CONSUMER SEARCH AND ONLINE DEMAND FOR DURABLE GOODS

Jun B. Kim* Bart J. Bronnenberg† Paulo Albuquerque‡

June 26, 2008

(Preliminary and incomplete, please do not cite or post)

Abstract

Using aggregate, product search data from Amazon.com, we jointly estimate consumer information search and online demand for durable goods. To estimate the demand and search primitives, we introduce an optimal sequential search process into a model of choice and treat the observed market-level product search data as aggregations of individual-level optimal search sequences. The model builds on the dynamic programming framework by Weitzman (1979) and combines it with a choice model. The model can accommodate highly complex demand patterns at the market level, and at the individual level the model has a number of attractive properties in estimation, including closed-form expressions for the probability distribution of alternative sets of searched goods and breaking the curse of dimensionality. Using numerical experiments, we verify the model's ability to identify the heterogeneous consumer tastes and the distribution of search cost from product search data. Empirically, the model is applied to the online market for camcorders and is used to answer manufacturer questions about market structure and competition and to address policy maker issues about the effect of recommendation tools on consumer surplus outcomes. We find that consumer online search for camcorders at Amazon.com is typically limited to less than 10 choice options, and that this affects the estimates of own and cross elasticities. In a policy simulation, we also find that the majority of the households benefit from the Amazon.com's online recommendations via lower search costs. However, lowering search cost through product recommendations to popular product pages may cause worse choice outcomes or higher total search cost for households with atypical preferences.

Keywords: cost-benefit analysis, optimal sequential search, demand for durable goods, information economics, consideration sets

* Jun B. Kim is a Ph.D. student at the UCLA Anderson Graduate School of Management.
† Bart J. Bronnenberg is Professor of Marketing, and CentER Research Fellow, Tilburg University.
‡ Paulo Albuquerque is Assistant Professor of Marketing at the Simon School of Business, University of Rochester

1 Introduction

Online demand for consumer durables and search goods is large and rapidly growing. Comscore (2007) estimates that non-travel U.S. online consumer spending in 2006 reached $102.1 billion. Jupiter Media Metrix (2004) estimated U.S. online consumer spending of $65 billion in 2004, with $20.2 billion on durable consumer search goods, and another $8.3 billion on information goods. The Comscore report shows that the fastest growing e-commerce categories include durable search goods such as video consoles, consumer electronics, furniture, appliances, and equipment, as well as information goods such as books and magazines, music, and software. These categories saw annual growth rates for 2007 of 25% to 50% range. PC World reported in 2007 that the "appeal of online shopping is growing. Between August 2006 and the same month a year later, 14 percent of the $159 billion that U.S. shoppers spent on consumer electronics was spent online, up from 5 percent a year earlier, according to the Consumer Electronics Association."

In this paper, we seek to understand online demand and product information acquisition for durable search goods and/or information goods at Amazon.com using aggregated histories of search behavior. Our approach is to treat browsing behavior as the outcome of an optimal sequential search process across choice options for which the consumer has different expectations and uncertainties. In addition, these choice options need not all be equally accessible and may be offered to consumers at different search costs (for instance through the use of seller sponsored recommendation engines). Recognizing that the three demand primitives – expectations, uncertainties, and search cost – can be changed by interested parties, e.g., manufacturers or policy makers, the substantive goal of this paper is to analyze the impact of differential search cost, e.g., product recommendations, on the search and choice decisions of the consumers, and on the competitive structure of the online market for consumer durable goods. Methodologically, we introduce an optimal sequential search process into a model of choice and identify the demand parameters of interest from the search data.

Table 1 presents an example of viewing data for camcorders obtained from Amazon.com. The table lists products that were viewed by consumers conditional on viewing a particular or focal product, the Sony DCR-DVD108. In addition, the order in which these products are listed is determined by an Amazon.com algorithm that uses the frequency of same-session viewing of the focal product and the other products listed[1]. The table does not list all existing 300+ camcorder

[1] This data generating mechanism is explained separately in the data section.

options and reflects the fact that some options are never viewed together with the focal product by any consumer in the same online session. The data in Table 1 exist for each of the camcorder options as the focal product. Because the products in the view list are rank ordered, we refer to these data as the view-rank data. The paper shows that across viewed items, the view-rank data are informative about substitution, and that from viewed to non-viewed items, the data imply either low or lack of substitution. For durable goods, without meaningful observations of consumer switching, the premise of this paper is that the view-rank data are in the spirit of revealed measures of substitution.

A related premise is that the view-rank data can be used to estimate the demand systems. This would be of interest to practitioners and policy makers because the Amazon.com view-rank data are accessible to anyone, and contain cross-product information that is not present in reports of sales volume or market share.

The general approach in the paper is to model the view-rank data as the aggregation across consumers of individual-level optimal search sequences, in which each consumer tries to maximize her expected utility minus total search cost for the camcorder options that she inspects. At the individual level, our approach yields a probabilistic model of the optimal sequence of search, is not subject to the curse of dimensionality, and is purposely suited to be estimable using view-rank data.

Using data experiments, we find that the model is successful at identifying the parameters of a choice-based demand system with random effects. In addition, the model correctly identifies search cost, and search set size. Finally, we show that the model produces the correct predictions of market shares, despite the fact that market shares are not part of our data.

From an application of our model to the Amazon.com camcorder category, we find the following results. The median (average) search set contains 6 (7.2) products and the estimated distribution of search set size has a long tail to the right. We estimate that the cost of search is significant and is subject to consumer heterogeneity. We find that the search cost is lowered for products on the Amazon.com web site that have many incoming links, measured as the number of times a particular product is recommended on the pages of other products. We also find that online competition between many products is effectively 0, because many pairs of products are not searched jointly by consumers. We estimate that more than 50% of product pairs are viewed by less than 0.1% of the population. This implies severe limits on substitution, which

View-rank	Brand	Media format	Optical Zoom	···	Price
1	SONY	DVD	25	···	$443.32
2	PANASONIC	MINIDV	32	···	$248.11
3	SONY	DVD	20	···	$539.00
4	SONY	HD	10	···	$665.20
5	SONY	HD	40	···	$509.84
6	SONY	MINIDV	40	···	$299.99
7	SONY	MINIDV	25	···	$363.88
8	PANASONIC	DVD	30	···	$347.55
9	SONY	MINIDV	20	···	$257.43
10	CANON	DVD	25	···	$345.99
11	SONY	MINIDV	10	···	$552.42
12	HITACHI	DVD	10	···	$378.45
13	SONY	DVD	10	···	$790.22
⋮	⋮	⋮	⋮		⋮
37	SONY	DVD	10	···	$752.75
38	CANON	DVD	35	···	$354.78
39	CANON	DVD	35	···	$376.57
40	PANASONIC	MINIDV	32	···	$289.39
41	SONY	HD	25	···	$554.14
42	JVC	MINIDV	32	···	$488.88
43	PANASONIC	DVD	32	···	$361.81

Table 1: Product options searched at Amazon.com, in May 2007, given search of a Sony Camcorder with DVD media format, 40 × optical zoom, 2.5-Inch swivel screen, etc., selling at $328

in turn cause for many cross-elasticities to be numerically zero. Finally, our results show that not everyone benefits from selectively lowering search costs on products that are popular (e.g., through recommendations). Indeed, those consumers who have atypical preferences and who are susceptible to product recommendations, will often be worse off.

This paper is organized as follows. The next section reviews the literature. Section 3 outlines the model. Section 4 presents the data and discusses the Amazon.com US Patent on which the data generation is based. Section 5 explains model operationalization and estimation. Section 6 presents evidence from numerical experiments to show that the model is identified. Section 7 presents empirical results. Section 8 contains two policy experiments. Section 9 concludes.

2 Background

Because of non-zero search cost, product proliferation, and preference dispersion in most industries, marketing scholars and economists have long recognized that consumers do not in general search or consider the universal choice set in an industry (e.g., Hauser and Wernerfelt 1989; Howard and Sheth 1969; Nelson 1970; Stigler 1961). The recent popularity of the choice based demand system has brought renewed attention to the issue of modeling choice sets and the concern exists that not taking into account the limited nature of choice sets leads to biased estimates of demand (Bruno and Vilcassim 2008; Chiang, Chib, and Narasimhan 1999; Goeree 2008). Papers in this tradition specify a probability of a product being known (Goeree 2008) or accessible (Bruno and Vilcassim 2008) that is not the outcome of an optimal search process but simply constitutes a consumer's response to firms' actions. In this paper, we advocate that such responses can only be measured in the context of how they affect the consumer's search strategies. In addition, if one has access to outcomes of search behavior, as we do here, those become informative of important demand primitives when viewed through the lens of optimal information search.

Understanding consumer information search has been an important topic both in marketing and economics and hence research on consumer information acquisition abounds. Starting with Stigler (1961), early research on consumer information acquisition focused on consumers searching for price-quotes in homogeneous goods markets at some effort. Extending the scope of consumer search to issues of market outcomes, several authors theorized that limited consumer information search may have a significant impact on market structure (Diamond 1971; Nelson 1974; Anderson and Renault 1999). In this paper, we model consumer search behavior not only to evaluate market

structure issues, but also to evaluate the impact of changing search costs by firms or by online sellers on consumer utility.

We model the consumer's willingness to search for choice options by assuming that the consumer is motivated to search only if she benefits from doing so. Hence, we are interested in modeling search behavior as the outcome of a reasoned process. There is already a tradition in the consideration set literature to represent consideration sets as the outcome of non-sequential search (Lattin and Roberts 1991; Mehta, Rajiv and Srinivasan 2003). This tradition rests on the fixed sample strategy proposed in Stigler (1961) as an optimal search policy for a consumer in a commodities goods market under price uncertainty.

In contrast, McCall (1965) and Nelson (1970) argue that a sequential search strategy is optimal in terms of total cost[2] and since we additionally believe that online search is more correctly captured as a sequential process, we will model online search for information in this study as a sequential process and use theory of optimal sequential search. Seminal contributions to sequential search theory have been made by Weitzman (1979), in the case of single agent problems and by Reinganum (1982, 1983) in the case of multiple agent problems. We seek to implement the optimal search strategies of these papers into a single-agent random utility choice model.

In contrast to a large volume of theoretical work, there has been relatively limited empirical research on consumer information search using secondary data. Two recent exceptions are papers on empirical search for commodities (Hong and Shum 2006) and for differentiated products (Hortaçsu and Syverson 2004). In the former, the authors devise a model that translates the price dispersion into heterogeneous search cost across population. In the latter, the authors develop a model to translate the utility distribution into heterogenous search cost. Moraga-Gonzáles (2006) contains a comprehensive review of several empirical applications. In our case, like Hortaçsu and Syverson (2004), we model search for differentiated products, but unlike them, we have collected direct measures of search outcomes, allowing us to estimate a more general demand model. For instance, in contrast to the homogeneous demand model in Hortaçsu and Syverson (2004), we estimate both heterogeneous consumer preferences and search costs in a differentiated product category.

With our choice model that includes optimal sequential search, we seek to explore the influence of online retailers' product recommendations on consumer search behavior and choice

[2] Actually, block-sampled search strategies have been argued to be even better (see e.g., Morgan and Manning 1985). However, in online search such strategies can not be executed and therefore they are not considered here.

outcomes. Given the popularity and ubiquity of recommendations at many online stores, it is of practical and academic interest to investigate how recommendations affect the consumer information and product search decisions. In behavioral work, Huang and Chen (2006) report that the recommendations of other consumers influence the choices of subjects more effectively than recommendations from an expert. Senecal and Nantel (2004) also show that a retailer's recommendations will ultimately affect demand.

3 A demand model with costly sequential product search

3.1 Utility

Our modeling assumptions at the individual level are as follows. Consumer i has a utility for product $j = 1, ..., J$ that is equal to

$$u_{ij} = V_{ij} + e_{ij} \tag{1}$$

with

$$\begin{aligned} V_{ij} &= X_j b_i \\ b_i &\sim N(b, B) \\ e_{ij} &\sim N(0, \sigma_{ij}^2). \end{aligned}$$

We assume the matrix B is diagonal. The outside good is the $(J+1)^{st}$ alternative, and the consumer is aware of the option not to buy. This option does not require a search and is available at no cost.

The utility function thus contains an expectation V_{ij} and an unknown component of utility, e_{ij}. Our interpretation is that this decomposition partitions what the consumer knows and does not know into V_{ij} and e_{ij}, and the consumer's goal of search is to resolve e_{ij}. This is not a limiting assumption.[3] Knowledgeable consumers may have lower variance e_{ij}'s and less knowledgeable consumers may have higher-variance e_{ij}'s. We assume that e_{ij} is zero-mean. Many important attributes for a product are accessible from landing pages or general category information displays without retrieving the product web page,[4] facilitating the existence of an expectation V_{ij}. When

[3] Our interpretation is consistent with Nelson (1970) who defines consumer search as an information problem to fully evaluate the utility of each option.

[4] In the digital camcorder category at Amazon.com, a consumer already has access to important product characteristics in camcorder such as brand, price, media format, zoom, pixel number, and the dimension.

consumers request the product detail web page, they see more details about the product which resolves e_{ij}.

Resolving e_{ij} upon search comes at some cost. We introduce product and individual specific search cost, c_{ij}, which we interpret mainly as time spent on discovering and evaluating the product.[5] We model search cost as a log normally distributed random effect

$$c_{ij} \propto \exp\left(L_j \gamma_i\right), \tag{2}$$

with

$$\gamma_i \sim N(\gamma, \Gamma)$$

where the matrix Γ is diagonal. The lognormal specification ensures that the sign of c_{ij} is consistent with theory, i.e., positive. The cost attributes L_j describe the accessibility of product j. For instance, it may contain the number of links into product j's page or the number of times it is recommended, etc.

The consumer's search and choice process are the outcome of her desire to maximize expected utility minus total search cost. This involves contrasting the marginal benefit and marginal costs of search. The objective of the analyst is to estimate b, B, γ, and Γ from data.[6]

3.2 A model of sequential search

In sequential search, a consumer decides to stop or continue search each time after having searched a product. The theory of optimal sequential search states that consumers only continue search if the marginal benefits of doing so outweigh the marginal costs.

Utility u_{ij} of consumer i for product j is $V_{ij} + e_{ij}$. Define u_i^* at any stage of the search process as the highest utility among the searched product thus far. The consumer's expected marginal benefit from search of product j is

$$\mathcal{B}_{ij}\left(u_i^*\right) = \int_{u_i^*}^{\inf} \left(u_{ij} - u_i^*\right) f\left(u_{ij}\right) du_{ij}, \tag{3}$$

[5] Search cost is different between consumer packaged goods and consumer durables. For packaged goods in which experience is more easily obtained, mental maintenance and processing cost constitute the majority of search cost (Lattin and Roberts 1991). For one-time purchases such as consumer durable goods, it is more likely that search costs are determined by the time spent on searching for more information and the need for evaluation. Therefore in the context of digital camcorders, we interpret marginal search cost as the opportunity cost of time invested in identifying and evaluating another candidate product.

[6] In the empirical analysis, we will assume that $\sigma_{ij}^2 = 1$, but in the modeling section we wish to keep the level of product uncertainty general.

where $f(\cdot)$ is the probability density distribution of u_{ij}. The marginal benefit is the expectation of the utility for j given that it is higher than u_i^*, multiplied by the probability that u_{ij} exceeds u_i^*.[7] Note that the benefit of search only depends on the arrangement of utility above u_i^*. The left tail of the utility distribution below u_i^* can be arbitrarily rearranged without affecting search or choice.

The goal of the consumer is, given the current best option**,** to maximize expected utility minus incurred search cost over a set of options that, at the individual level, are characterized by product specific mean utilities, V_{ij}, product specific search costs, c_{ij}, and possibly product specific uncertainties, captured by σ_{ij}^2. This implies that the consumer continues search if there exists at least one j such that

$$c_{ij} < \mathcal{B}_{ij}\left(u_i^*\right), \tag{4}$$

i.e., if the expected marginal benefit of searching is larger than the marginal cost, c_{ij}.

The optimal sequential search strategy can be formalized as follows. First, partition the set of options into $S_i \cup \bar{S}_i$, with S_i containing all searched options and $\bar{S}_i$ containing all non-searched options. All decision relevant information about S_i is contained in $u_i^* = \max_{j \in S_i}\left\{u_{ij}, 0\right\}$, provided we assign 0 to the utility of not buying anything.

At any point in the search process, the state of the system is given by $\left(u_i^*, \bar{S}_i\right)$. Define the value function $W\left(u_i^*, \bar{S}_i\right)$ as the expected (discounted) value of following an optimal search policy, from the current state going forward. This value function must satisfy the following Bellman equation (Weitzman 1979)

$$W(u_i^*, \bar{S}_i) = \max(u_i^*, \max_{j \in \bar{S}_i}(-c_{ij} + \beta_i \cdot [\underbrace{F(u_i^*) \cdot W(u_i^*, \bar{S}_i - \{j\})}_{u_{ij} \leq u_i^*} + \underbrace{\int_{u_i^*}^{\infty} W(u_{ij}, \bar{S}_i - \{j\}) f(u_{ij}) du_{ij}]}_{u_{ij} > u_i^*})) \tag{5}$$

This equation says that from state $\left(u_i^*, \bar{S}_i\right)$, the consumer can terminate search and collect u_i^*, or the consumer can search any $j \in \bar{S}_i$. In the latter case, the consumer gets in expectation

[7] This can be seen by writing (3) alternatively as

$$\mathcal{B}_{ij}\left(u_i^*\right) = \left(1 - F_j\left(u_i^*\right)\right) \times \int_{u_i^*}^{\inf} \left(u_{ij} - u_i^*\right) \frac{f\left(u_{ij}\right)}{\left(1 - F_j\left(u_i^*\right)\right)} du_{ij},$$

which is the multiplication of the chance that the utility draw is larger than u_i^* and the expected value of a truncated draw from the distribution of u_{ij} above u_i^*

$F(u_i^*) \cdot W(u_i^*, \bar{S}_i - \{j\}) + \int_{u_i^*}^{\infty} W\left(u_{ij}, \bar{S}_i - \{j\}\right) f\left(u_{ij}\right) du_{ij}$, which she seeks to maximize across j. Because all online search in a single session is conducted in a short time span, we set the discount rate β_i to 1.

3.3 The optimal strategy

The solution to the above dynamic program is to continue searching until a utility u_i^* is discovered that is larger than some limit, which in turn depends on how much option value is still left in the unsearched set. This limit depends on a quantity that is called a "reservation utility". To define this concept, each consumer i has a reservation utility z_{ij} for each product j that – if she had already found a product with that utility – leaves her indifferent between searching and not searching j. In other words, the reservation utility z_{ij} obeys the following equation (see also equation 4, above):

$$c_{ij} = \mathcal{B}_{ij}\left(z_{ij}\right) = \int_{z_{ij}}^{\inf} \left(u_{ij} - z_{ij}\right) f\left(u_{ij}\right) du_{ij}. \tag{6}$$

Thus, the reservation utilities solve $z_{ij} = \mathcal{B}_{ij}^{-1}\left(c_{ij}\right)$. For the moment, assume that the reservation utilities can be computed and are unique. The estimation section establishes that $\mathcal{B}_{ij}$ is monotonic and a separate appendix provides the details of computation.

The optimal search strategy (see, e.g., Weitzman 1979) that solves the consumer's maximization problem of equation (5) has three components, a selection rule, which determines the ordering of the search sequence, a stopping rule, which determines the length of the search sequence, and a choice rule.

1. Selection rule: Compute all reservation utilities z_{ij}, and sort them in descending order. If a product is to be searched, it should be the product with the highest reservation utility z_{ij} among the products not yet searched.

2. Stopping rule: Stop searching when the highest utility obtained so far, u_i^*, is larger than $\max_{j \in \bar{S}_i}\left(z_{ij}\right)$ among the unsearched items.

3. Choice rule: Once search stops, collect u_i^* by choosing the maximum utility alternative in S_i.

We note that this search and choice process can accommodate that some consumers do not search at all. Indeed, consumers for whom $\max_j\left(z_{ij}\right) < 0$ for all j will not find it worth their time

to search brands. They will choose the outside good.[8] The same process can also accommodate that some consumers just browse but do not buy. For such consumers, $\max_{j \in S_i}(z_{ij}) > 0$, but $\max_{j \in S_i}(u_{ij}) < 0$. These two statements are not in conflict, as will be seen below. These consumers will also choose the outside good.

Two important points need to be made. First, given a choice set, the choice model above is not a probit model. For instance, given the stopping rule above, search beyond item k is continued only if the utility draw for e_{ik} is low enough. This implies that conditional on observing a specific choice set, the e_{ij} are not distributed normal with mean 0 and variance σ_{ij}^2. Therefore, given search, choice probabilities do not follow a standard probit.

Second, Chiang, Chib and Narasimhan (1999) mention that identification of choice sets (or in this case: search sets) is subject to the curse of dimensionality. Indeed, in a non-sequential search process, with J possible alternatives, there exist 2^J possible search sets. This large number of permutations would render the computation of the search frequency of any given product impossible with universal choice set sizes of $J = 300+$ at Amazon.com. However, an important computational windfall of the sequential search process is that it is not subject to the curse of dimensionality. Given the selection rule above, there are only J possible choice sets at the individual level. Given a set of individual level parameters, there will be an ordering of the choice alternatives along their reservation utilities z_{ij}, and the consumer optimally samples these choice alternatives in descending order. Thus, if the z_{ij} can be computed, the contents of a search set of size m is known. In sum, whereas, across consumers, the model allows for the existence of any of the 2^J possible search sets as a consequence of consumer heterogeneity, at the individual level only J of these sets can be an outcome of the optimal sequential search process that belongs to a particular vector of individual parameters.

Before completing the model, we investigate some properties of the search sequence by means of an example.

3.4 Some characteristics of the search sequence

In Figure 1, we plot the relation between c, σ^2, and z, under the assumption of normality of e_{ij}. The left hand panel varies search cost from 0 to 2. For reference, in this example we choose $V_{ij} = 1$, and $\sigma_{ij}^2 = 1$. The z_{ij}, the reservation utility, or more intuitively the relative attractiveness

[8] Note that because we estimate our model on search data, we assume that all consumers search. However, the model can actually accommodate non-search behavior.

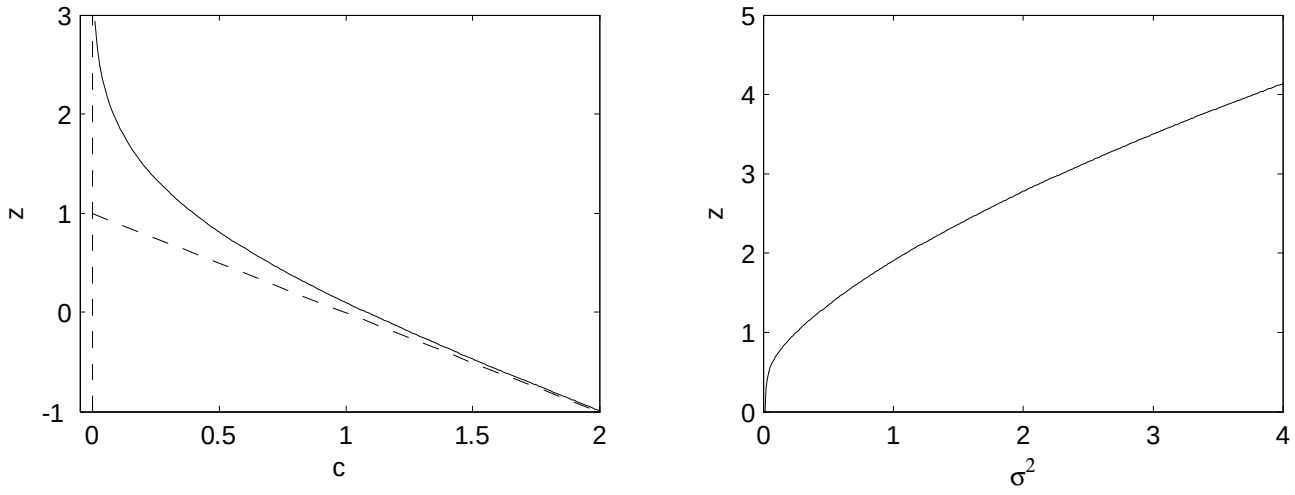

Figure 1: The relation between search cost, c, product uncertainty, σ^2, and search attractiveness, z.

of searching j, is decreasing in its search cost. As search cost increases, z_{ij} goes to $V_{ij} - c_{ij}$. This implies that as search costs increase relative to product uncertainty, the attractiveness of search tends to go to the expected utility net of search cost. On the other hand, if search costs are low relative to product uncertainty, or product uncertainty is high relative to search cost, z_{ij} goes to infinity. Indeed, if it is free to search, the option value (upside) of searching any product that has utility support on R^+ is infinite.

The relation between σ^2 and z in the right hand side panel shows the option value of noisy prospects. For reference, in this graph $V_{ij} = 1$ and $c_{ij} = 0.1$. As outlined above, in sequential search, the search value of a product is determined by its upside. That is, anything lower than the current maximum u_i^* is irrelevant. Per consequence, the reservation utility z_{ij} is increasing in product variance. As a natural consequence, if novice consumers are characterized by having high σ_{ij}^2 relative to V_{ij}, they will tend to have higher z_{ij} and thus search more than consumers who have more experience. For completeness, we note that z_{ij} increases linearly in V_{ij}.

3.5 Inclusion probabilities and set occurrence

Our data (see the next section) are a function of the frequency with which products are being viewed or searched, and therefore we seek to derive the probability π_{ij} that a given product j is included in the optimal search set of consumer i.

Consider that we know z_{ij} and V_{ij} for each individual and product. With some abuse of

notation, denote the rank of z_{ij} by r, with $r(1)$ returning the index j of the highest ranked z_{ij} and $r(J)$ returning the index j with the lowest ranked z_{ij} for individual i. From these definitions, $\pi_{i,r(1)}$ is the inclusion probability of the product with the highest ranked z_{ij} for consumer i, and $\pi_{i,r(j)}$ is the inclusion probability of the product with the j^{th} highest ranked z_{ij}.

The contents of set S_{ik} is fully determined by the selection rule (ranking on z) and the stopping rule (the size k). The probability $\pi_{i,r(j)}$ that product $r(j)$ is in the set, is equal to the probability that the first $j-1$ draws of utilities all fell short of $z_{i,r(j)}$ (which is less than $z_{i,r(j-1)}$ by the selection rule above). Thus, the inclusion probability of product $r(j)$ is

$$\begin{aligned} \pi_{i,r(j)} &= \Pr\left(\max_{k=1}^{j-1}\left(V_{i,r(k)}+e_{i,r(k)}\right) < z_{i,r(j)}\right) \\ &= \prod_{k=1}^{j-1} F\left(z_{i,r(j)}-V_{i,r(k)}\right) \ , \quad j>1 \end{aligned} \tag{7}$$

with $\pi_{i,r(1)}=1$[9] and $F(\cdot)$ is the cumulative probability distribution of e_{ij}, which in our case is the normal distribution with mean 0 and variance σ_{ij}^2.

There are three useful properties of these probabilities of inclusion.

1. First, it is trivial to show that $\pi_{i,r(j)} > \pi_{i,r(j+1)}$, or the inclusion probability of the $(j+1)^{th}$ product is always less than the inclusion probability of the j^{th} product.

2. Second, given the sequential nature of search and the selection rule of the optimal strategy, the probability that $r(j)$ and $r(j+k)$ occur together in a set is equal to the probability that $r(j+k)$ is in the set.

$$\pi_{i,\{r(j) \text{ and } r(j+k)\}} = \pi_{i,r(j+k)} = \min\left(\pi_{i,r(j)}, \pi_{i,r(j+k)}\right), \tag{8}$$

   where the last step is from the first property. In the estimation section, we will use the last formulation of this property, when we need to determine the probability that two product j and k are jointly in the set.

3. Third, given the sequential nature of choice and the independence of the $e_{ij,}$ the probability that the set S_{ik} occurs can be computed as follows. First, recall that S_{ik} is the optimal set of size k for individual i. The probability that S_{ik} occurs is equal to the probability that search continues beyond $r(k-1)$ minus the probability of continuing search beyond $r(k)$.

[9] Because our data are predicated on the occurrence of search, consumers search at least one product.

This is equal to the chance that $r(k)$ is in the choice set minus the chance that $r(k+1)$ is in the choice set of consumer i. Thus

$$\Pr\left(S_{i,r(k)}\right) = \pi_{i,r(k)} - \pi_{i,r(k+1)}, \tag{9}$$

This concludes the statement of the individual-level model. The aggregation to the level at which Amazon.com reports its data is explained in the estimation section. For completeness, it is also explained that an alternative to the approach in this subsection is to use draws of the e_{ij} and compute realizations of the process. This would lead to less computation at the individual level, but at the aggregate level we would have to use a frequency estimator for market level behavior, whereas using the model above, we can integrate over a probability model with far greater precision.

4 Data

4.1 The view-rank data

We have collected, on a daily basis, the view-rank data for all camcorder products from May 2006 until October 2007. The data are also updated on a daily basis by Amazon.com. To ensure that the analysis is based on a sufficiently large sample of viewing behaviors, and because we do not have information about the temporal window used by Amazon.com in computing the view-rank data, we aggregated the view-rank data to the monthly level.[10] We use data from the month of May 2007.

In total, we extracted the top 200 camcorders from the Amazon.com website, based on sales-rank. We removed the niche players Samsonic and DXG who cater to the lowest-price tier only with different types of camcorders and which have very low sales-rank. We also removed from the analysis those camcorders on which we had no observations of media format, and all camcorders of professional grade. After applying these data filters, we are left with 113 choice options from 10 manufacturers. The summary statistics of the products are shown in Table 2.

[10] We use average sales price in our analysis. In the data, we observe that the positions of products in the view-rank lists fluctuate over time. This calls for an averaging mechanism for the different positions of a product in the lists over time. For this averaging procedure, we use the percentile ranking similar to Bajari, Fox and Ryan (2007). In the percentile ranking, the product with the highest rank among J products is coded as J, not 1. Then we normalize the rank of product j at time t as

$$\hat{r}_{jt} = \frac{r_{jt}}{\max_k\{r_{kt}\}} \tag{10}$$

Attributes	Ranges
Brand	Sony (33), Panasonic (21), Canon (16), JVC (15), other (28)
Media Formats	MiniDV (36), DVD (34), FM (24), HD (19)
Price	$ 478 (mean), $ 273 (std. dev.)
Form	Compact (21), Conventional (92)
High-Definition	Yes (14), No (99)
Pixel	1.77M (mean), 1.51M (std. dev)
Zoom	17.6 (mean), 12.0 (std. dev.)

Table 2: Description of the choice options in the empirical data (with frequency of occurrence in parenthesis)

All 113 products have their own view-rank lists, i.e., all of the products have a list from which we observe which other products are closely related, in the order of decreasing relationship. On average, a given product appears 39 out of 113 times on other products' view list with a standard deviation of 26. The minimum number of appearances is 0 while the maximum is 109.

Table 3 gives the results of a descriptive regression of the number of appearances on the viewlists.[11] Note that Sony, Panasonic, and Canon appear most frequently of the view-rank lists. Further, hard drive storage, high definition, and pixel size improve the number of appearances, while higher price reduces it. We conclude that the number of appearances on the view-ranks depend on demand drivers such as product attributes and prices.

We point out the rich information embedded in the Amazon view-rank data. For every focal product k, Amazon.com provides a list of top N most related products among the remaining $J-1$ products.[12] Also, product k may appear on the view-rank lists of other $J-1$ products. Therefore, the data reveal a complex pattern of relations between a given product k and the other $J-1$ products.

Lastly, we discuss the type of consumers who we believe are represented in the product search data. Moe (2003) classifies online store browsing behavior of consumers into four different categories - directed buying, search and deliberation, hedonic browsing, and knowledge building. She also classifies the contents of e-commerce web pages into three different categories: product, category, and information pages. She reports that the consumers in directed buying mode will

Once we compute $\hat{r}_{jt}$, the percentile ranking of the product j at t, we compute the average ranking of product j as the mean of the daily percentile ranking as $\hat{r}_j = \frac{1}{T}\sum_t \hat{r}_{jt}$.

[11] In this regression, all categorical variables are "effects coded," i.e., rather than setting the response to one level of these variables to 0, we let the response to that level be the negative sum of the effects of all remaining category levels. For instance, the effect for the brand Mustek is the negative sum of the effects of all remaining brands.

[12] During the data collection period, Amazon.com listed up to 45 products that are related to the focal product.

Variable	β	std.err.
Intercept	53.32	20.65
Sony	34.40	6.62
Panasonic	22.73	5.77
Canon	24.06	6.12
JVC	-0.47	6.82
Samsung	8.01	6.48
Sanyo	-37.51	13.22
Aiptek	-28.34	9.48
Pure Tech.	-15.70	11.12
Hitachi	15.65	11.80
Mustek[a]	-22.81	–
MiniDV	-0.94	3.50
DVD	-10.31	3.88
FM	-10.20	8.48
HD[a]	21.46	–
Compact	-0.09	5.85
Non-Compact[a]	0.09	–
High Definition	12.80	3.42
Non High Definition[a]	-12.80	–
Zoom	0.29	0.263
Screen Size	-5.60	7.73
Pixel	8.19	2.19
Price	-58.41	12.31
Link	1.87	0.28
R^2	0.673	

[a]categorical variable are effects-coded

Table 3: Descriptive regression of the frequency of product appearance against product characteristics

frequently visit the product page while the consumers in the mode of search and deliberation focus on both product and category pages. Hedonic browsers focus on category pages while consumers in knowledge building will focus on information pages. Montgomery et. al (2004) also identify that the focus of the consumers in the buying mode is product detail pages. Amazon.com's product search data are based on the number of consumers who requested *product detail pages* from the Amazon.com server. Therefore, consistent with previous research, we conjecture that Amazon.com's product search data predominantly reflect the behaviors of consumers in either a buying or search phase with a vested interest in the product category.

4.2 Other measures of search at Amazon.com

We now discuss other data that are available at Amazon.com. At each product detail page, Amazon.com lists up to four top products purchased by consumers who searched the product in the current detail page. These product links serve as shortcuts to other closely relevant products, thereby reducing consumers' search costs for potentially attractive products. We use this information as an explanatory variable that affects search cost. For instance, we hypothesize that a product with a large number of incoming links to its page will have a smaller search cost compared to products with a low number of incoming links. Hereafter, we use L_j as the total number of incoming links to product j.

4.3 Amazon.com's generation of view-rank data

Amazon lists a set of closely related products for each focal product in the order of decreasing "strength" of relationship. According to the Amazon.com US Patent *6,912,505 B2* (Linden et al. 2005), the strength of the relationship between two products, j (focal product) and k (related product), is measured by a commonality index, (CI_{jk}), defined as

$$CI_{jk} = \frac{n_{jk}}{\sqrt{n_j} \cdot \sqrt{n_k}}, \tag{11}$$

where n_j and n_k are the numbers of consumers who viewed products j and k, respectively, and n_{jk} is the number of consumers who viewed products j and k together in the same session. Note that n_j, $n_k \geq n_{jk}$ and that the commonality index is bounded between 0 and 1. The higher the commonality index is, the stronger the relationship is between two products. Amazon.com orders its view data, exemplified in Table 1, according to the computed CI_{jk} for each product. So, if the commonality index between product j and k is larger than between product j and ℓ, k appears

before ℓ on the view list for j. If we represent the view-rank of k over ℓ on the view list of j by $(j,k) \succ (j,\ell)$ then

$$(j,k) \succ (j,\ell) \iff CI_{jk} > CI_{j\ell} \tag{12}$$

From the view lists in our data, these inequalities are observed directly[13]. We treat these pair-wise inequalities as the dependent variables in our analysis. To this end we use the indicator variables, $I_{j,k\ell}$, defined as

$$I_{j,k\ell} = \begin{cases} 1 & \text{if } (j,k) \succ (j,\ell) \\ 0 & \text{otherwise} \end{cases} \tag{13}$$

For each product j, there are $0.5 \times (J-1) \times (J-2)$ unique inequalities defined by (12). Therefore, across J products, we theoretically have $0.5 \times J \times (J-1) \times (J-2)$ observed inequalities or pairwise conditional view-ranks. In principle, these data therefore contain a lot of information about substitution patterns, because given our sequential search model, the pairwise data are informative of the degree to which two products are related in search, and therefore in expected utility V_{ij} (as well as other factors such as search cost).

As mentioned above, our empirical study involves 113 products. For these products, and taking into account that Amazon.com may truncate view-lists, the total number of observed pairwise ranks is 401,647. The rank information may contain less information compared to continuous data such as the share information. But, there are many observations of view-ranks and given the search model, such data are informative about substitution and consumer heterogeneity.

5 Estimation

5.1 General approach

The general approach to estimating the parameters of our model is as follows. Given a set of parameters and draws from the heterogeneous distributions, we first use the optimal sequential search model to make forecasts of the unobserved commonality indices CI_{jk} in Equation 11. Below we show how to compute these forecasts. Second, we assume that the model's forecast of CI_{jk} differs from its true value by a median-zero error process. Third, we construct a goal function that counts the number of times that the model correctly predicts the cases where $I_{j,k\ell} = 1$ (see Equation 13). Finally, we maximize this goal function using the Differential Evolution algorithm

[13] Not all pairwise combinations are observed. Some products are viewed so infrequently that the view list is truncated by Amazon.com. While, we do not know the view-rank among products that are not on the view list, we do know the view ranks among all listed products, and the view ranks of pairs of listed and non-listed products.

(Price and Storn 1997). Our estimator is a maximum score estimator (Manski 1975). This estimator is semi-parametric and maximizes the matches between actual and predicted relations or inequalities.

We decompose the true commonality index between two products, j and k, as

$$CI_{jk} = \widehat{CI}_{jk}(X, \Theta) - \epsilon_{jk} \tag{14}$$

where CI_{jk} is the unobserved, true commonality index, $\widehat{CI}_{jk}(X, \Theta)$ is the model's prediction of the commonality index given data X and parameters Θ, and ϵ_{jk} is median zero error process, which include measurement or sampling errors that we add directly to the dependent variables (see Bresnahan 1987 and Bajari et al. 2007 for a similar approach).

A major advantage of the maximum score estimator is that it does not impose knowledge about functional form of the error distribution. Specifically, we merely assume that the error term $\epsilon_{jk}|X$ in Equation 14 is independent and identically distributed across k, within a given product j. We also assume that the measurement errors on CI_{jk} are independent (but not necessarily identically distributed) with respect to j. This allows for heteroskedasticity of unknown magnitude among the CI's across different focal products and thus for a more flexible error distribution than a spherical assumption in which the measurement error on CI_{jk} is assumed to be i.i.d. across j and k.

Next, we tie the above assumptions to our observations. Unfolding the relations among the products into a set of pairwise view-ranks, e.g., product j is viewed more often with k than with ℓ, we obtain

$$\begin{aligned} I_{j,k\ell} &= 1(CI_{j\ell} < CI_{jk}) \\ &= 1(\widehat{CI}_{j\ell}(X, \Theta) - \epsilon_{j\ell} < \widehat{CI}_{jk}(X, \Theta) - \epsilon_{jk}) \\ &= 1(\epsilon_{j,k\ell} < g_{j,k\ell}(X, \Theta)) \end{aligned} \tag{15}$$

where $\epsilon_{j,k\ell} = \epsilon_{jk} - \epsilon_{j\ell}$ and $g_{j,k\ell}(X, \Theta) = \widehat{CI}_{jk} - \widehat{CI}_{j\ell}$. Since the error term ϵ_{jk} is assumed to be i.i.d. with a median equal to zero for a given j, the difference between two such error terms given j will also have a zero median, i.e., $\text{med}(\epsilon_{j,k\ell}|j, X)=0$. A condition for consistent identification in the maximum score estimator is that the response probabilities are known almost everywhere (Manski 1988). We further require one continuous variable in X to break ties in the score, and for this purpose we use price. Finally, we note that the maximum score estimator does not depend on

the functional form $g_{j,k\ell}$ for its identification (Fox 2007) and can be nonlinear in terms of model parameters (Matzkin 1993).

Next, we discuss the sources of the measurement errors ϵ_{jk}. First, it is possible that there is sampling or measurement error in Amazon.com's computation of the CI_{jk}. Second, although we mainly model consumers engaged in optimal search behavior, the Amazon.com data may contain traces of consumers in other modes such as hedonic browsing. Third, we aggregate and average a month's data to generate the pairwise ranks among the products in the product search data. Combined, these forces may introduce measurement error in the dependent variable.

The score function in our empirical analysis is defined as

$$s\left(\Theta\right)=\sum_{j}\sum_{k\neq j}\sum_{\ell\neq j,k}1(I_{j,k\ell}=1,\widehat{CI}_{jk}(X,\Theta)>\widehat{CI}_{j\ell}(X,\Theta)) \tag{16}$$

The score function increases by one when a simulated pairwise rank corresponds to the observed pairwise rank on the view-rank lists which we show as an example in Table 1. The maximum score estimator is

$$\Theta^{*}=\arg\max_{\Theta}s\left(\Theta\right) \tag{17}$$

Below, we provide computational details on how to compute the forecasts $\widehat{CI}_{jk}$. We also explain how the reservation utilities z_{ij} can be computed efficiently in estimation.

5.2 Computational details

The commonality index From the definition of the commonality index used by Amazon.com, the estimator for CI_{jk} is

$$\widehat{CI}_{jk}\left(X,\Theta\right)=\frac{\hat{n}_{jk}}{\sqrt{\hat{n}_{j}}\sqrt{\hat{n}_{k}}}. \tag{18}$$

where $\hat{n}_j$ is equal to the forecasted number of simulated individuals that has searched j (given X and Θ) and $\hat{n}_{jk}$ is equal to the forecasted number of individuals that has jointly searched j and k. We can approximate $\widehat{CI}_{jk}$ to an arbitrary degree of precision by computing it on the basis of the simulated search histories of many pseudo households (draws from the heterogeneity distributions). In terms of our model, the prediction $\hat{n}_j$ is equal to the sum across individuals of the probability that product j is included in the search set, i.e., using Equation (7),

$$\hat{n}_{j}=\sum_{i=1,...,I}\pi_{ij},\ j=1,...,J. \tag{19}$$

where I is the total number of simulated individuals. Further, the prediction $\hat{n}_{jk}$ is equal to the sum across individuals of the probability that products j and k are both included in the search set, i.e., using Equation 8,

$$\hat{n}_{jk} = \sum_{i=1,...,I} \min\left(\pi_{ij}, \pi_{ik}\right), \; j,k = 1,...,J. \tag{20}$$

Therefore, the prediction for CI_{jk} is equal to

$$\widehat{CI}_{jk}\left(X,\Theta\right) = \frac{\sum_i \min\left(\pi_{ij}, \pi_{ik}\right)}{\sqrt{\sum_i \pi_{ij}}\sqrt{\sum_i \pi_{ik}}} \tag{21}$$

Note that since $0 < \widehat{CI}_{jk}\left(X,\Theta\right) < 1$ by construction, the predictor is robust.

Computing reservation utilities The right hand side of Equation 21 involves aggregations of probability distributions of optimal search sets. These optimal choice sets involve individual level optimal search sequences over product options sorted in descending order of reservation utilities, z_{ij}. To compute the reservation utilities z_{ij} in estimation, we develop the following results in Appendix A. First, the reservation utilities follow

$$z_{ij} = V_{ij} + \zeta\left(\frac{c_{ij}}{\sigma_{ij}}\right) \times \sigma_{ij}, \tag{22}$$

where $\zeta\left(\frac{c_{ij}}{\sigma_{ij}}\right)$ is a scalar function that translates standardized search cost c_{ij}/σ_{ij} into a multiplier on σ_{ij}. Thus, given the assumptions of the model, the reservations utilities are simply the expected utilities V_{ij} plus a function of search cost c_{ij} times the uncertainty about the product σ_{ij}.

Second, the appendix shows that the function $\zeta\left(x\right)$ solves the following implicit equation

$$x = \left(1 - \Phi\left(\zeta\right)\right)\left(\lambda\left(\zeta\right) - \zeta\right), \tag{23}$$

where Φ is the cumulative standard normal distribution, and λ is the standard normal Hazard rate, $\phi\left(\zeta\right)/\left(1-\Phi\left(\zeta\right)\right)$ in which ϕ is the standard normal probability distribution function. The function $x\left(\zeta\right)$ in Equation 23 is further shown to be continuous and monotonic. Hence, the inversion to the function $\zeta\left(x\right)$ exists. Although the precise solution of Equation 23 is computationally expensive, it can be solved once for a large set of x outside the estimation algorithm. That is, $\zeta\left(x\right)$ does not need to be solved in estimation because it does not directly involve any model parameters. Armed with a table of x and $\zeta\left(x\right)$, we can – during estimation – substitute $x = \frac{c_{ij}}{\sigma_{ij}}$ and look up $\zeta\left(x\right)$ from the table, possibly using an interpolation step if the table of $\zeta\left(\;\right)$ only covers a neighborhood of $x = \frac{c_{ij}}{\sigma_{ij}}$. These computational steps in estimation can be made arbitrarily

precise and are inexpensive. Thus, we do not need to iteratively solve the reservation utilities in estimation.

Third, this decomposition of z_{ij} has intuitive appeal. If search cost c_{ij} is low, relative to product uncertainty σ_{ij}, $\zeta\left(\frac{c_{ij}}{\sigma_{ij}}\right)$ can be shown to be large and positive. Thus, in this case, the reservation utility or attractiveness of search, z_{ij}, is equal to V_{ij} plus multiples of σ_{ij}. With low cost, consumers focus on the upside of the utility distribution. On the other hand, if c_{ij} is large, then $\zeta\left(\frac{c_{ij}}{\sigma_{ij}}\right)$ turns out to become negative and the reservation utility z_{ij} is less than V_{ij}.[14] In this case, the stopping rule of optimal search will be met earlier, because it is likely that the realized utility draws of u_{ij} is greater than the low reservation utilities z_{ik} of products in the set of unsearched products k.

5.3 Discussion of identification

In this section, we discuss the empirical identification in an informal manner. We first discuss the identification of heterogeneity in consumer tastes. The random-effects in choice-based demand model is identified from deviations in substitution patterns between the homogenous and heterogenous model. It is well known that such deviations are difficult to identify from sales data(e.g., Albuquerque and Bronnenberg 2008, Petrin 2003). Thus, from just observing sales, it may be difficult to estimate heterogenous tastes. However, in our case, we have direct information about whether two products are searched together in a single session. As shown in Berry et. al (2004), having information about other alternatives considered is essential in identifying consumer heterogeneity.

In a sense, the critical unobserved quantities in our model are n_j and n_{jk} where $j, k = 1, \ldots, J$. Once we have the estimates for these quantities, given our results from the optimal search sequence, we can infer the relative popularity of product j and its perceived similarity with k. However, directly estimating n_j and n_{jk} is subject to the curse of dimensionality since the number of parameters required to estimate is $\frac{J(J-1)}{2}$. For $J = 113$, which is the number of products in our empirical analysis, this number is 6,328. Instead, we parameterize and estimate n_j and n_{jk} as an

[14] We note that these observations are not unique to the Normal distribution. We have derived z_{ij} for the Uniform distribution also. For this case, if utilities are distributed uniform on $[V_{ij} - \sigma_{ij}, V_{ij} + \sigma_{ij}]$, we also obtain

$$z_{ij} = V_{ij} + \zeta\left(\frac{c_{ij}}{\sigma_{ij}}\right)\sigma_{ij},$$

with $\zeta() = 1 - 2\sqrt{()}$. In other words, the decomposition in Equation (22) is virtually identical for the Uniform distribution.

aggregate outcome of rational consumers making optimal search decisions and hence reduce the number of parameters in a drastic manner.

Given that we use a score estimator, we can only set-identify the model parameters. However, since the number of pairwise view-ranks is very large (i.e, more than 400,000), we are able to almost point-identify the model parameters.

5.4 Inference

Horsky and Nelson (2006) introduce two methods to statistically test the significance of the mathematical programming based estimators in which the dependent variables are pairwise comparison data. The first statistical test compares model fit with and without the parameter of interest to check its significance. The second method computes the standard error by Jack Knife or Bootstrap. Since the second type is much more computationally intensive, we use the first type. The idea behind the first type is to compute and compare the proportion of correctly predicted inequalities between the two models. If $s(\Theta)$ is the score given the parameter vector Θ in Equation (16), this proportion is $p = \frac{s(\Theta)}{s_{\max}}$. We compute this proportion for the full model (p_f) and for a restricted model (p_r) in which we fix a given parameter to 0. We can statistically test the loss of fit using the standard test for the difference between two proportions. Thus, in the full model, we estimate all the model parameters freely while in the restricted model we keep the test parameter to 0. We reject the null hypothesis that j^{th} parameter is 0 if (Horsky and Nelson 2006)

$$\text{PR} = \mid \frac{p_f - p_r}{\sqrt{(\frac{p_f + p_r}{2})(1 - \frac{(p_f + p_r)}{2})\frac{2}{M}}} \mid \geq \Phi(\alpha/2) \tag{24}$$

where we use standard normal distribution (Φ) for testing since we have a large number of observations. With n parameters to be estimated, the above testing requires n re-estimations of the restricted models, one for each parameter.

6 Data experiment

We conducted a numerical experiment to verify that the model can be identified from view-rank data. To this end, we created 32 product options, from 5 attributes that were arbitrarily named "Sony," "Panasonic," "Zoom $> 10\times$," "Media: mini DV," and "Media: Hard Drive." In addition, we added a continuous attribute "price." Finally, we assigned a value for Product Links, L_j to each of the 32 options.

We chose a random effects specification on all product attributes, a fixed effects specification for search cost, $c_{ij} = \exp(\gamma_0 + \gamma_1 L_j)$, and assumed a set of values for the parameters. As is often the case in empirical models of choice, we need to fix the variation of the e, which led our choice of $\sigma_{ij}^2 = 1$ as a normalization.[15] The assumed values for the parameters were chosen with similar magnitudes as those we obtained from preliminary empirical results. Next, we drew 50,000 pseudo households from the distribution of parameters and, for each of these pseudo households, computed the V_{ij}, z_{ij} and other relevant quantities to obtain the optimal search sequence and choice. We then aggregated these sequences according to the recipe used by Amazon.com[16] to obtain the lists of view-ranks similar to those exemplified in Table 1.

We used the generated view-ranks in estimation. To estimate the model, we generated 3,000 pseudo households. At each candidate parameter vector, we compute the draws of the expected utilities V_{ij} and the reservation utilities z_{ij} to compute the inclusion probabilities π_{ij}. Note again that the z_{ij} takes search cost into account, i.e., low search cost leads to high z_{ij}, etc. The 3,000 draws for the π_{ij} are aggregated and computed according to Equation 21, the forecast of the commonality index given a set of parameter values. We use this forecast to predict the view-rank lists which in turn can be matched against the generated view-rank lists.

The goal function in Equation 16 was maximized using the Differential Evolution (DE) algorithm of Storn and Price (1997). This algorithm is a heuristics-based, direct search method for function maximization.

We discuss two runs of the model. The two runs are based on two separately generated data sets and they also differ in how many iterations the DE algorithm was run. In neither case were improvements in the score observed in the last 25 iterations. Table 4 shows the results. We observe that more than 96% of the pairwise view-ranks were correctly fitted by the model. This high value may be due to the controlled conditions of the experiment and the absence of any specification error.

The correlation between the actual and the estimated parameters is very high, around 0.995.

[15] Note that there is potential for estimating some aspects of these variance terms. Indeed, differences in product uncertainty across options would manifest into low V_{ij} that are searched frequently. Therefore, we envision that the model may become even more general than currently represented once we combine the view-rank with sales-rank data.

[16] There are two ways to do this step. We can generate search histories and use a frequency estimator to compute $\hat{n}_j$ and $\hat{n}_{jk}$. We can also compute the π_{ij} and $\pi_{i(j \text{ and } k)}$ directly without drawing search histories. Both methods of data generation were used with similar results albeit that the second method is likely more precise with fewer pseudo households.

parameter		True values	Run 1 (100 iterations)	Run 2 (150 iterations)
mean effects	Sony	1	1.01	0.87
	Panasonic	-1	-0.57	-1.09
	Zoom $> 10\times$	1	0.89	0.90
	Media: mini DV	-1	-1.20	-1.33
	Media: Hard Drive	2	1.86	1.95
	Price	-2	-1.80	-1.69
heterogeneity	Sony	1	1.01	0.99
	Panasonic	2	2.05	1.93
	Zoom $> 10\times$	1	0.93	0.92
	Media HD	2	1.98	1.77
	Media mini DV	1	1.04	1.04
	Price	1	0.99	1.00
cost	base cost	-2	-2.40	-2.26
	effect of links	-3	-3.04	-3.17
fraction of correct view-ranks		1	0.964	0.969
correlation with true parameters		1	0.994	0.996

Table 4: Estimation results from the numerical experiment

This suggests that the model parameters are identifiable from view-rank data. We call attention to the estimated values of heterogeneity. In both runs, the variances of the distribution of the random effects are very well recovered. This means that the taste variation in the model is well identified. We have argued above, and the simulation results seem to agree, that this is because the view-rank data allow us to directly *observe* which pairs of products substitute well at an aggregate level.

Finally, the model also correctly reproduces the value of the cost parameters. In this example, search costs enter the model as fixed effects. In the empirical results, we will also investigate heterogeneous search cost specifications.

We conclude from this numerical experiment that the model parameters are identified from the view-rank data.[17]

[17] It is perhaps noteworthy to mention that the model also very accurately forecasts market shares. That is, whereas we have not used market shares in the analysis (and we do not have market shares empirically), the simulated data environment of this section allows us to generate market shares using a frequency estimator that counts the number of times a searched product has the highest utility draw. We can subsequently check that the model reproduces these market shares. The correlation between generated and forecasted market shares is close to 1.

7 Empirical analysis

7.1 Specification

We use a random coefficients discrete choice model to represent the utility component of a consumer's product search decisions. In the utility specification, we represent a product as a bundle of characteristics. We do not include a product-specific intercept in the utility function. Utility is modeled as

$$u_{ij} = X_j\beta_i - \alpha_i P_j + e_{ij}, \tag{25}$$

where X_j is a row vector of product j's characteristics and P_j is j's price. β_i is a K-dimensional (column) vector that represents the individual-specific sensitivities to product characteristics. e_{ij} is a random error term with $N(0, \sigma^2)$ and is i.i.d. across individuals and products. We set $\sigma^2 = 1$ for all i and j. We include $K = 8$ product characteristics in the utility specification. Jointly, the responses to the various levels of these 8 characteristics are represented by 18 parameters. We additionally specify random coefficients on all these effects. Further, we impose a theory-driven restriction on the price coefficient. Thus,

$$\begin{aligned} \log(\alpha_i) &\sim N(\beta_p, \sigma_p^2) \\ \beta_i &\sim N(\beta_0, \Sigma), \end{aligned}$$

where Σ is a diagonal matrix containing the variances of the random effects, σ_k^2. Search cost is specified as

$$c_{ij} = \exp(\gamma_{0i} + \gamma_{1i} L_j),\ \gamma_{0i} \sim N(\gamma_0, \sigma_{\gamma 0}^2),\ \gamma_{1i} \sim N(\gamma_1, \sigma_{\gamma 1}^2).$$

The random effects on search cost reflect the different search behaviors or strategies across consumers. For instance, consumers who prefer navigation tools such as sales-ranking or filters will be less responsive to the product links, hence will have low γ_{1i}, while consumers who heavily rely on the products links will have large negative γ_{1i}.

7.2 Parameter estimates

In the estimation, we again run the DE algorithm until we do not observe the score improvement for more than 25 iterations (which involve 640 function evaluations each). It typically takes $100 - 150$ iterations for this criterion to be met.

The estimated parameters are shown in Table 5. The estimated parameters satisfy about 86% of the view-rank inequalities. For reference (not shown), we also estimated the model with

only one random coefficient (price). With this model, the estimates satisfy about 46% of all the inequalities in the pairwise view-rank data. Compared to this benchmark, the full random effects model fits the data much better. Note again that heterogeneity has a large impact on the fit of the model.

The estimated parameters have face validity. For instance, Sony, the most popular brand, has the largest mean brand effect. In general, well known brands have higher mean brand effects than other smaller or unknown brands. Second, the number of incoming product links decreases the search cost, as can be inferred from the sign of the coefficient γ_1, which is negative. The number of pixels determines the picture quality and is known as an important attribute. Also, among the four media formats, we find that DVD and miniDV are more popular than the other two formats. We compare our findings with Gowrisankaran and Rysman (2007) who also estimated demand parameters in the camcorder category. Our finding on consumer sensitivity to pixel[18] is consistent with their dynamic demand analysis. We find a large degree of consumer heterogeneity present in the preference for brands, media formats, form (compact), high definition, and price.

Finally, applying the inference procedure of Horsky and Nelson (2006), we find that all of the estimated parameters are significant.[19]

7.3 Search set analysis

We next interpret the nature of search from the estimation results by analyzing the implied size and composition of the individual optimal search sets. To this end, we compute the size and composition of the choice sets drawing 30,000 pseudo-households from the set of estimated population density of parameters. For each of these pseudo-households, we compute expected utilities V_{ij} and reservation utilities z_{ij}. We next draw their optimal search sets and use various sample statistics to report on the size and contents of the individual level search sets.

Figure 2 shows the distribution of the estimated number of products searched per individual. The mean of the search set size distribution is 7.2 while its median value is 6. We note that about 11% of individuals are estimated to just search one product. The population distribution of search set size has a long right tail. Our random effects model is flexible enough to produce the complex patterns of Figure 2. We conclude that the model implies search sets that are of realistic

[18] We have normalized all our covariates to lie between 0 and 1 while the Gowrisankaran and Rysman have taken log on their continuous variables so that they fall in similar ranges.

[19] This is preliminary and awaits further analysis.

Variable	mean effect (β)	heterogeneity (σ)
Sony	6.410	1.812
Panasonic	4.141	2.668
Canon	4.171	1.621
JVC	3.376	1.200
Samsung	2.066	2.583
Sanyo	1.716	6.302
Aiptek	−0.483	0.723
Pure Tech.	1.057	1.506
Hitachi	2.365	1.790
Mustek	-24.820^a	–
MiniDV	1.287	2.220
DVD	0.879	2.838
FM	−2.264	0.726
HD	0.099^a	–
Compact	−3.061	4.358
Hi-Def	0.392	1.886
Zoom	−0.010	0.086
Screen Size	−0.312	0.015
Pixel	0.489	0.397
log (Price)	1.118	3.182
search base cost (γ_0)	−3.140	0.065
effect of incoming links (γ_1)	−3.038	0.117
score	346,459	
percentage of inequalities satisfied	86.3%	

[a] categorical variables are effects-coded.

Table 5: Estimation results

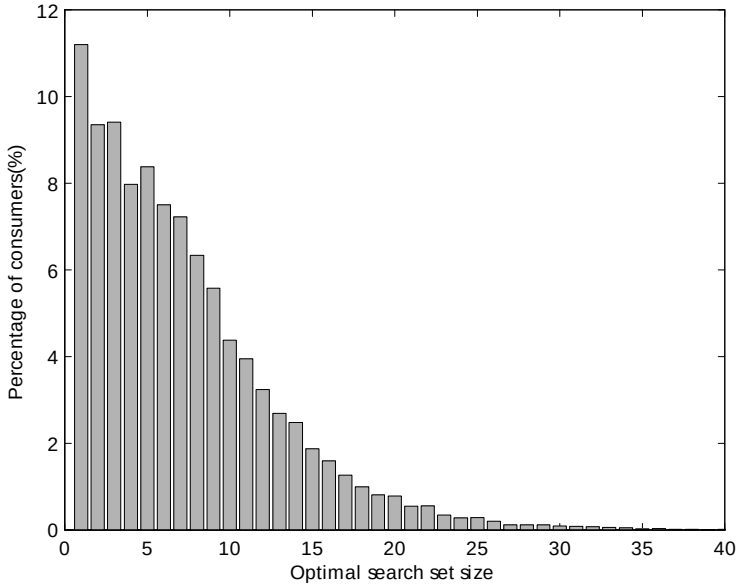

Figure 2: The estimated population distribution of set sizes

size, and that consumers generally do not search too many products.

We now comment on some aspects of the contents of the inferred optimal search sets. First, we present the brand level search set membership. This is important information to manufacturers and allows them to infer a brand's search frequency among consumers. It is also informative about relative "brand strength" or "brand presence" during the pre-purchase phase. The left hand panel of Figure 3 reveals that Sony accounts for about 50% of all products searched at Amazon.com. This bestows on Sony a large "mind share." Panasonic is a distant second followed by Canon. Given that Panasonic offers about a third more products than Canon does, it seems that many Panasonic products are searched at a relatively low frequency. The right hand panel of Figure 3 shows the average share of search volume per product by each brand. The first bar in the graph shows that a typical Sony product has an average share of 1.55% of the total product search volume. Thus, from the two graphs, we conclude that the Sony dominates search process of consumers both at the brand level, as well as on a per-product basis.

Next, we look at which brands are more frequently searched together. The left hand panel in Figure 4 shows the joint search frequency between the Sony and all other brands. The second bar indicates that about 42% of all consumers who search at least one Sony product also search at least one Panasonic product. This is the most frequently searched pair for Sony. In the right

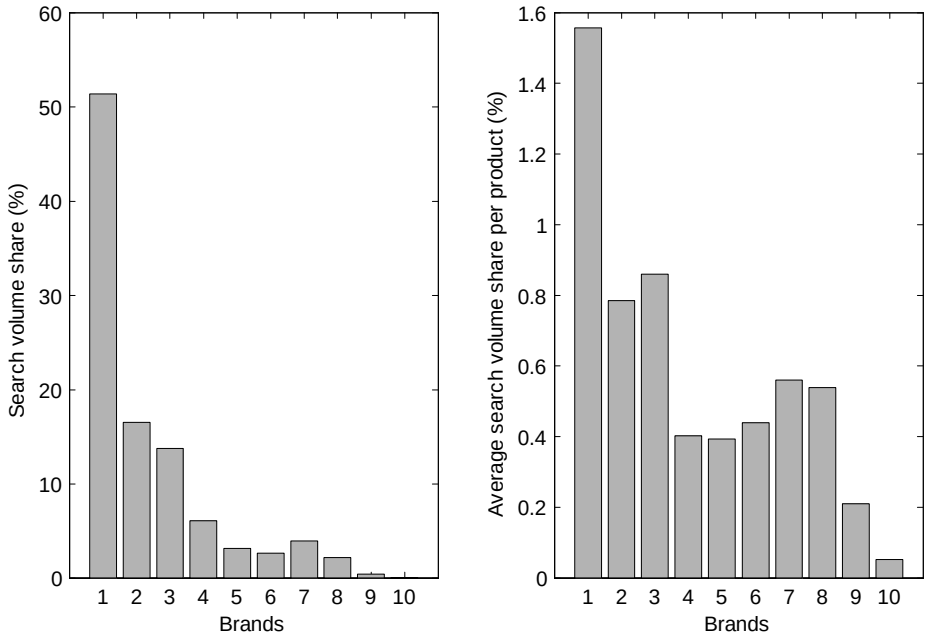

Figure 3: Search volume share by manufacturers and by products. The brands are 1 (Sony), 2 (Panasonic), 3 (Canon), 4 (JVC), 5 (Samsung), 6 (Sanyo), 7 (Aiptek), 8 (Pure Technology), 9 (Hitachi), and 10 (Mustek).

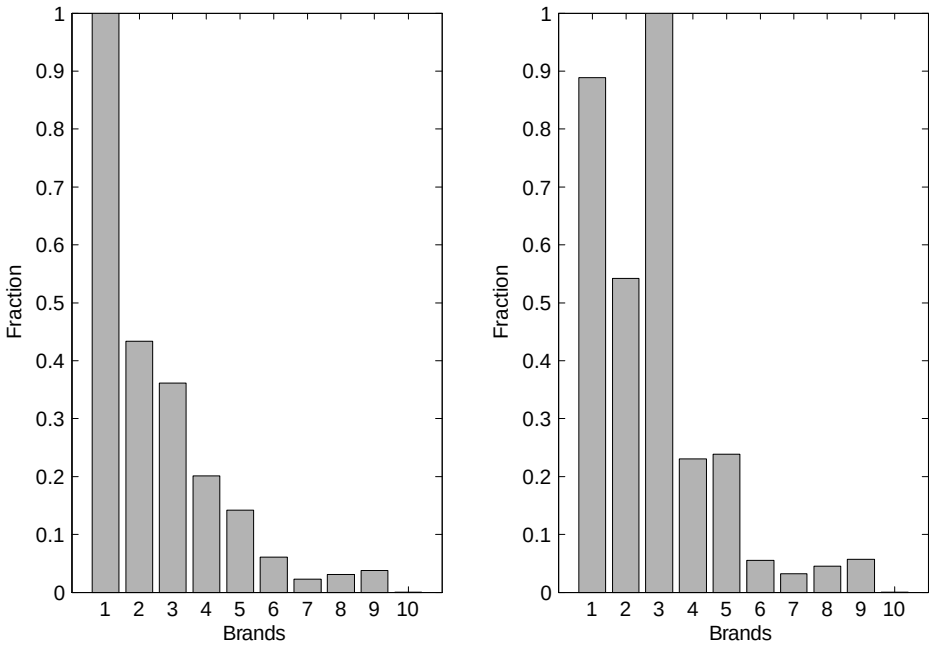

Figure 4: Joint search share conditional on search on Sony (Left) and Canon (Right). The brands are 1 (Sony), 2 (Panasonic), 3 (Canon), 4 (JVC), 5 (Samsung), 6 (Sanyo), 7 (Aiptek), 8 (Pure Technology), 9 (Hitachi), and 10 (Mustek).

hand panel of Figure 4, we show the joint search frequency between Canon and all other products. We see that almost 90% of Canon searchers also search at least one Sony product while slightly more than 50% of them search at least one Panasonic product. Note that the conditional search shares are asymmetric, i.e., 35% of Sony searchers also search for a Canon, but almost 90% of Canon searchers will also search for a Sony. Taken together, our results emphasize that joint search frequencies are generally low. This questions whether assuming full information prior to choice is realistic even at the *brand* level among the top brands.

Since our model is at the individual level, we can infer the joint search frequency among the *products* in a more granular manner. With 113 products, there are 6328 product pairs. Even with heterogeneous tastes, of these pairs, a total of 51% is searched by less than 0.1% of the population. Further, an additional 14% is viewed by more than 0.1% but less than 1%, and another 14% is viewed by more than 1% but less than 10%. No product pairs are predicted to be viewed by more than 20% of the population. We conclude from these numbers that the majority of products is not searched jointly by a meaningful fraction of the population. This limited consumer

search potentially has important implications on demand estimation and on price competition. To investigate this further, we next look at price elasticities.

7.4 Price elasticity

Search We now look at substitution patterns by analyzing price elasticities. We investigate the impact of price on product search volume. To this end, we define the elasticity of search volume with respect to price as,

$$\epsilon_{i,j}^{\text{search}} = \frac{\%\text{ search frequency change in j}}{\%\text{ price change for i}} \tag{26}$$

For instance, $\epsilon_{1,3}^{\text{search}}$ quantifies the percentage change in search frequency of product 3 (a Sony Camcorder with HD media format, 10 × optical zoom, 2.7-Inch swivel screen, etc., at $654) with respect to percentage change in price of product 1 (a Sony Camcorder with DVD media format, 25 × optical zoom, 2.5-Inch swivel screen, etc., at $360). The top panel of Figure 5 shows the elasticity of search as well as the cross elasticity for product 1. First, note that its own elasticity is negative at -0.85. This means that if we increase price of product 1 by 1%, its own search frequency will drop by 0.85%. The price increase of product 1 results in a lower z_{ij}, letting its close competitors enter the search set. For a selection of competing products, we see that a price increase in product 1 has a positive influence on search frequency. For instance, if we make product 1 more expensive, the most closely matched Sony products and products with DVD media format are searched more. Second, not all products are affected by the product 1's price change. This makes sense in our modeling framework since products that are not close substitutes of product 1 will have a very low chance of entering the optimal search set together with product 1 even in the presence of product 1's price increase.

Demand A price change affects consumer demand in two ways. First, as explained above, a price change affects the optimal search sets of consumers. Second, the price change directly affects the consumer choice from the affected optimal search set. We compute own- and cross-price elasticities to further understand the competition among products. In this computation, we first predict the optimal search set for each individual i by computing V_{ij} and z_{ij}. Next, we predict the demand for product j by counting the number of times j was the highest utility product option in i's optimal search set.

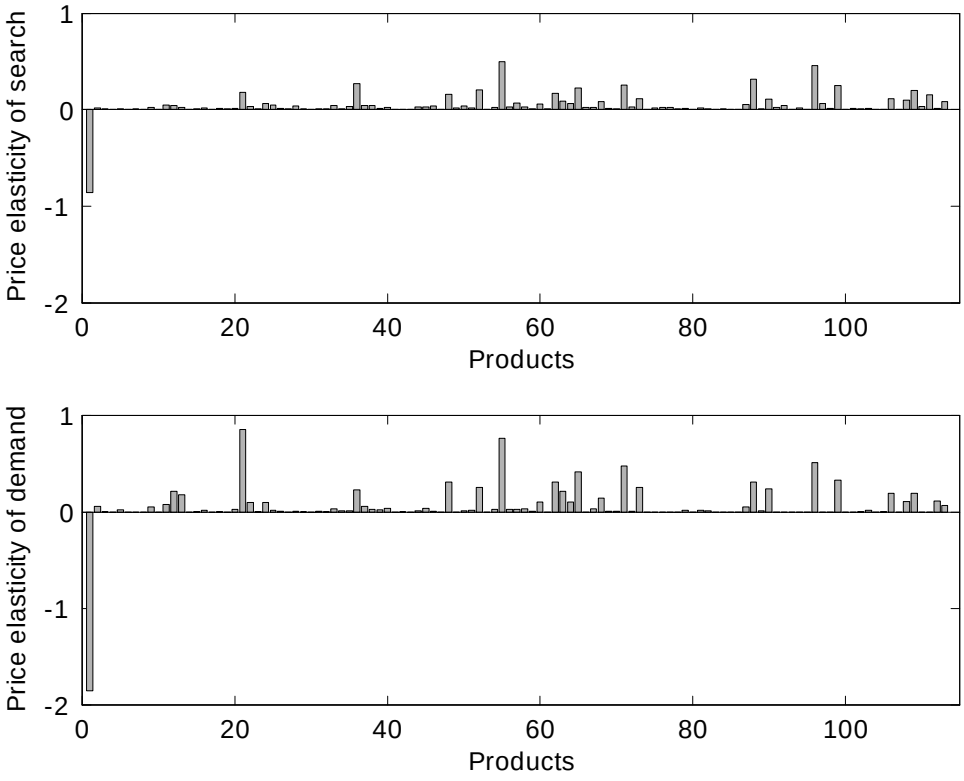

Figure 5: Price elasticity of search (top) and of demand (bottom) for the product 1 (SONY, DVD, Zoom 40, and $360).

The bottom panel of Figure 5 shows the demand elasticities. Own elasticity for product 1 is -1.8 and is significantly more negative than search elasticity. Indeed, whereas search can happen for reasons of low cost or of high V_{ij}, choice given search is impacted by a high V_{ij} but not by search cost. In other words, a price increase leads to both a direct share loss and an indirect share loss via search volume loss. Cross-elasticities are numerically zero for product pairs that are never viewed together. Also, we note that product 21 is predicted to have the largest cross elasticity with product 1. This has face validity, since product 21 is the most similar Sony to product 1, sharing the same media format and being the closest in price given media format. The findings strongly support the view that not all products are in direct competition with each other. Our approach can be used to predict and identify the set of products of direct competition both in the product search and in the product choice stages.

8 Counterfactual simulations

8.1 The effect of product links on consumer surplus

Providing product links selectively lowers search costs for some products but not for all. The net effect of such product recommendation is *a priori* not clear. On one hand, lowering search cost may increase consumer surplus if it facilitates consumers in finding their preferred products with less costs. On the other hand, it may lower consumer surplus if search costs are lowered on the wrong products or if lowered search costs result in disproportionately more search. To investigate these issues, we now analyze the role of product recommendations on consumer surplus. We do so by evaluating the effects of Amazon.com's links on consumers' search set formation and their subsequent choices. For this purpose, we simulate the optimal search sets and choices across population with and without Amazon.com's product links. We then compute the aggregate change in the net surplus across the population. The net surplus of a consumer with respect to a search set S_i is defined as the highest utility in the search set less the total search cost incurred in the formation of i's search set S_i.

$$NS(S_i) = \max_{j \in S_i} \{u_{ij}\} - \sum_{j \in S_i} c_{ij}. \tag{27}$$

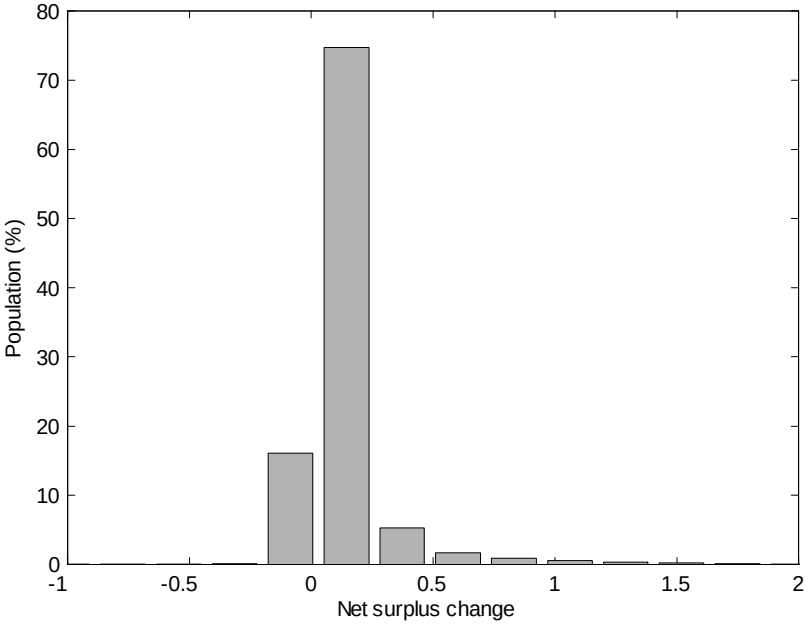

Figure 6: The change in surplus from providing consumers with product links.

The difference between the net surplus with and without the Amazon.com's product links for the entire population is computed as

$$\Delta_{NS} = \sum_{i=1}^{I} \Delta_{NS,i} \tag{28}$$

$$= \sum_{i=1}^{I} NS(S_i^*|L = \{L_j\}) - NS(S_i^*|L = \emptyset) \tag{29}$$

where L is the set of L_j, the number of links for product $j = 1, \ldots, J$ and $(S_i^*|L)$ is i's optimal search set given links L. The first term computes the net surplus across consumers under the presence of the Amazon.com's links while the second term computes the net benefits across consumers in a hypothetical case where Amazon.com does not provide any product links. Our main question is whether consumers are better off or not in the presence of Amazon.com's links to other products.

The distribution of $\Delta_{NS,i}$ is shown in Figure 6. From this simulation, we infer that the majority of consumers benefits from the reduced search cost through the presence of product links. About 84.3% of consumers experience positive net surplus while 7.6% of consumers experience negative net surplus. The rest of the consumers keep the surplus level unchanged.

	$\Delta u_i^* < 0$	$\Delta u_i^* = 0$	$\Delta u_i^* > 0$	Total
$\Delta c_i < 0$	0.4	76.9	7.1	84.4
$\Delta c_i = 0$	0.0	8.1	0.0	8.1
$\Delta c_i > 0$	0.0	7.3	0.2	7.5
Total	0.4	92.3	7.4	1.00

Table 6: Breakdown of utility and cost changes

We also find that in the absence of Amazon's links, the consumers generally search less. The median and mean search set size is 2 and 2.83, respectively. This makes sense because higher search cost discourages consumers from conducting more search.

To explain our findings, Table 6 provides a detailed breakdown of how the product links, L_j, change utility and search cost. From this table, we see that 76.9% of consumers fall in the cell where they choose the same products but incur lower total search costs ($\Delta u_i^* = 0$, $\Delta c_i < 0$) with Amazon.com's links. The finding that the majority of the consumers benefits in the presence of Amazon.com's links is intuitive. However, there is also a consumer segment (7.3%) who are worse off under the presence of the Amazon.com's links. From Table 6, the consumers who fall in the cell ($\Delta u_i^* = 0$, $\Delta c_i > 0$) are worse off since they choose the same utility products but incur higher search costs in the presence of Amazon.com's product links. For this group, the lower search cost disproportionately increases the reservation utilities, and hence, these individuals will search for disproportionately more products. So, whereas the per-product search cost may go down, total search cost goes up. The group of consumers who are worse off in the presence of links is summarized as those who are (1) highly sensitive to Amazon's recommendations, (2) whose demand parameters do not conform to those of "average" population and (3) who are interested in the product space that is crowded with relatively many products. It is important to understand the second point. The links at Amazon.com represent the market level demand of the consumers. If a consumer's preference conforms to the average preference, the links provide her with a "true" guide to the products she may be interested in. However, if her demand parameters deviate from that of the market average, she keeps receiving wrong recommendations. This justifies the personalization of the recommendations.

Next, we also notice that there is another consumer segment (7.1%) that achieves both higher quality choice ($\Delta u_i^* > 0$) and lower search cost ($\Delta c_i < 0$) in the presence of Amazon's links. These are individuals with (1) high baseline search cost (who would not search a lot in absence

of links) and (2) a large sensitivity to the links. Also, their preferences are consistent with what is recommended (i.e., popular products).

There are common factors and differences between the two groups who are worse off and who benefit the most. Both groups are highly sensitive to the Amazon.com's product links. The major difference is that the consumer who is worse off with the links was looking for a feature not commonly popular among the majority of the population. On the other hand, the consumer who benefited most has a preference that conforms to that of the average utility parameters. Therefore, we find that the recommendations based on market level information may negatively affect the consumers whose ideal product is far away from those of the "aggregate" consumer.

8.2 Market structure under full and limited search

The literature on information search has argued that limited information search can have a profound impact on market structure. One such example is when popular products are being recommended, thereby overstating their popularity at the expense of less popular and less recommended products. We empirically investigate this topic by comparing the market shares under two different scenarios: (1) the limited degree of search implied by the estimated search cost and (2) full search implied by zero search costs[20]. Given the ever growing presence of recommendations at many online stores, this counterfactual exercise helps manufacturers understand the direct impact of search costs on their products' performance.

Figure 7 shows the percentage difference in the market shares between the limited and the full search scenarios. A positive number in this figure means that the share under the limited (directed and sponsored) search is higher than under the full search. The products on the horizontal axis are sorted by sales-rank with popular products on the left, and low selling products on the right. The first conclusion from the top graph is that limited search on the Amazon.com website benefits the better selling items and harms the poorer selling items. The bottom graph shows that the number of incoming links is generally larger for better selling products. Combining the top and bottom panels in this figure, we further see that the percentage share difference is greater for the products that have less incoming links. Indeed, the presence of product links forms a double jeopardy to lower value products. Namely, not only do these products suffer from low preferences,

[20] If we assume that a consumer conducts a complete search over the entire product space, our model reduces to a standard probit model. By simulating, as before, a very large number of consumers using the estimated parameters but now with zero search cost, we can compute market shares under the full search scenario.

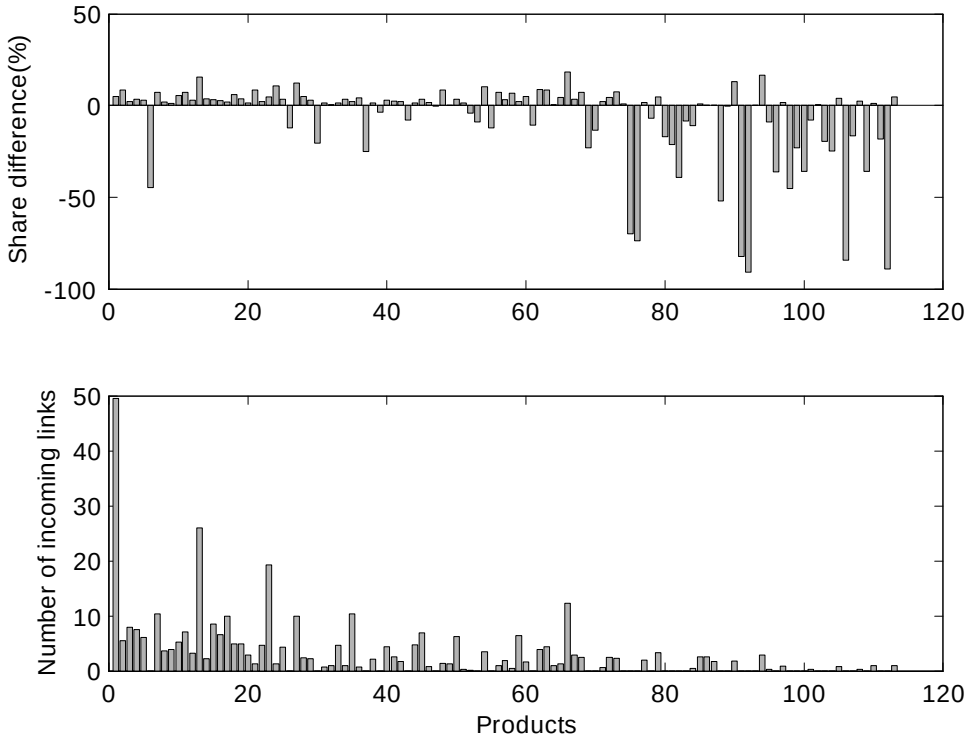

Figure 7: The impact of limited product search on market shares. The top panel shows the forecast of market share under limited search less the share under full search. The bottom panel shows the number of incoming links for the products.

they additionally suffer from high search cost (in the sense of not having many incoming links).

We find that the online market for camcorders is more concentrated with than without using product links. The Herfindahl concentration index with product links is equal to 0.0297. Without product links it becomes 0.0288 (a percentage drop of 3%). This may not seem like a large difference, but as Figure 7 illustrates, in general, shares of popular brands increase with positive search cost, and decrease with zero-search cost, frequently by as much as 10 to 15%. On the other hand, the demand for many lower selling camcorders would increase manifold if search cost were absent. In this sense, non-zero search cost tends to concentrate the online market for camcorders into demand for popular items. We note that this "polarization" effect of recommendation based on past popularity may be larger once we simulate over multiple time periods and allow the recommendation effect to accumulate and settle in. We intend to take this up in future research.

9 Discussion and conclusion

We study online consumer search and choice behavior in a durable goods context. Because evaluating a product for purchase takes time and effort, a consumer who seeks to maximize expected utility minus total search cost, needs to decide which product to search first, second, etc., and when to stop search. We have proposed a model of optimal search and choice for the analysis of online demand and consumer surplus in the case of durable search goods. Our model can estimate demand primitives for heterogeneous households, as well as a distribution of household specific search cost and its dependence on product recommendations or product links.

Our modeling framework has a number of important virtues. First, it constitutes an internally consistent theory of information search and demand for durable goods, integrated into a random utility choice framework. Second, the model has closed form expressions for the probability that a given brand is included in the search set. It also has closed form expressions for the probability distribution of entire search sets or consideration sets. Third, the model is not subject to the "curse of dimensionality." Rather, because aggregation of individual optimal search sequences and choices to market level demand outcomes is computationally efficient, the model allows for the occurrence of all possible search sets at the market level (a "blessing of dimensionality"), whereas, at the individual level, only as many search sets can be optimal as there are products to choose from. Fourth, the model is uniquely suited for the analysis of demand systems from widely available product search data, click-trail data, viewing frequencies, coincidence of pairs of

brands being viewed in the same session, or observed consideration sets.

We show with data experiments, that the model is identified and that the estimation strategy works well. From these data experiments, we conclude that it is possible to estimate demand systems with heterogeneity from product viewing data.

From an application of the model to the online market for camcorders at Amazon.com, we find that consumers do not search many products. This means that the majority of product pairs are never searched or considered together in an online choice setting. We find that whereas most consumers are better off with product links, households with atypical preferences may experience worse choice outcomes net of search costs.

We see three areas of further development of our model. First, the current model uses only the view-rank data. Amazon.com also provides sales-rank data. Recently, Bajari and Fox (2007) show how these data can be used in demand estimation. We hope that the combination of both types data may lead to further improvements in model estimation. For instance, we believe that it is possible to estimate aspects of the uncertainty that is associated with a particular product, i.e., to estimate σ_{ij} or factorizations thereof. This is because different σ_{ij} will shift the view-ranks, but should not impact sales-rank (see also Bajari and Fox 2007). Thus, where as view-rank data allow for the identification of the random effects choice based demand system, the combination of view and sales-rank data, may provide us with an opportunity to estimate some aspects of product uncertainty.

A second, but related, avenue for further development is to assume only partial consumer knowledge of attribute information and study how this affects choice and search outcomes. That is, in our model, the goal of search would remain to resolve the unknown component of utility e_{ij}, but we could allow that this component involves (part) of the product attributes. This way, it becomes possible to rigorously study the consequences for demand and market structure of lack of information and of expectations about prices or other attributes. We also see application of our model in research on the effects of advertising, e.g. in lowering search cost, on consideration.

A final avenue for future research is to analyze the long run implication of recommendations of popular products for industry concentration and the demand for new products. That is, if purchases are influenced by recommendation, future recommendations will depend on current recommendations, in addition to current demand.

A Derivation of the reservation utilities

The reservation utilities z_{ij} can be expressed rewriting the implicit equation 6 into

$$c_{ij} = \int_{z_{ij}}^{\inf} (u_{ij} - z_{ij}) f(u_{ij}) du_{ij} = (1 - F(z_{ij})) \left[\int_{z_{ij}}^{\inf} (u_{ij} - z_{ij}) \frac{f(u_{ij})}{1 - F(z_{ij})} du_{ij} \right], \quad (A.1)$$

with $F(z_{ij})$ equal to the cumulative probability distribution of u_{ij} evaluated at z_{ij}. The term in parenthesis after the second equality sign is the probability that, upon search, u_{ij} exceeds z_{ij}, whereas the term in square brackets is the expected value of the truncated distribution of $u_{ij} - z_{ij}$ given that u_{ij} larger than z_{ij}. Using the assumption of normality $u_{ij} \sim N(V_{ij}, \sigma_{ij})$ and substituting the expectation of a truncated normal distribution (e.g., Johnson and Kotz 1970) in equation A.1, we obtain,

$$c_{ij} = \left(1 - \Phi\left(\zeta_{ij}\right)\right) \left(V_{ij} - z_{ij} + \sigma_{ij} \frac{\phi\left(\zeta_{ij}\right)}{1 - \Phi\left(\zeta_{ij}\right)} \right),$$

with ϕ and Φ the standard Normal density and CDF, and $\zeta_{ij} = \frac{z_{ij} - V_{ij}}{\sigma_{ij}}$. Dividing both sides of the last equation by σ_{ij}, we need to solve z_{ij} out of the equation,

$$x_{ij} = \left(1 - \Phi\left(\zeta_{ij}\right)\right) \left(\frac{\phi\left(\zeta_{ij}\right)}{1 - \Phi\left(\zeta_{ij}\right)} - \zeta_{ij} \right),$$

where $x_{ij} = \frac{c_{ij}}{\sigma_{ij}}$. Finally, write the standard normal hazard rate $\frac{\phi(\zeta_{ij})}{1-\Phi(\zeta_{ij})}$ by $\lambda\left(\zeta_{ij}\right)$. It is noted that this hazard rate is the inverse of Mills' Ratio (Johnson and Kotz 1970). Dropping subscripts because this equation holds for any i and j, we obtain the following implicit equation.

$$x = (1 - \Phi(\zeta)) (\lambda(\zeta) - \zeta). \quad (A.2)$$

This equation is identical to equation (23) in the main text. Note that if we know ζ, we can compute $z = V + \zeta \times \sigma$ from the definition of $\zeta = \frac{z - V}{\sigma}$, above.

There are a number of properties of this equation that deserve further mentioning. First, and importantly, we propose to solve ζ out of equation (A.2), which expresses a function between two variables x and ζ, not directly involving any model parameters.

Second, from Barrow and Cohen (1954), we use that the derivative of the hazard rate (the inverse of Mills' Ratio) can be implicitly expressed as, $\lambda'(\zeta) = \lambda(\zeta)(\lambda(\zeta) - \zeta)$. Using this result and taking derivatives on both sides of A.2 with respect to ζ, we obtain that

$$\frac{\partial x}{\partial \zeta} = -(1 - \Phi(\zeta)). \quad (A.3)$$

Thus the derivative of x with respect to ζ is negative. Thus, x is a decreasing function of ζ. This implies also that ζ is a decreasing function of x. From monotonicity, solutions to equation (A.2) yields unique pairs of x and ζ.

Third, from results on Mills' ratio, $\lambda(\zeta) - \zeta > 0$ and $\lim_{\zeta \to \infty} \lambda(\zeta) = \zeta$. Therefore, as one expects, x, the normalized cost of search (cost , c, divided by product uncertainty, σ, is always positive. Also, one obvious solution to this equation is $x = 0$ and $\zeta = \inf$. Indeed, at 0 cost, the attractiveness to search an item is determined by the maximum upside of product utility which is $+\inf$.

The results above justify that we can construct a table of combinations of x and ζ that solve this equation. This table does not depend on model parameters, and therefore it can be solved once and it can subsequently be used in estimation, possibly with an interpolation step. Namely, at any stage in the estimation, we can use the current value for σ_{ij} and c_{ij} to compute $x_{ij} = \frac{c_{ij}}{\sigma_{ij}}$. Given x_{ij}, we can look up $\zeta(x_{ij})$ that solves equation A.2 and compute z_{ij} from the definition of z^*_{ij} and the current values for V_{ij} and σ_{ij}, i.e.,

$$z_{ij} = V_{ij} + \zeta\left(\frac{c_{ij}}{\sigma_{ij}}\right) \times \sigma_{ij} \tag{A.4}$$

With reference to equation (22) in the text, in this appendix we have shown that z_{ij} can be decomposed into the expected utility V_{ij} and fixed function of normalized search cost $\frac{c_{ij}}{\sigma_{ij}}$, which translates how product uncertainty is valued in search. The computational steps involved are trivial and inexpensive. The table of x and ζ can be solved fast for an arbitrarily fine grid on x. As this table is constructed outside the estimation, the marginal impact of extra precision in computing the z's on estimation time is 0.

References

[1] Albuquerque, Paulo and Bart J. Bronnenberg (2008),"Measuring Consumer Heterogeneity using Aggregated Data: An Application to the Frozen Pizza Industry," *Marketing Science*, forthcoming,

[2] Anderson, Simon P., and Régis Renault (1999), "Pricing, Product Diversity, and Search Costs: A Bertrand-Chamberlin-Diamond Model," *The RAND Journal of Economics*, 30 (4), 719-35.

[3] Bajari, Patrick and Jeremy Fox (2007), "Measuring the Efficiency of an FCC Spectrum Auction", Universty of Chicago, working paper.

[4] Bajari, Patrick., Jeremy Fox, and Stephen P. Ryan (2007), " Linear Regression Estimation of Discrete Choice Models with Nonparametric Random Coefficients", *American Economic Review*, 97 (2), 459-63.

[5] Barrow D. F., and A.C. Cohen (1954), "On Some Functions Involving Mill's Ratio," *The Annals of Mathematical Statistics,* 25 (2), 405-08.

[6] Berry, Steven, James Levinsohn, and Ariel Pakes (2004), "Differentiated Products Demand Systems from a Combination of Micro and Macro Data: The New Car Market," *Journal of Political Economy*, 112 (1), 68-105

[7] Bresnahan, Timothy F. (1989), "Competition and Collusion in the American Automobile Industry: The 1955 Price War", *The Journal of Industrial Economics*, 35 (4), 457-82.

[8] Bruno, Hernăn, A., and Naufel Vilcassim (2008), "Structural Demand Estimation with Varying Product Availability," *Marketing Science,* forthcoming.

[9] Chiang, Jeongwen, Siddhartha Chib and Chakravarthi Narasimhan (1999), "Markov Chain Monte Carlo and Models of Consideration Set and Parameter Heterogeneity," *Journal of Econometrics*, 89, 223-248.

[10] Diamond, Peter A., (1971), "A Model of Price Adjustment," *Journal of Economic Theory*, 3 (2), 158-68.

[11] Ellickson, Paul, B., Stephanie Houghton, and Christopher Timmins (2007), "Estimating Network Economies in Retail Chains: A Revealed Preference Approach", Duke University, working paper.

[12] Fox, Jeremy (2007), "Semiparametric Estimation of Multinomial Discrete Choice Models Using a Subset of Choices", *The RAND Journal of Economics, 38 (4), 1002-19.*

[13] Goeree, Michelle (2008), "Limited Information and Advertising in the US Personal Computer Industry", *Econometrica*, forthcoming.

[14] Gowrisankaran, Gautam, and Marc Rysman (2007), "Dynamics of Consumer Demand for New Durable Goods," *Washington University in St. Louis*, working paper.

[15] Gilbride, Timothy J. and Greg M. Allenby (2004) , "A Choice Model with Conjunctive, Disjunctive, and Compensatory Screening Rules," *Marketing Science*, 23 (3), 391-406.

[16] Hauser, John R., and Birger Wernerfelt (1990), "An Evaluation Cost Model of Consideration Sets," *The Journal of Consumer Research*, 16 (4), 393-408.

[17] Hong, Han and Matthew Shum (2006), "Using price distributions to estimate search cost," *The RND Journal of Economics,* 37(2), 257-75.

[18] Hortaçsu, Ali., and Chad Syverson (2004), "Product Differentiation, Search Costs, and Competition in the Mutual Fund Industry: A Case Study of S&P 500 Index Funds", *Quarterly Journal of Economics*, 119 (2), 403-56.

[19] Howard J. A., and J. N. Sheth (1969), "The theory of buyer behavior". , John Wiley & Sons Inc., New York, NY.

[20] Huang, Jen-Hung and Yi-Fen Chen (2006), "Herding in online product choice," *Psychology and Marketing,*

[21] Johnson, N. L., and S. Kotz (1970), "Distributions in statistics: continuous univariate distributions," 1, Wiley, New York.

[22] Roberts, John H., and James M. Lattin (1991), "Development and Testing of a Model of Consideration Set Composition," *Journal of Marketing Research*, 28 (4), 429-40.

[23] Manski, Charles F. (1975), "Semiparametric Analysis of Random Effects Linear Models from Binary Panel Data", *Econometrica*, 55(2), 357-62.

[24] Manski, Charles F. (1988), "Identification of Binary Response Models", *Journal of The American Statistical Association*, 83 (403), 729-38.

[25] Matzkin, Rosa. (1993), "Nonparametric Identification and Estimation of Polychotomous Choice Models," *Journal of Econometrics*, 58(1-2), 137-68.

[26] McCall, John J. (1965),"The Economics of Information and Optimal Stopping Rules", *The Journal of Business*, 38 (3), 300-17.

[27] Mehta N., Rajiv S. and Kannan Srinivasan (2003), "Price Uncertainty and Consumer Search: A Structural Model of Consideration Set Formation" *Marketing Science* , 22(1), 2003, 58-84

[28] Moe, Wendy (2003), "Buying, Searching, or Browsing: Differentiating Between Online Shoppers Using In-store Navigational Clickstream," *Journal of Consumer Psychology*, 13 (1 & 2), 29-39.

[29] Montgomery Alan, Shibo. Li, Kannan. Srinivasan, and John.C. Liechty (2004), "Modeling Online Browsing and Path Analysis Using Clickstream Data," *Marketing Science*, 23 (4), 579

[30] Moraga-Gonzáles, José, Louis, (2006), "Estimation of Search Cost", University of Groningen, working paper.

[31] Nelson, Philip (1970), "Information and Consumer Behavior," *The Journal of Political Economy*, 78 (2), 311-29.

[32] Nelson, Philip (1974), "Advertising as Information," *The Journal of Political Economy*, 82 (4), 729-54.

[33] Petrin, Amil. (2002), "Quantifying the Benefits of New Products: The Case of the Minivan," *Journal of Political Economy,* 110 (4), 705-29

[34] Reinganum, J. (1982), "Strategic Search Theory," *International Economic Review*, 23, 1-15.

[35] Reinganum, J. (1983), "Nash Equilibrium Search for the Best Alternative," *The Journal of Economic Theory*, 30, 139-152

[36] Senecal, Sylvain and Jacques Nantel (2004), "The influence of online product recommendations on consumers' online choices," *Journal of Retailing,* 80 (2), 159-69.

[37] Stigler, George J. (1961), "The Economics of Information," *The Journal of Political Economy*, 69 (3), 213-25.

[38] Weitzman, Martin L (1979), "Optimal Search for the Best Alternative," *Econometrica*, 47 (3), 641-54