*

Proposition 6 follows from our earlier results combined with a pair of simple observations. Firstly, the determinant operator is a polynomial on the affine subspace of $\Re^{n \times n}$ containing the so-called "Birkhoff polytope" of bistochastic matrices. And secondly, any real-valued (and nonconstant) polynomial function on a Euclidean space is nonzero almost everywhere (see, e.g., Caron and Traynor [9]). Furthermore, the same logic applies equally well to the ρ-model (substituting the matrix $\mathbf{R}$ for $\mathbf{C}$), yielding the following counterpart result.

Proposition 7. *In the ρ-model with strict preferences and observed preference heterogeneity, for almost all distributions τ of preferences the raw moments $\langle m_j \rangle_{j=1}^{n}$ are uniquely determined by the aggregate choice shares $\langle p(k) \rangle_{k=1}^{n}$.*

Hence we can conclude that under preference heterogeneity Propositions 2–3 continue to hold (generically), provided the distribution of tastes is known.

4.2 Unknown taste distribution

In this section we turn to the problem of identifying cognitive and preference heterogeneity simultaneously, when the taste distribution cannot be treated as known. This

[18] A matrix is said to be *bistochastic* if each of its row and columns contain nonnegative entries that sum to one. The Birkhoff-von-Neumann theorem states that the set of bistochastic matrices equals the convex hull of the set of permutation matrices.

will be a common situation in empirical applications of the present framework. Accordingly, the treatment of this case will move the discussion towards an econometric specification of our models.

With unknown tastes, the information contained in aggregate choices is insufficient to reveal the distributions of both preference and cognitive characteristics nonparametrically. Indeed, if we use this dataset to infer n probability masses or moments of F, as in Section 3.4, then no degrees of freedom remain to pin down cardinal features of the taste distribution. In a parameterized setting like that of Section 3.3, on the other hand, we can in principle use the choice shares to identify both the cognitive and preference distributions as long as the map from parameter vectors to datasets remains one-to-one. This joint elicitation is illustrated in the following example.

Example 7. [*γ-model; Poisson distribution*] Let the consideration capacity be distributed as in Example 3, and for $q \in [0,1]$ let the preference order be $1 \succ 2 \succ \cdots \succ n$ with probability q and $1 \succ 2 \succ \cdots \succ n-2 \succ n \succ n-1$ with probability $1-q$. As before, we have the default share $p\,(d) = e^{-\mu}$ and hence the cognitive parameter $\mu = -\log p(d)$. Moreover, we can compute

$$p(n-1) = \left[\frac{\mu}{n} + q\frac{\mu^2}{n[n-1]}\right] e^{-\mu}, \tag{55}$$

$$p(n) = \left[\frac{\mu}{n} + [1-q]\frac{\mu^2}{n[n-1]}\right] e^{-\mu}. \tag{56}$$

Combining these equations and substituting for μ then yields an expression

$$q = \frac{1}{2}\left[\frac{p(n-1)-p(n)}{p(d)} \times \frac{n[n-1]}{[\log p(d)]^2} + 1\right] \tag{57}$$

for the taste parameter in terms of observable choice shares. □

We proceed now to examine a parameterized model appropriate for empirical applications, in which the preference distribution is generated by random utility. Specifically, suppose that agent i's utility for alternative k is given by

$$u_{ik} = \sum_{j=1}^{J} \beta_j x_{kj} + \epsilon_{ik}; \tag{58}$$

where each x_{kj} is a characteristic of the alternative, β_j is the associated preference parameter, and the error term ϵ_{ik} is extreme value (i.e., standard Gumbel) distributed independently across agents and alternatives. In this setting we observe the characteristics and the aggregate choice shares, and we wish to estimate both the preference parameters and the attention distribution.

Adopting the γ-model, we can write the analog of Equation 5 for the present scenario as

$$p_\gamma(k) = \sum_{r=1}^{n} \frac{\binom{n-r}{\gamma-1}}{\binom{n}{\gamma}} \sum_{h:\varphi_h(k)=r} \tau_h. \tag{59}$$

Here the utility realizations could place alternative k in any ordinal position r, and the inner sum is over the rankings h that do this. Writing $\mathcal{A}_\gamma(k) = \{A \subseteq M : k \in A \wedge |A| = \gamma\}$ for the collection of subsets of the menu that have cardinality γ and include alternative k, we can use Equation 58 to express the probability in Equation 59 as

$$p_\gamma(k) = \frac{1}{\binom{n}{\gamma}} \sum_{A \in \mathcal{A}_\gamma(k)} \frac{\exp \sum_j \beta_j x_{kj}}{\sum_{\ell \in A} \exp \sum_j \beta_j x_{\ell j}}. \tag{60}$$

We then compute

$$\frac{\partial p_\gamma(k)}{\partial x_{kj}} = \frac{\beta_j}{\binom{n}{\gamma}} \sum_{A \in \mathcal{A}_\gamma(k)} \frac{[\exp \sum_{\bar{j}} \beta_{\bar{j}} x_{k\bar{j}}][\sum_{\ell \in A \setminus k} \exp \sum_{\bar{j}} \beta_{\bar{j}} x_{\ell \bar{j}}]}{[\sum_{\ell \in A} \exp \sum_{\bar{j}} \beta_{\bar{j}} x_{\ell \bar{j}}]^2}, \tag{61}$$

and it follows that

$$\frac{\partial p(k)/\partial x_{kj}}{\partial p(k)/\partial x_{k1}} = \frac{\sum_{\gamma=1}^{n} \frac{\partial p_\gamma(k)}{\partial x_{kj}} \pi(\gamma)}{\sum_{\gamma=1}^{n} \frac{\partial p_\gamma(k)}{\partial x_{k1}} \pi(\gamma)} = \frac{\beta_j}{\beta_1} \tag{62}$$

for each characteristic (indexed by j) of alternative k. Therefore, by observing the *changes* in the aggregate choice shares induced by changes in the characteristics, we are able to identify the corresponding preference parameters up to a scaling factor.[19] We can then use the *levels* of the shares to reveal the attention distribution, via the methods described in Section 3. Observe finally that the preceding analysis applies equally well to the ρ-model, since it is a particular case of the γ-model (see Section 2.4).

5 Monte Carlo simulation of cognitive heterogeneity

5.1 The attention models with logit tastes

In this section we demonstrate how our two models of cognitive heterogeneity can be put to use for applied purposes. We present a pair of Monte Carlo simulations with data generated by the γ- and ρ-models, and show that in both cases an econometric implementation of our theoretical results succeeds in retrieving the true parameters. In both of these exercises we use a logit random utility specification for preferences, as in Equation 58, while the distribution of cognitive parameters is retrieved nonparametrically from the aggregate choice shares.

The models employed in our simulations can be described more formally as follows. The "γ-logit" model takes the form

$$\underbrace{\begin{bmatrix} p\,(d) \\ p\,(1) \\ \vdots \\ p\,(n) \end{bmatrix}}_{\overline{\mathbf{p}}} = \underbrace{\begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & p_1(1) & \cdots & p_n(1) \\ \vdots & \vdots & \ddots & \vdots \\ 0 & p_1(n) & \cdots & p_n(n) \end{bmatrix}}_{\mathbf{D}} \underbrace{\begin{bmatrix} \pi\,(0) \\ \pi\,(1) \\ \vdots \\ \pi\,(n) \end{bmatrix}}_{\overline{\boldsymbol{\pi}}}, \tag{63}$$

where each $p_\gamma\,(k)$ is a choice probability from Equation 60. We can use $\mathbf{Q}$ from Equa-

[19] Note that the relations in Equation 62 hold separately for each alternative k, so the preference parameters are overidentified to a degree proportional to n. This creates an avenue for testing our models of cognition (as in, e.g., Abaluck and Adams [1]).

tion 44 to construct the augmented matrix

$$\overline{\mathbf{Q}} = \begin{bmatrix} 1 & \nu' \\ \mathbf{0} & \mathbf{Q} \end{bmatrix}, \tag{64}$$

where $\mathbf{0}$ is the $n \times 1$ zero vector and $\nu = \langle [-1]^j \binom{n}{j} \rangle_{j=1}^n$. Recalling that $\overline{\mathbf{m}} = \langle 1, m_1, \ldots, m_n \rangle$, we then have that $\overline{\boldsymbol{\pi}} = \overline{\mathbf{Q}}\overline{\mathbf{m}}$ by Equations 40 and 44 and the fact that $p(d) = \pi(0)$. The "ρ-logit" model now appears as

$$\overline{\mathbf{p}} = \mathbf{D}\overline{\boldsymbol{\pi}} = \mathbf{D}[\overline{\mathbf{Q}}\overline{\mathbf{m}}] = [\mathbf{D}\overline{\mathbf{Q}}]\overline{\mathbf{m}}. \tag{65}$$

Identification of the logit parameters $\langle \beta_j \rangle_{j=1}^J$ in the matrix $\mathbf{D}$ is discussed in Section 4.2. The attention distribution $\overline{\boldsymbol{\pi}}$ in Equation 63 is then (generically) identified by Proposition 6, and the moment distribution $\overline{\mathbf{m}}$ follows from $\overline{\boldsymbol{\pi}}$ since $\overline{\mathbf{Q}}$ is invertible.

It can be observed that Equation 63 is an example of a finite mixture model (FMM), where the weights are given by the capacity probabilities in $\overline{\boldsymbol{\pi}}$. Note also that the full-attention logit model is the limiting case with choice probabilities given by the last column of $\mathbf{D}$, where all n options are considered with certainty.

In our simulations we set $n = 4$ (alternatives) and $J = 2$ (characteristics). We construct an economy comprised of one hundred "markets," each populated by 10 000 "individuals." In each market m we normalize $x_{11}^m = x_{12}^m = 0$ and choose the (two) characteristics of the other (three) alternatives by independent Gaussian draws from

$$x_{21}^m, x_{32}^m \sim N(0,1), \tag{66}$$

$$x_{22}^m, x_{41}^m, x_{42}^m \sim N(1,1), \tag{67}$$

$$x_{31}^m \sim N(2,1). \tag{68}$$

Preferences have the logit form in Equation 58, with taste parameters $\langle \beta_1, \beta_2 \rangle = \langle 1, 1 \rangle$.

5.2 The γ-logit model: Estimation

Our first simulation employs the γ-model of limited attention, with the consideration capacity distribution given by $\pi(\gamma) = \frac{1}{5}$ for $\gamma = 0, 1, \ldots, 4$. To generate a sample for market m, we draw the cognitive type γ_i of each individual i independently from the capacity distribution and use Equation 60 to calculate the associated choice probabilities $p^m_{\gamma_i}(k)$ for $k = 1, \ldots, 4$. (Of course, $p^m_{\gamma_i}(d) = 1$ if and only if $\gamma_i = 0$.) These probabilities are used to draw individual i's chosen alternative, and we then sum over the population of market m to obtain the aggregate choice shares. This procedure is repeated to generate five separate simulated samples.

Each simulated sample consists of the characteristics $\langle x^m_{21}, x^m_{22}, x^m_{31}, x^m_{32}, x^m_{41}, x^m_{42}\rangle^{100}_{m=1}$ and the aggregate choice shares $\langle p^m(d), p^m(1), p^m(2), p^m(3), p^m(4)\rangle^{100}_{m=1}$. The parameters $\langle \beta_1, \beta_2\rangle$ and $\langle \pi(\gamma)\rangle^4_{\gamma=0}$ are estimated robustly by means of the expectation-maximization (EM) algorithm.[20] Across all samples the estimated taste and attention parameters are very close to the true values, as indicated by the bias and root mean squared error (RMSE) figures reported in Table 1.

	β_1	β_2	$\pi(0)$	$\pi(1)$	$\pi(2)$	$\pi(3)$	$\pi(4)$
value	1.0000	1.0000	0.2000	0.2000	0.2000	0.2000	0.2000
bias	0.0019	0.0014	0.0000	0.0005	−0.0009	0.0004	0.0000
RMSE	0.0152	0.0150	0.0012	0.0046	0.0070	0.0096	0.0080

Table 1: Estimation of the simulated γ-logit model with $\pi(\gamma) = \frac{1}{5}$ for $\gamma = 0, 1, \ldots, 4$.

5.3 The ρ-logit model: Estimation

To simulate the ρ-model, we assume that $\rho \sim \text{Beta}(3, 3)$ and in each market m draw the cognitive type ρ_i of each individual i independently from this distribution. The associated choice probabilities are then calculated as $p^m_{\rho_i}(k) = \sum^4_{\gamma_i=0} \binom{n}{\gamma_i} \rho_i^{\gamma_i} [1 - \rho_i]^{n-\gamma_i} p^m_{\gamma_i}(k)$ for $k = 1, \ldots, 4$. The chosen alternative is drawn from these probabilities, and summing yields the aggregate shares in market m. Once again we generate five separate simulated samples.

[20]See McLachlan and Peel [30] for details of finite mixture models and the EM algorithm.

We estimate $\langle \beta_1, \beta_2 \rangle$ and $\langle \pi(\gamma) \rangle_{\gamma=0}^{4}$ as above using the EM algorithm, and then proceed to compute the estimated moments via $\overline{\mathbf{m}} = \overline{\mathbf{Q}}^{-1} \overline{\boldsymbol{\pi}}$. Table 2 reports bias and RMSE figures for these estimates, showing that both the taste parameters and the moments of the cognitive distribution are retrieved accurately.

	β_1	β_2	m_1	m_2	m_3	m_4
value	1.0000	1.0000	0.5000	0.2857	0.1786	0.1190
bias	-0.0005	0.0001	0.0001	0.0001	0.0002	0.0002
RMSE	0.0074	0.0071	0.0013	0.0017	0.0017	0.0014

Table 2: Estimation of the simulated ρ-logit model with $\rho \sim$ Beta $(3, 3)$.

With the estimated moments of ρ in hand, for each simulated sample we can use maximum entropy methods to approximate the full cognitive distribution and compare this to the underlying distribution Beta $(3, 3)$. These pairs of (estimated and true) distributions are shown in the five panels of Figure 1, and demonstrate the accuracy of the recovery procedure.[21]

5.4 Related empirical literature

Our simulation exercise contributes to a growing literature on the estimation of consideration sets from demand data. Sovinsky Goeree [40] investigates the impact of marketing on the consideration set, using advertising data to separate utility and attentional components of demand. Van Nierop et al. [32] propose a model of brand choice that accommodates both stated and revealed consideration-set data, and apply this framework to an online experiment simulating a variety of merchandising strategies. Abaluck and Adams [1] build a very general econometric framework that exploits asymmetries in the matrix of cross-partial derivatives to identify consideration-set effects. Barseghyan et al. [4] develop a model of consideration set in a consumer choice setting under risk. The model is very general in that it makes no assumption on the consideration set formation mechanism, yet obtain a partial identification of the con-

[21]Here we employ the MaxEnt algorithm developed in Rajan et al. [34], which also discusses the relative merits of various approaches to solving the moment problem computationally. Note that MaxEnt *estimates* the support of the distribution, rather than imposing it, which explains the small but positive densities outside of $[0, 1]$ in Figure 1.

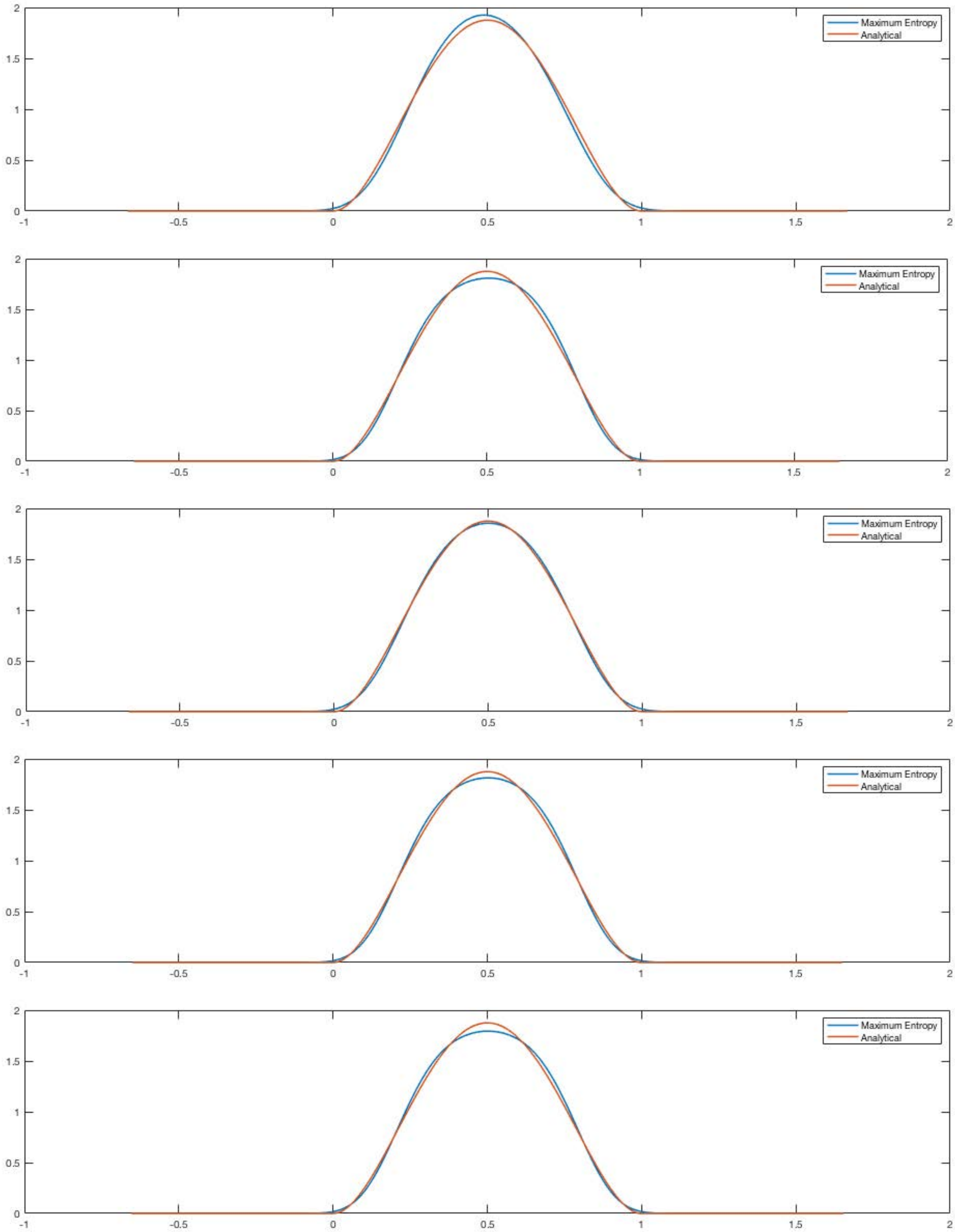

Figure 1: Estimated (blue) and true (red) probability density functions for the cognitive distribution, computed for each of the five simulated samples generated from the ρ-logit model.

sideration sets. Barseghyan et al. [5] propose a class of consideration set models based on Manzini and Mariotti [26] specific to choice under risk, and show that with appropriate restrictions the model is semi-parametrically identified. Crawford et al. [13] devise a model-free identification strategy based on reducing the menu of alternatives to a "sufficient set" of those that are certain to be considered. Lu [21] describes an approach to estimating multinomial choice models that employs known upper and lower bounds on the consideration set. Honka et al. [19], among others, model consideration sets as the outcome of a search process, while Gaynor et al. [17] exploit institutional changes to identify consideration sets in hospital choice.[22]

Our exercise is distinct from this literature in that we use a different identification strategy: We rely on our theoretical results to establish correspondences between the observed choice shares and the unobserved cognitive parameter (i.e., the consideration probability or capacity). This enables us to retrieve the type distribution from aggregate choice data via raw moments (in the ρ-model) or probability masses (in the γ-model).

6 Concluding comments

This paper contributes to the theoretical literature on boundedly rational decision making by outlining a methodology for inferring the distribution of cognitive characteristics in a population using aggregate choice data. A major advantage of our approach is that it assumes a fixed menu of alternatives. In contrast, much earlier work in this area assumes knowledge of a single individual's choices from a family of overlapping menus. While both theoretical frameworks yield results that can be brought to bear on data, our view is that the fixed-menu approach is closer to the practice of empirical research on discrete choice. We also show that our results can be adapted for estimation in the context of simulated choice data, in order to validate the tractability of the methodology for empirical work.

[22] The search literature typically deals with datasets that include information about the composition of a consumer's consideration set, though there are exceptions. For example, in Hastings et al. [18] exposure to a sales force influences the probability that financial products are considered.

A second message of the paper is that both the "consideration probability" ρ-model and the "consideration capacity" γ-model are surprisingly tractable within the fixed-menu framework. In both models the aggregate choice shares are linear in quantities that are highly informative about the cognitive distribution; namely, low-cardinality choice set probabilities in the γ-model and low-order raw moments in the ρ-model. These systems are recursive—provided all preferences are strict—and easily solved for the quantities in question. Indeed, our theoretical results show that for large menus the cognitive distribution is essentially fully identified, while for smaller menus we can still infer substantial useful information (and typically the full distribution in parameterized settings).

Finally, we mention three possible ways to build on the work reported in this paper. One is to generalize the models of consideration set formation that we have studied; for example, by allowing non-uniform consideration probabilities in the ρ-model, or by relaxing the assumption that all consideration sets with the same cardinality are equally likely to occur in the γ-model.[23] Another is to bring additional models of bounded rationality—incorporating phenomena such as computational constraints and reference points—into the present framework. And a third is to enrich the econometric specification used in our simulation exercise, allowing more precise control of the interaction between cognitive and taste heterogeneity.

References

[1] Jason Abaluck and Abi Adams (2017). What do consumers consider before they choose? Identification from asymmetric demand responses. Unpublished.

[2] Jose Apesteguia and Miguel A. Ballester (2013). Choice by sequential procedures. *Games and Economic Behavior* 77:90–99.

[3] Nicholas Baigent and Wulf Gaertner (1996). Never choose the uniquely largest: A characterization. *Economic Theory* 8:239–249.

[23] On the extension to non-uniform consideration probabilities, see Brady and Rehbeck [6].

[4] Levon Barseghyan, Maura Coughlin, Francesca Molinari, and Joshua Teitelbaum (2018). Heterogeneous consideration sets and preferences. Unpublished.

[5] Levon Barseghyan, Francesca Molinari, and Matthew Thirkettle (2018). Discrete Choice Under Risk With Limited Consideration. Unpublished.

[6] Richard L. Brady and John Rehbeck (2016). Menu-dependent stochastic feasibility. *Econometrica* 84:1203–1223.

[7] Andrew Caplin and Mark Dean (2015). Revealed preference, rational inattention, and costly information acquisition. *American Economic Review* 105:2183–2203.

[8] Andrew Caplin, Mark Dean, and Daniel Martin (2011). Search and satisficing. *American Economic Review* 101:2899–2922.

[9] Richard Caron and Tim Traynor (2005). The zero set of a polynomial. Unpublished.

[10] Matias D. Cattaneo, Xinwei Ma, Yusufcan Masatlioglu, and Elchin Suleymanov (2017). A random attention model. Unpublished.

[11] Vadim Cherepanov, Timothy Feddersen, and Alvaro Sandroni (2013). Rationalization. *Theoretical Economics* 8:775–800.

[12] Anna Cohen and Arie Yeredor (2011). On the use of sparsity for recovering discrete probability distributions from their moments. *Proceedings of the 2011 IEEE Statistical Signal Processing Workshop*.

[13] Gregory S. Crawford, Rachel Griffith, and Alessandro Iaria (2018). Preference estimation with unobserved choice set heterogeneity using sufficient sets. Unpublished.

[14] Geoffroy de Clippel, Kfir Eliaz, and Kareen Rozen (2014). Competing for consumer inattention. *Journal of Political Economy* 122:1203–1234.

[15] Henrique de Oliveira, Tommaso Denti, Maximilian Mihm, and Kemal Ozbek (2017). Rationally inattentive preferences and hidden information costs. *Theoretical Economics* 12:621–654.

[16] Federico Echenique, Kota Saito, and Gerelt Tserenjigmid (2017). The perception-adjusted Luce model. *Mathematical Social Sciences*, forthcoming.

[17] Martin Gaynor, Carol Propper, and Stephan Seiler (2016). Free to choose? Reform, choice, and consideration sets in the English National Health Service. *American Economic Review* 106:3521–3557.

[18] Justine S. Hastings, Ali Hortaçsu, and Chad Syverson (2017). Sales force and competition in financial product markets: The case of Mexico's social security privatization. *Econometrica* 85:1723–1761.

[19] Elisabeth Honka, Ali Hortaçsu, and Maria Ana Vitorino (2017). Advertising, consumer awareness, and choice: Evidence from the U.S. banking industry. *Rand Journal of Economics* 48:611–646.

[20] P. Kumaraswamy (1980). A generalized probability density function for double-bounded random processes. *Journal of Hydrology* 46:79–88.

[21] Zhentong Lu (2016). Estimating multinomial choice models with unobserved choice sets. Unpublished.

[22] R. Duncan Luce (1959). *Individual Choice Behavior: A Theoretical Analysis*. Wiley.

[23] R. Duncan Luce and Patrick C. Suppes (1965). Preferences, utility and subjective probability. In: R. Duncan Luce, Robert R. Bush, and Eugene Galanter, eds., *Handbook of Mathematical Psychology*, Wiley.

[24] Nathaniel Macon and Abraham Spitzbart (1958). Inverses of Vandermonde matrices. *The American Mathematical Monthly* 65:95–100.

[25] Paola Manzini and Marco Mariotti (2007). Sequentially rationalizable choice. *American Economic Review* 97:1824–1839.

[26] Paola Manzini and Marco Mariotti (2014). Stochastic choice and consideration sets. *Econometrica* 82:1153–1176.

[27] Yusufcan Masatlioglu and Daisuke Nakajima (2013). Choice by iterative search. *Theoretical Economics* 8:701–728.

[28] Yusufcan Masatlioglu, Daisuke Nakajima, and Erkut Y. Ozbay (2012). Revealed attention. *American Economic Review* 102:2183–2205.

[29] Daniel L. McFadden (2001). Economic choices. *American Economic Review* 91:351–378.

[30] Geoffrey McLachlan and David Peel (2000). *Finite Mixture Models*. Wiley.

[31] Laurence R. Mead and Nikos Papanicolaou (1984). Maximum entropy in the problem of moments. *Journal of Mathematical Physics* 25:2404–2417.

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

[33] Efe A. Ok, Pietro Ortoleva, and Gil Riella (2014). Revealed (p)reference theory. *American Economic Review* 105:299–321.

[34] Arvind Rajan, Ye Chow Kuang, Melanie Po-Leen Ooi, Serge N. Demidenko, and Herman Carstens (2018). Moment-constrained maximum entropy method for expanded uncertainty evaluation. *IEEE Access* 6:4072–4082.

[35] John H. Roberts and James M. Lattin (1991). Development and testing of a model of consideration set composition. *Journal of Marketing Research* 28:429–440.

[36] John H. Roberts and James M. Lattin (1997). Consideration: Review of research and prospects for future insights. *Journal of Marketing Research* 34:406–410.

[37] Yuval Salant and Ariel Rubinstein (2008). (A, f): Choice with frames. *Review of Economic Studies* 75:1287–1296.

[38] Allan Shocker, Moshe Ben-Akiva, Bruno Boccara, and Prakash Nedungadi (1991). Consideration set influences on consumer decision making and choice: Issues, models, and suggestions. *Marketing Letters* 2:181–198.

[39] Christopher A. Sims (2003). Implications of rational inattention. *Journal of Monetary Economics* 50:665–690.

[40] Michelle Sovinsky Goeree (2008). Limited information and advertising in the U.S. personal computer industry. *Econometrica* 76:1017–1074.

[41] Kenneth E. Train (2009). *Discrete Choice Methods with Simulation*. Cambridge University Press.

[42] Christopher J. Tyson (2008). Cognitive constraints, contraction consistency, and the satisficing criterion. *Journal of Economic Theory* 138:51–70.

[43] Christopher J. Tyson (2013). Behavioral implications of shortlisting procedures. *Social Choice and Welfare* 41:941–963.