Marketing Science

Online Demand Under Limited Consumer Search

Jun B. Kim, Paulo Albuquerque, Bart J. Bronnenberg,

Please scroll down for article—it is on subsequent pages

Online Demand Under Limited Consumer Search

Jun B. Kim
College of Management, Georgia Institute of Technology, Atlanta, Georgia 30308, jun.kim@mgt.gatech.edu

Paulo Albuquerque
Simon Graduate School of Business, University of Rochester, Rochester, New York 14627,
paulo.albuquerque@simon.rochester.edu

Bart J. Bronnenberg
CentER, Tilburg School of Economics and Management, University of Tilburg, 5037 AB Tilburg,
The Netherlands, bart.bronnenberg@uvt.nl

Using aggregate product search data from Amazon.com, we jointly estimate consumer information search and online demand for consumer 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 [Weitzman, M. L. 1979. Optimal search for the best alternative. *Econometrica* **47**(3) 641–654] and combines it with a choice model. It can accommodate highly complex demand patterns at the market level. 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 search costs 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 selectively lowered search costs on consumer surplus outcomes. We demonstrate that the demand estimates from our search model predict the actual product sales ranks. We find that consumer search for camcorders at Amazon.com is typically limited to 10–15 choice options and that this affects estimates of own and cross elasticities. In a policy simulation, we also find that the vast majority of the households benefit from Amazon.com's product recommendations via lower search costs.

Key words: cost-benefit analysis; optimal sequential search; demand for durable goods; information economics; consideration sets
History: Received: February 6, 2009; accepted: March 9, 2010; processed by Teck Ho. Published online in *Articles in Advance* June 24, 2010.

1. Introduction

Online demand for consumer durables and search goods is substantial and rapidly growing. Comscore (2007) estimates that nontravel U.S. online consumer spending in 2006 reached $102.1 billion. Jupiter Media Metrix estimated U.S. online consumer spending of $65 billion in 2004 (Rush 2004), with $20.2 billion on durable consumer search goods and another $8.3 billion on information goods. The 2007 Comscore report shows that the fastest-growing e-commerce categories include durable consumer 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 in the range of 25% to 50%. *PC World* reported in 2007 that the "appeal of online shopping" is growing; between August 2006 and August 2007, 14% of the $159 billion that U.S. shoppers spent on consumer electronics was spent online, up from 5% a year earlier,

according to the Consumer Electronics Association (Bertolucci 2007).

In this paper, we seek to understand online demand by studying the product information acquisition for durable search goods and/or information goods at Amazon.com using aggregated histories of search behavior, which provides us with a unique opportunity to directly observe category-level consumer browsing behaviors. Our premise is that we can learn about the preferences of consumers by studying their "shopping behaviors." That is, because the examination and inspection of goods or services come at the cost of the consumer's time and effort, search outcomes become informative about what the consumer wants. Observing the particular region of the attribute space in which a consumer invests her time browsing products may teach us something about her preferences. Specifically, our proposal 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 limited search on choice decisions of the consumers and on the competitive market structure. From a methodological standpoint, 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 a 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.[1] Table 1 does not list all existing 300+ camcorder options and reflects the fact that some options are seldom or 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. This paper shows that, across viewed products, the view-rank data are informative about substitution, and that from viewed to nonviewed products, the data imply either low substitution or a lack of substitution. For durable goods, where meaningful observations of consumer switching are usually very limited, 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 system. This would be of interest to practitioners and policy makers because the Amazon.com view-rank data are publicly available and contain cross-product information that is not present in reports of sales volume or market shares.

The general approach in this 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, taking into account the search costs of the alternatives that she inspects. At the individual level, our approach yields a probabilistic model of optimal search set formation, is not subject to the curse of dimensionality, and is purposely suited to be

Table 1 Product Alternatives for a Sony Camcorder Searched at Amazon.com in May 2007, Given Search of a Sony Camcorder

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

Notes. Camcorder features include DVD media format, 40x optical zoom, 2.5-inch swivel screen, etc. The camcorder sells for $328. HD, hard drive.

estimable using the 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.

From an application of our model to the Amazon.com camcorder category, we find the following results. The median (average) search set contains 11 (14) products, with about 40% of consumers searching for fewer than 5 products out of over 90 total products. We estimate that the cost of search is significant and is subject to consumer heterogeneity. The search cost is lowered for products that appear more frequently at Amazon.com, measured by the total number of references to the product. We also find that online competition between many products is effectively zero, because many products are not jointly searched by consumers. In fact, when looking at the estimated frequency of coviewership of two products, we find that the large majority of all possible product pairs, about 70%, is viewed together by less than 5% of the population. This implies limits on substitution, which in turn causes many cross-price elasticities to be numerically zero. Finally, our results show that almost everyone benefits from the product references that selectively lower search costs at Amazon.com.

The remaining portion of this paper is organized as follows. The next section reviews the background literature. Section 3 outlines the model.

---

[1] This data-generating mechanism is explained separately in §4.

Section 4 presents the data and discusses the Amazon.com's data generation process. Section 5 explains model operationalization and estimation and discusses empirical identification. Section 6 presents evidence from numerical experiments to show that the model is identified. Section 7 presents empirical results, model robustness checks, and model validations. Section 8 contains two policy experiments and describes managerial implications. Section 9 concludes.

2. Background

Marketing scholars and economists have long recognized that consumers do not in general search or consider the universal choice set, due to reasons such as nonzero search cost, product proliferation, and preference dispersion (e.g., Hauser and Wernerfelt 1990, 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 concerns exist that not taking into account the limited nature of choice sets leads to biased estimates of demand (Bruno and Vilcassim 2008, Chiang et al. 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 response to firms' actions. In this paper, we advocate that such responses can be measured in the context of how they affect the consumer's search strategies and that 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 in both 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 on consumer surplus.

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. There is already a tradition in the consideration

set literature to represent consideration sets as the outcome of nonsequential search (Roberts and Lattin 1991, Mehta et al. 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 because 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 the 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 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 the population. In the latter, the authors develop a model to translate the utility distribution into heterogenous search cost. 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, we believe that information about which products tend to be viewed together allows us to estimate heterogeneous consumer preferences in a differentiated product category.[3]

With our choice model that includes optimal sequential search, we seek to explore the influence of retailer product recommendations, a mechanism to selectively lower search costs, on consumer search behavior as well as and its impact on market structure. 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

---

[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 cannot be executed, and therefore they are not considered here.

[3] For a comprehensive review of several empirical applications, see Moraga-Gonzáles (2006).

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 retailer recommendations will significantly 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 $j = 1, \ldots, J$ that is equal to

$$u_{ij} = V_{ij} + e_{ij}, \qquad (1)$$

with

$$V_{ij} = X_j b_i,$$
$$b_i \sim \mathrm{N}(b, B),$$
$$e_{ij} \sim \mathrm{N}(0, \sigma_{ij}^2),$$

where X_j is a row vector of product characteristics and b_i is a vector that represents individual-specific sensitivities to product characteristics. We assume that 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 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}.[4] The most relevant attributes, whose values are defined by V_{ij}, are accessible from general-category information displays without retrieving the product detail Web page,[5] thus facilitating the existence of an expectation V_{ij} prior to search. Before accessing a product page, knowledgeable consumers may have lower variance e_{ij}s and less-knowledgeable consumers may have higher variance e_{ij}s. When 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.[6] We model search cost as a lognormally distributed random effect

$$c_{ij} \propto \exp(L_j \gamma_i), \qquad (2)$$

with

$$\gamma_i \sim \mathrm{N}(\gamma, \Gamma),$$

where the matrix Γ is diagonal. The lognormal specification ensures that the sign of c_{ij} is positive, consistent with theory. The cost attributes L_j describe, for instance, the accessibility of product j and are assumed to be known by the consumers. For instance, it may contain the appearance frequency of product j at the store or the number of times it is recommended.

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 cost of search. The objective of the analyst is to estimate b, B, γ, and Γ from data.[7]

3.2. A Model of Sequential Search

In sequential search, a consumer decides to stop or continue a search each time after having searched a product. The theory of optimal sequential search states that consumers only continue a 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 products thus far. The consumer's expected marginal benefit from the search of product j is

$$\mathcal{B}_{ij}(u_i^*) = \int_{u_i^*}^{\infty} (u_{ij} - u_i^*) f(u_{ij}) \, du_{ij}, \qquad (3)$$

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^*.[8] Note that the

---

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

[5] In the digital camcorder category page at Amazon.com, consumers have access to important product characteristics in camcorders, such as brand, price, media format, zoom, pixel number, and the dimension for all products in this category.

[6] 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 (Roberts and Lattin 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 search cost as the opportunity cost of time invested in identifying and evaluating another candidate product.

[7] In the empirical analysis, we will assume that $\sigma_{ij}^2 = 1$, but in this section we wish to keep the level of product uncertainty general. The issue of estimating product-level uncertainty, although potentially important, is beyond the scope of this paper.

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

$$\mathcal{B}_{ij}(u_i^*) = (1 - F(u_i^*)) \times \int_{u_i^*}^{\infty} (u_{ij} - u_i^*) \frac{f(u_{ij})}{(1 - F(u_i^*))} \, du_{ij},$$

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 product-specific uncertainties, captured by σ_{ij}^2. This implies that the consumer continues search if there exists at least one j such that

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

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 nonsearched options. All decision-relevant information about S_i is contained in $u_i^* = \max_{j \in S_i}\{u_{ij}, e_{i(J+1)}\}$, provided we assign zero to the deterministic component of utility for the outside good.

At any point in the search process, the state of the system is given by $(u_i^*, \bar{S}_i)$. Define the value function $W(u_i^*, \bar{S}_i)$ 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):

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

where $F(\cdot)$ is the cumulative density function (CDF) of u_{ij}.

This equation shows that from state $(u_i^*, \bar{S}_i)$, the consumer can terminate search and collect u_i^* or can search any $j \in \bar{S}_i$. In the latter case, the consumer gets an expectation $F(u_i^*) \cdot W(u_i^*, \bar{S}_i - \{j\}) + \int_{u_i^*}^{\infty} W(u_{ij}, \bar{S}_i - \{j\}) f(u_{ij}) \, du_{ij}$, which she seeks to maximize across j. Because all online searches in a single session are conducted in a short time span, we set the discount rate β_i to one.

Now we discuss some important modeling assumptions in our proposed search model. First, our model

is a full-information model in which the consumers are assumed to have full knowledge about the products and their attribute values.[9] This allows the consumers to form V_{ij} for all products prior to search and to use them in computing the reservation utilities during the sequential search process. In §7, we conduct a series of robustness tests in which we relax the aforementioned assumption. In these robustness tests, we assume that consumers have partial knowledge about the products and investigate whether this partial knowledge assumption meaningfully affects our model estimates. Second, we assume that $E(e_{ij} \cdot e_{kl}) = 0$ and thus that the correlation of the unobserved portion of the utility across individuals and products is zero. The corresponding consumer behavior underlying this assumption is that the consumers have well-defined preferences prior to search and do not learn about the products during the search process.[10] Third, we do not include any context or reference effect in the proposed model of optimal consumer search. Although we acknowledge that a more comprehensive model should include such effects, we use the cost-benefit framework as the first-order approximation of the optimal consumer search. Last, we do not model Amazon.com's (potentially) strategic behavior in setting prices and product information (e_{ij}).[11]

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)):

$$c_{ij} = \mathcal{B}_{ij}(z_{ij}) = \int_{z_{ij}}^{\infty} (u_{ij} - z_{ij}) f(u_{ij}) \, du_{ij}. \qquad (6)$$

Thus, the reservation utilities solve $z_{ij} = \mathcal{B}_{ij}^{-1}(c_{ij})$. The estimation section establishes that $\mathcal{B}_{ij}$ is monotonic,

---

[9] In §7, we check the robustness of our results against assumptions of more limited consumer knowledge.

[10] We discuss this topic in detail in the next section.

[11] Amazon.com does not differentiate prices or product information across consumers. We allow for flexible product fixed effects in the model, thereby reducing the potential for endogeneity biases stemming from potentially strategic prices or supply of product information. We acknowledge that the issue is important and warrants future study. The issue of how much and which product information Amazon.com should supply to maximize profits is interesting but is outside the scope of this paper.

---

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^*.

and the appendix provides the details of computation of z_{ij}, including its existence and uniqueness.

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 greater than $\max_{j \in \bar{S}_i}(z_{ij})$ among the unsearched items.

(3) *Choice rule*: Once the 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 $\max_j(z_{ij}) < 0$ for all j will not find it worth their time to search brands. They will choose the outside good.[12] 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 consumers will also choose the outside good.

We assume that the optimal selection and stopping rules mentioned above are derived assuming that information obtained by searching one product does not affect the knowledge of other products. That is, we do not assume consumer learning during the search process. From the modeling perspective, we assume that e_{ij} are independent given V_{ij} during the search process. The current consumer behavior literature advocates that this is a reasonable assumption for consumers engaging in search processes (Moorthy et al. 1997).

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, a 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} do not follow a normal distribution with a mean of 0 and a variance of σ_{ij}^2. Therefore, given a search, choice probabilities do not follow a standard probit.

Second, Chiang et al. (1999) mention that identification of choice sets (or in this case, search sets) is subject to the curse of dimensionality. Indeed, in a nonsequential search process, with J possible alternatives,

there exist $2^J - 1$ 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 greater than or equal to 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 optimal 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 and measures the change in the reservation utility z_{ij}. For reference, in this example we choose $V_{ij} = 1$ and $\sigma_{ij}^2 = 1$. The reservation utility z_{ij}, or, more intuitively, the relative attractiveness 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 panel shows the option value of uncertain 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}.

---

[12] Note that because we estimate our model with search data, we assume that all consumers search at least one product. However, the model actually accommodates nonsearch behavior.

Figure 1   The Relation Between Search Cost c, Product Uncertainty σ^2, and Search Attractiveness z
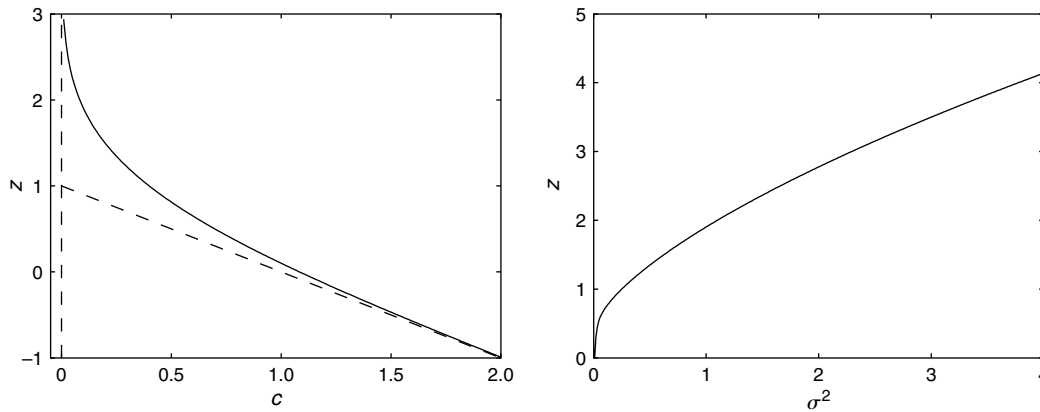
3.5.   Inclusion Probabilities and Set Occurrence

Our data 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 jth highest-ranked z_{ij}.

The contents of set S_{ik} are 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 mentioned previously). Thus, the inclusion probability of product $r(j)$ is

$$\pi_{i,r(j)} = \Pr\left(\max_{k=1}^{j-1}(V_{i,r(k)} + e_{i,r(k)}) < z_{i,r(j)}\right) \quad (7)$$

$$= \prod_{k=1}^{j-1} F(z_{i,r(j)} - V_{i,r(k)}), \quad j > 1, \quad (8)$$

where $\pi_{i,r(1)} = 1$[13] 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 that the inclusion probability of the $(j+1)$th product is always less than the inclusion probability of the jth product.

(2) Second, given the sequential nature of search and the selection rules 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:

$$\pi_{i,\{r(j) \text{ and } r(j+k)\}} = \pi_{i,r(j+k)} = \min(\pi_{i,r(j)}, \pi_{i,r(j+k)}), \quad (9)$$

where the last step is from the first property. In §5, we will use the last formulation of this property when we need to determine the probability that two products, 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 $r(k)$. 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

$$\Pr(S_{i,r(k)}) = \pi_{i,r(k)} - \pi_{i,r(k+1)}. \quad (10)$$

This concludes the statement of the individual-level model. The aggregation to the level at which Amazon.com reports its data is explained in §5. For completeness, it is also explained that an alternative to using the probabilities of Equation (8) in estimation would be 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. Using the model above, however, we can integrate over a probability model with far greater precision.

4.   Data

4.1.   The View-Rank Data

We have collected the view-rank data for all camcorder products from May 2006 until October 2007

---

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

on a regular basis.[14] 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 use the products that appear throughout the data collection period and aggregated their rank orders in the view-rank data to the monthly level. We use data from the month of May 2007.

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 view lists over time. For this averaging procedure, we use the percentile ranking similar to Bajari et al. (2008). 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

$$\hat{r}_{jt} = \frac{r_{jt}}{\max_k\{r_{kt}\}}. \tag{11}$$

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; $\hat{r}_j = (1/T)\sum_t \hat{r}_{jt}$.

At first, we extracted the top 200 camcorders from the Amazon.com website, based on sales rank. We removed the smaller players such as Aiptek, 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[15] and all camcorders of professional grade. After applying these data filters, we are left with 91 choice alternatives. The summary statistics of the products are shown in Table 2.

All 91 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 24 out of 90 times on other product's view list with a standard deviation of 18. The minimum number of appearances is 0, whereas the maximum is 83.

Table 3 gives the results of a descriptive regression of the number of appearances on the view lists. Note

Table 2   Description of the Choice Options in the Empirical Data

Attributes	Ranges
Brand	Sony (31), Panasonic (19), Canon (15), JVC (15), other (11)
Media formats	MiniDV (33), DVD (30), FM (9), HD (19)
Price	$530 (mean), $263 (std. dev.)
Form	Compact (8), Conventional (83)
High definition	Yes (14), No (77)
Pixel	1.67M (mean), 1.45M (std. dev.)
Zoom	19.8 (mean), 10.9 (std. dev.)

Note. Frequency of occurrence appears in parentheses.

Table 3   Descriptive Regression of the Frequency of Appearance in View Lists Against Product Characteristics

Variable	β	Std. err.
Intercept	−10.64	22.71
Sony	51.80	15.16
Panasonic	48.61	14.56
Canon	44.61	14.09
JVC	30.42	14.51
Samsung	36.09	12.79
MiniDV	−13.42	4.21
DVD	−22.58	4.34
FM	−18.69	9.00
Compact	1.80	8.88
High definition	16.10	5.38
Zoom	0.17	0.20
Screen size	1.28	6.28
Pixel	6.49	1.64
Price	−36.40	9.28
Frequency on purchase list[a]	1.65	0.23
R^2	0.68	

[a]This is the number of times a product A is listed as the one ultimately purchased on the viewing pages of all other products B.

that Sony, Panasonic, and Canon appear most frequently in the view-rank lists. Furthermore, high definition and pixel size improve the number of appearances, whereas higher price reduces it. We conclude that the number of appearances on the view-rank data depends on demand drivers such as product attributes and prices.

We point out the rich information embedded in the Amazon.com 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.[16] 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.

Last, we discuss the type of consumers who we believe are represented in the product search data.

---

[14] We used an automated process for the data collection. Every other day, using the category- and product-specific URLs, we downloaded the product details and product-viewing pages of all the products in the camcorder category. We then parsed all the downloaded files and retrieved the necessary data used in the empirical analysis. We repeated this process over time and constructed a longitudinal database.

[15] The media formats of the products in the data include flash memory and hard drive.

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

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 frequently visit the product page, whereas the consumers in the mode of search and deliberation focus on both product and category pages. Hedonic browsers focus on category pages, whereas 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 the buying phase or the 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. On each product detail page, Amazon.com lists up to four top products purchased by past consumers who searched the product in the current detail page. These product references serve as shortcuts to other closely relevant products, thereby reducing consumers' search costs for potentially attractive products. We use the number or frequency of appearances of each product aggregated across other product pages as an explanatory variable that affects search cost. For instance, we hypothesize that a product that appears frequently at the store level will have a smaller search cost compared to products that do not. Hereafter, L_j denotes the frequency of appearances of product j over all product pages.[17]

4.3. Amazon.com's Generation of View-Rank Data
Amazon.com lists a set of closely related products for each focal product in the order of decreasing "strength" of relationship. According to the Amazon.com U.S. Patent 6,912,505 (Linden et al. 2001), 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

$$\mathrm{CI}_{jk} = \frac{n_{jk}}{\sqrt{n_j} \cdot \sqrt{n_k}}, \qquad (12)$$

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 zero and one. The higher the commonality index is, the stronger the relationship is between two products. Amazon.com orders its view-rank data, exemplified in Table 1, according to the computed CI_{jk} for each product. Therefore, if the commonality index between product j and k is larger than that 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 $(j, k) \succ (j, l)$, then

$$(j, k) \succ (j, l) \longleftrightarrow \mathrm{CI}_{jk} > \mathrm{CI}_{jl}. \qquad (13)$$

From the view lists in our data, these inequalities are observed directly.[18] We treat these pairwise inequalities as the dependent variables in our analysis. To this end, we use the indicator variables, $I_{j,kl}$, defined as

$$I_{j, kl} = \begin{cases} 1 & \text{if } (j, k) \succ (j, l), \\ 0 & \text{otherwise.} \end{cases} \qquad (14)$$

For each product j, there are $0.5 \times (J - 1) \times (J - 2)$ unique inequalities defined by (13). 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 previously, our empirical study involves 91 products. For these products, and taking into account that Amazon.com may truncate view lists, the total number of observed pairwise ranks is 168,336. The rank information may contain less information compared to continuous data such as the share information. However, there are many observations of view ranks, and given the search model, such data are informative about substitution and consumer heterogeneity.

---

[17] It is important to ensure that there is no multicollinearity between L_j and other product characteristics we use in our empirical analysis. To verify this, we regressed L_j on the product attributes such as brands, media formats, prices, etc. We find that the model fit is relatively low ($R^2 = 0.12$) and none of the coefficients are significant, thereby supporting a lack of or low level of multicollinearity between L_j and other product characteristics.

[18] Not all pairwise combinations are observed. Some products are viewed so infrequently that the view list is truncated by Amazon.com. Although 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 nonlisted products.

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 (12). Below we show how to compute these forecasts. Next, we assume that the model's forecast of CI_{jk} differs from its true value by a zero-mean error process and construct a nonlinear least-square estimator using the observed $\{I_{j,kl}\}$ as dependent variables.

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

$$CI_{jk} = \widehat{CI}_{jk}(\Theta; X) - \epsilon_{jk}, \qquad (15)$$

where CI_{jk} is the unobserved, true commonality index; $\widehat{CI}_{jk}(\Theta; X)$ is the model's prediction of the commonality index given data X and parameters Θ; and $\epsilon_{jk} \overset{iid}{\sim} N(0, v^2/2)$. This error term allows for aggregate-level prediction errors, similar to Bresnahan (1987) and Bajari et al. (2007).

Next, we tie the above model 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

$$Pr(I_{j,kl} = 1) = Pr(CI_{jl} < CI_{jk}) = Pr(\epsilon_{j,kl} < \widehat{CI}_{jk} - \widehat{CI}_{jl}), \quad (16)$$

where $\epsilon_{j,kl} = \epsilon_{jk} - \epsilon_{jl}$. Given the assumptions about the distribution of ϵ_{jk}, the error term, $\epsilon_{j,kl}$, is an independent and identically distributed (i.i.d.) normal random variable with a mean of 0 and a variance of v^2. Therefore,

$$Pr(I_{j,kl} = 1) = \Phi\left(\frac{\widehat{CI}_{jk}(\Theta; X) - \widehat{CI}_{jl}(\Theta; X)}{v}\right), \qquad (17)$$

where Φ is the CDF for the standard normal distribution.

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 nonlinear least-square estimator in our empirical analysis is defined as

$$\{\Theta^*, v^*\} = \underset{\{\Theta, v\}}{\arg\min} \sum_{(j,k,l) \in \mathbf{S}} [Pr(I_{j,kl} = 1) - 1]^2, \qquad (18)$$

where $j, k, l = 1, \ldots, J$, and $\mathbf{S} = \{(j, k, l) \mid I_{j,kl} = 1, j \neq k \neq l\}$. In the next subsection, 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.

The "minimum" operator in Equation (9) may generate discontinuities in the objective function. To find the global optimum value, we first use differential evolution (Price et al. 2005, Bajari et al. 2008), a stochastic search algorithm, to obtain good initial estimates. Second, with these estimates as starting values, we obtain the final parameter values using a gradient-based search algorithm. We tested this approach using Monte Carlo simulations and found that it works very well to obtain the global minimum of our objective function.

Finally, for our empirical analysis, we lack data on the size of the outside goods in search and choice. In addition, the nature of our dependent variables—relative joint search popularity—does not allow us to directly measure the size of the outside good. Hence, in our empirical analysis, we are confined to inference on our model conditional on consumer search and choice within the category. This is a limitation of the data and not of our model, because the modeling approach explicitly accommodates outside goods, both in search and in choice (see §3.3).

5.2. Computational Details

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

$$\widehat{CI}_{jk}(\Theta; X) = \frac{\hat{n}_{jk}}{\sqrt{\hat{n}_j}\sqrt{\hat{n}_k}}, \qquad (19)$$

where $\hat{n}_j$ is equal to the forecasted number of simulated individuals that have searched j (given X and Θ), and $\hat{n}_{jk}$ is equal to the forecasted number of individuals who have 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 pseudohouseholds (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 (8),

$$\hat{n}_j = \sum_{i=1,\ldots,I} \pi_{ij}, \quad j = 1, \ldots, J, \qquad (20)$$

where I is the total number of simulated individuals. Furthermore, 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 (9),

$$\hat{n}_{jk} = \sum_{i=1, \dots, I} \min(\pi_{ij}, \pi_{ik}), \quad j, k = 1, \dots, J. \qquad (21)$$

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

$$\widehat{\text{CI}}_{jk}(\Theta; X) = \frac{\sum_i \min(\pi_{ij}, \pi_{ik})}{\sqrt{\sum_i \pi_{ij}}\sqrt{\sum_i \pi_{ik}}}. \qquad (22)$$

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

5.2.2. Computing Reservation Utilities. The right-hand side of Equation (22) 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 the appendix. First, the reservation utilities follow

$$z_{ij} = V_{ij} + \zeta\left(\frac{c_{ij}}{\sigma_{ij}}\right) \times \sigma_{ij}, \qquad (23)$$

where $\zeta(c_{ij}/\sigma_{ij})$ 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 $\zeta(x)$ solves the following implicit equation:

$$x = (1 - \Phi(\zeta))(\lambda(\zeta) - \zeta), \qquad (24)$$

where Φ is the cumulative standard normal distribution and λ is the standard normal hazard rate, $\phi(\zeta)/(1 - \Phi(\zeta))$, in which ϕ is the standard normal probability distribution function. The function $x(\zeta)$ in Equation (24) is further shown to be continuous and monotonic. Hence, the inversion to the function $\zeta(x)$ exists. Although the precise solution of Equation (24) is computationally expensive, it can be solved once for a large set of x outside the estimation algorithm. That is, $\zeta(x)$ does not need to be solved in estimation because it does not directly involve any model parameters. Armed with a table of x and $\zeta(x)$, we can—during estimation—substitute $x = c_{ij}/\sigma_{ij}$ and look up $\zeta(x)$ from the table using an interpolation step if the table of $\zeta()$ only covers a neighborhood of $x = c_{ij}/\sigma_{ij}$. 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 an intuitive appeal. If search cost c_{ij} is low relative to product uncertainty σ_{ij}, $\zeta(c_{ij}/\sigma_{ij})$ 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 $\zeta(c_{ij}/\sigma_{ij})$ turns out to become negative and the reservation utility z_{ij} is less than V_{ij}.[19] 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} are greater than the low reservation utilities z_{ik} of products in the set of unsearched products k.

5.3. Inference

We use resampling techniques for our standard error computation. Statistical inference based on the bootstrap resampling technique (Efron and Tibshirani 1993) has been widely used in many disciplines. In marketing, Horsky and Nelson (2006) used the resampling to compute the standard errors in the presence of dependencies among the dependent variables.

The basic idea behind the bootstrap is very simple. We first generate a random sample of same size by drawing, with replacement, from the original pairwise rank indicators. Such a random sample is considered as a perturbation from the original data. Next, we estimate the model parameters with the drawn random sample. We repeat the random sampling and estimation, and use the distribution of parameter estimates to compute the standard errors. Using bootstrap samples, the standard error of a parameter β_k is computed as

$$s(\beta_k) = \sqrt{\frac{1}{B-1} \sum_{b=1}^{B} \left(\hat{\beta}_k^b - \frac{1}{B}\sum_{b=1}^{B} \hat{\beta}_k^b\right)^2}, \qquad (25)$$

where $\hat{\beta}_k^b$ is an estimate of β_k from bootstrap sample b and B is the total number of bootstrap samples.

In our context, we generate a random sample from the view-rank data, $\{(j, k)^b \mid j, k = 1, \dots, J, j \neq k\}$, by sampling with replacement from the original view-rank data $\{(j, k) \mid j, k = 1, \dots, J, j \neq k\}$, and we estimate $\{\Theta^b, v^b\}$ using a bootstrap sample $\{(j, k)^b \mid \cdot\}$. We repeat the estimation with a set of bootstrap resamples and use the distribution of $\{\Theta^b, v^b\}$ to infer the standard errors of $\{\Theta^*, v^*\}$.

---

[19] 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 (23) is virtually identical for the uniform distribution.

5.4. Discussion of Identification

We discuss the empirical identification of the proposed model. First, we discuss the identification of consumer preferences. We note that the mean consumer preferences for product attributes (b) are identified by the correlation between the search frequencies of products inferred by aggregating the view-rank data *across* all products and the frequencies of underlying products' attributes. That is, the mean effect of an attribute is measured by how often the products that share the same attribute appear in the view-rank data relative to other products that do not share the attribute. We infer the search frequency of a product from the total number of times a product appears *across* other product's view lists and its relative rank positions to other products *within* the view list of each focal product in the view-rank data.

Second, there are two information sources for the identification of consumer heterogeneity (B). The consumer heterogeneity is partially identified by the discrepancy between our predicted search frequencies of products based solely on the mean consumer preferences and the (observed) search frequencies. Another unique source for the identification of heterogeneous consumer tastes in our data is the different relational strengths among the products. How often products are searched together or not searched together in the view-rank lists is very informative about how substitutable the underlying product attributes are. This unique information is critical in estimating the consumer heterogeneity.

Third, we discuss the identification of consumer search costs. In our search cost specification, we conjecture that the search cost is a function of "appearance frequency" of products. From the perspective of empirical identification, we can treat the appearance frequency of products as another product attribute. Therefore, we can extend our discussion on the identifications of mean and heterogeneous consumer preferences to those of any parameters that relate to search costs. We also depend on the nonlinear functional form in the reservation utility for the separate identification of consumer preference and search cost parameters. That is, the consumer preferences enter the equation in a nonlinear manner (because we need to integrate the utility out over the truncated error distribution) in Equation (3), whereas the search cost enters the equation in a linear manner. This nonlinearity helps us separately identify consumer preferences and search costs. Last, we point out that the mean consumer search cost is identified by the average search set size implied by the sparsity of products in the view-rank lists across products. Consumers will search all the products under zero search costs. Under this scenario, all the commonality indices will be unity, and we will observe uniform patterns in the view-rank data across all the products. This is not the case with the observed view-rank data. Therefore, the mean search cost is identified by the average consumer search set size that best matches the sparsity of an *average product* in the view-rank lists across products.

6. Data Experiment

We conduct a numerical experiment to verify that the model can be identified from view-rank data. To this end, we create 32 product options with attributes that were arbitrarily named "Sony," "Panasonic," "Zoom > 10×," "Media: MiniDV," and "Media: Hard drive." In addition, we add a continuous attribute, "Price." Finally, we assign a value for product appearance frequency, L_j, to each of the 32 options.

We choose a random-effects specification on all product attributes and a fixed-effects specification for search cost, $c_{ij} = \exp(\gamma_0 + \gamma_1 L_j)$, and we assume a set of values for the parameters. We use $\sigma_{ij}^2 = 1$ as a normalization.[20] The assumed values for the parameters are chosen with similar magnitudes as those we obtained from preliminary empirical results. Next, we draw 50,000 pseudohouseholds from the distribution of parameters and, for each of these pseudohouseholds, compute the V_{ij}, z_{ij}, and other relevant quantities to obtain the optimal search sequence and choice. We then aggregate these sequences according to the recipe used by Amazon.com[21] to obtain the lists of view ranks similar to those exemplified in Table 1.

Before we discuss a more detailed identification experiment below, we first conduct a series of preliminary view-rank data generation exercises to ensure that the view-rank data actually change as a function of the assumed parameter values. As a demonstration, we show that the different parameters for Media: Hard drive (hereafter HD) generate different view-rank data. with the value of the Media: HD parameter set at 2 (as in Table 4), we observe that HD products have an average rank position of 7.5 (with a standard deviation of 1.6) on the view lists *across* all the products. When we change its mean effect to 3, hence making the HD more attractive to consumers,

---

[20] 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 data with sales-rank data.

[21] There are two ways to perform 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 the second method is likely more precise with fewer pseudohouseholds.

Table 4   Estimation Results from the Numerical Experiment

Parameter	True values	Estimated values (std. err.)
Mean effects		
Sony	1	1.10 (0.19)
Panasonic	−1	−0.65 (0.72)
Zoom > 10×	1	1.05 (0.22)
Media: MiniDV	−1	−0.70 (0.62)
Media: Hard drive	2	2.13 (0.20)
Price	−2	−2.14 (0.46)
Heterogeneity		
Sony	1	0.92 (0.20)
Panasonic	2	2.11 (0.22)
Zoom > 10×	1	1.03 (0.09)
Media: MiniDV	1	0.95 (0.14)
Media: Hard drive	2	2.11 (0.22)
Price	1	0.78 (0.78)
Cost		
Base cost	−2	−1.97 (0.24)
Effect of links	−3	−3.09 (0.36)
Standard deviation of CI		
v	0.14	0.15 (0.01)
Number of observations	20	
Sum of squared errors	2,390	

HD product's average rank position moves up to 7.25 (2.0). This implies that these products will be searched more often by the consumers. When we set HD mean parameter to 1 (hence making it less attractive), the average rank position goes down to 7.9 (1.2). We also observe that when we increase the HD heterogeneity to 3, HD product's average rank goes down to 8 (2). We report similar observations with other attributes.

We now discuss the parameter recovery experiment in detail. To estimate the model, we generate 3,000 pseudohouseholds. 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 (22), the forecast of the commonality index given a set of parameter values. During the estimation, the sum of squared errors in Equation (18) is minimized. The recovered parameters and their corresponding standard errors are shown in Table 4.

As can be seen from Table 4, the recovered parameters are all close to the actual values (well within sampling errors). The correlation between the actual and the estimated parameters is very high, around 0.99. These observations suggest that the model parameters are identifiable from view-rank data. We call attention to the estimated values of heterogeneity. The variances of the distribution of the random effects are well recovered. This means that taste variation across consumers in the model is 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 seems to correctly reproduce the value of the cost parameters. 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.

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

$$u_{ij} = X_j \beta_i - \alpha_i P_j + e_{ij}, \qquad (26)$$

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 normally distributed random error term with mean 0 and variance σ^2 and is i.i.d. across individuals and products. We set $\sigma^2 = 1$ for all i and j. We include $K = 7$ product characteristics in the utility specification.[22] We additionally specify random coefficients on all these effects. For reasons of parsimony, we use one common heterogeneity parameter for all brands, as well as for all media formats. Further, we impose a theory-driven restriction on the price coefficient.[23] In particular,

$$\log(\alpha_i) \sim \mathrm{N}(\beta_p, \sigma_p^2),$$

$$\beta_i \sim \mathrm{N}(\beta_0, \Sigma_\beta),$$

[22] We include brand, media format, form, high definition, zoom, screen size, and pixels. It is possible that factors outside Amazon.com, such as advertising, could affect consumer search. However, given the nature of the data in our empirical analysis, we restrict our modeling effort to the product information available at Amazon.com.

[23] Past literature supports theory-driven empirical specification (Boatwright et al. 1999). However, we also estimate the proposed empirical model without the sign restriction on the price coefficient, treating it as a normally distributed random variable with mean and variance. The estimated mean and standard deviation are −0.8 and 2.2, respectively. This implies some consumers with positive price coefficients. However, the managerial implication remains substantially the same because computed own-price elasticities between these two models are very close with a correlation of 0.99. In addition, their magnitudes are very similar because the mean and median of the ratios of own-price elasticities between the two models are 1.08 and 1.04, respectively.

where Σ is a diagonal matrix containing the variances of the random effects, σ_k^2. Search cost is specified[24] as

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

$$\gamma_{1i} \sim N(\gamma_1, \sigma_{\gamma 1}^2),$$

where L_j is the appearance frequency of or references to product j at Amazon.com. 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 frequent appearances of products while searching, and hence this segment of consumers will have low γ_{1i}.

7.2. Parameter Estimates

The estimated parameters are shown in Table 5.[25] As mentioned in §5.1 on estimation, our model estimates and inference hereafter are conditional on consumer search and (inferred) optimal choice. As was noted in §5.1, this is not the limitation of our empirical model because our empirical model can explicitly accommodate outside goods both in search and in choice.

We check the face validity of the estimated parameters. Sony, the most popular brand, has the largest mean brand effect. Other well-known brands (Panasonic and Canon) have higher mean brand effects than lesser-known brands (JVC and Samsung) in the camcorder category. Second, the number of product appearance frequency at Amazon.com decreases the search cost, as can be inferred from the sign of the coefficient γ_1, which is negative. Among the media formats, we find that DVD and MiniDV are more popular than FM. We compare our findings with Gowrisankaran and Rysman (2009), who also estimated demand parameters in the camcorder category. Our finding on consumer sensitivity to pixel is consistent with their dynamic demand analysis that pixel is more important compared to zoom or screen size. We find a large degree of consumer heterogeneity present for brands, media formats, form (compact), high definition, and price.[26]

---

[24] We acknowledge that there may be other factors that affect consumer search cost, such as the way Amazon.com responds to consumer queries.

[25] For our estimation, we draw 1,500 consumers from the joint distribution. By testing different numbers of simulated consumers, we found that consumer size of over 1,000 provides stable parameter estimates.

[26] At the average price coefficient, the difference in valuation between Sony and Panasonic is $10.71, but the distribution of willingness to pay for Sony has a long right tail. We take this to mean that there are consumers of whom we infer that they are relatively price insensitive and would not buy a Panasonic over a Sony even at large price advantages of the former. We note that Meijer and Rouwendal (2006) and Sonnier et al. (2007) warn that inferences on ratios of two random coefficients in discrete-choice models should be made with some caution.

Table 5 Estimation Results

Variable	Mean effect (std. err.)	Heterogeneity (std. err.)
Sony	1.387 (0.363)	0.793[a] (0.047)
Panasonic	1.335 (0.349)	0.793 (0.047)
Canon	1.077 (0.351)	0.793 (0.047)
JVC	0.536 (0.317)	0.793 (0.047)
Samsung	0.453 (0.316)	0.793 (0.047)
MiniDV	−0.457 (0.101)	1.142[b] (0.083)
DVD	−0.585 (0.092)	1.142 (0.083)
FM	−1.439 (0.169)	1.142 (0.083)
Compact	−1.407 (0.374)	1.030 (0.141)
High definition	0.924 (0.133)	1.267 (0.091)
Zoom	0.006 (0.003)	0.019 (0.003)
Screen size	−0.307 (0.078)	0.224 (0.165)
Pixel	0.292 (0.040)	0.226 (0.028)
Log(price)	1.578 (0.319)	4.505 (0.767)
Search base cost (γ_0)	−5.076 (0.353)	0.064 (0.069)
Effect of appearance frequency (γ_1)	−1.365 (0.167)	0.001 (0.050)
Standard deviation of CI (v)	0.099 (0.002)	NA
Number of inequalities	168,336	
Sum of squared errors	11,693	

Note. Standard errors are in parentheses.
[a] Random-effects variance is common across brands.
[b] Random-effects variance is common across media formats.

7.3. Robustness of the Model Estimates

Our empirical model assumes that consumers have full knowledge about products in terms of their existence and their attribute values. We now relax this assumption and discuss how our model estimates change under the assumption that consumers have *partial* knowledge of the products. Our ultimate goal is to check the robustness of the model estimates under varying levels of consumer knowledge. To this end, we devise a 2×2 testing matrix. For each cell in the matrix, we assume different *contents* and *quantities* of consumer knowledge. In the *contents* dimension, the consumers have partial knowledge about the *existence of products* (limited product knowledge) or about the *attribute values* (limited attribute knowledge). In the *quantity* dimension, consumers have different *amounts* of knowledge. In *low* (*high*) condition, we assume that consumers are aware of 60% (80%) of all products' existence or attribute values. In the implementation, we randomly remove products or continuous attribute values across i and j. In the case of limited attribute knowledge, we assume that consumers impute the missing values using the category-level expectations over the remaining data. In the case of limited product knowledge, we assume that each consumer is aware of a random subset of products and conducts the optimal search on that subset.[27]

---

[27] Technically speaking, we set the utility of the randomly selected products to negative infinity.

Table 6 Parameter Correlations Under Varying Assumptions on Consumer Knowledge

	Low (60%)	High (80%)
Limited product	0.977	0.995
Limited attribute	0.969	0.990

We reestimate our models under varying assumptions and compare the parameter estimates thus obtained with those from our "full knowledge" model. We find high correlations between the parameter estimates in each of the four cells and very high correlations among the parameter estimates, ranging as high as 0.97~0.99. Table 6 shows the computed correlations among the parameter estimates for a 2×2 matrix. As noted, all cells exhibit high correlations, with the high-knowledge-level cells (80%) showing higher correlations than their low-knowledge counterparts. From this analysis, we conclude that our parameter estimates are robust to alternative assumptions about available products and levels of consumer knowledge. These results suggest that the estimates from the full model can be used to verify the implications of our model under a broader set of circumstances than under the strict completeness of information assumption that underlies the model development.

7.4. Model Validation

We conduct two out-of-sample model validations. First, we show how well our proposed search model predicts the actual sales ranks for the same month (May 2007). Because one of the appealing aspects of our proposed model is to study consumer demand using consumer search data, the proposed validation using the actual sales information is of critical interest. Second, we predict the consumer product search patterns for a different month and compare them against the observed consumer search data. This is a full out-of-sample validation including prediction of viewing patterns for new combinations of attributes.

Using the model estimates, we predict the market shares by aggregating the individual choices *inferred* from the optimal search processes. We compare our sales-rank predictions (obtained from the predicted market shares) against the actual sales ranks observable at Amazon.com. We report that the sales-rank correlation is 0.63, whereas the hit rates among all pairs of products, i.e., the fraction of correct predictions of the sales-rank comparison between a pair of products, is 0.72. It is informative to compare these numbers with the identical summary statistics from Roberts and Lattin (1991). Recall that Roberts and Lattin predicted market shares using individual-level consideration and choice data.[28] Using the individual consideration data, they achieve a sales-rank correlation and hit rate of 0.75 and 0.78, respectively. These numbers increase to 0.83 in both cases when choice data are used to predict share ranks, and so the predictive ability of the individual-level consideration data is comparable to that of the individual choice data. Thus, at least for the data in Roberts and Lattin, consumer consideration data appear to be a good proxy for aggregate sales predictions in the absence of such sales data. Comparing further, the predictive accuracy of our proposed model is similar to that from individual-level consideration data (hit rates of 0.72 versus 0.78) and to that from choice data (hit rates of 0.72 versus 0.83). We acknowledge that these are comparisons across different data sets but still submit that this is a notable achievement because our predictions are based on aggregate search data that are much coarser than the individual-level consideration data used in Roberts and Lattin.

As a second validation exercise, we predict and compare the consumer search patterns for the month of June 2007 from our model estimates. In contrast to the previous exercise, our goal in this exercise is to test the predictive ability of the model using the data from a different time duration. Among the 90 products in the June data, we observe seven exits from existing products and six entries of new products. In addition, we observe that overlapping products have different prices and appearance frequencies. We find that our model correctly predicts 89% of all the search patterns of pairwise inequalities $\{I_{j,kl}\}$ in the June data, and more important, it correctly predicts 88% of the view-rank lists of the six new products.[29] We believe that these validations, from two different data sources, demonstrate that our results are informative about demand in the camcorder category. We now report on several implications of our model estimates.

7.5. 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 pseudohouseholds from

---

[28] We obtain sales-rank predictions from their market share predictions in Table 3.

[29] We also used our demand estimates from the May 2007 data to predict the sales ranks for the month of June 2007. We find that our predicted sales rank correlates 0.60 with the actual sales ranks, and we observe a hit rate of 0.70. Both numbers are very close to the in-sample predictions stated in the previous paragraph. This out-of-sample validation further supports that the model's estimates are informative about demand.

Figure 2   The Estimated Population Distribution of Set Sizes
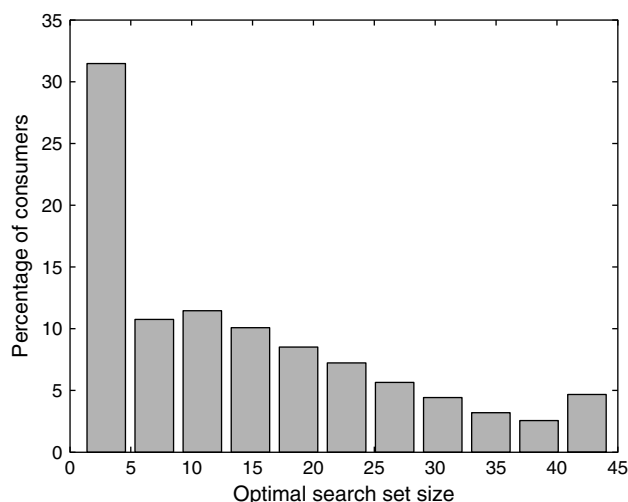
the set of estimated population density of parameters. For each of these pseudohouseholds, we compute expected utilities V_{ij} and reservation utilities z_{ij}. We next compute their optimal search sets and use various sample statistics to report on the size and contents of the individual-level search sets.

Figure 2 shows the distribution of the estimated number of products searched per individual. The mean of the search set size distribution is 14, whereas its median value is 11. A large proportion of the population, about 40%, has a small search set size of fewer than five products. Furthermore, the distribution of search set size is estimated to have a long right tail, as one might expect, given the relatively low cost of search.[30] The model implies search sets that are a priori realistic in size and that consumers generally do not search for very many products.

We now comment on some aspects of the contents of the inferred optimal search sets. First, we present a summary of 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 prepurchase phase. The left-hand panel of Figure 3 reveals that of all the products searched at Amazon.com, Sony accounts for about 35% of search volume. This bestows on Sony a large "mind share." Panasonic is second, followed by Canon. 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 1.1% of the total product search volume. Thus, from the two graphs, we conclude that the Sony dominates the search process of consumers at the brand level but is marginally behind Panasonic and Canon at per-product basis, which may be due its very long product line.

Next, we look at which brands are more frequently searched together. The left-hand panel in Figure 4 shows the joint search frequency between Sony and all other brands. The second bar indicates that more than 80% of all consumers who search at least one Sony product also search at least one Panasonic product. In the right-hand panel of Figure 4, we show the joint search frequency between Canon and all other products. We see that about 90% of Canon searchers also search at least one Sony or Panasonic product. Note that the conditional search shares are asymmetric; i.e., 75% of Sony searchers also search for a Canon, but 90% all of Canon searchers will also search for a Sony. Taken together, our results emphasize that joint search frequencies among the top three brands are high, whereas those with the rest of the brands are low. This questions whether assuming full information across all the *brands* in consumer choice is realistic.

Because our model is defined at the individual level, we can infer the joint search frequency among the *products* in a more granular manner. With 91 products, there are 4,095 product pairs. We find that 25% of the product pairs are searched by less than 1% of the population, and an additional 65% are viewed by more than 1% but less than 10%. Virtually all product pairs (98.5%) are predicted to be viewed by less than 20% of the population. We conclude from these numbers that the majority of products are 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.

7.6.   Substitution and Cross Elasticities

A price change affects consumer demand in two ways. First, it affects which brands enter 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 the products. To do so, we first predict the optimal search set for each individual i by computing V_{ij} and z_{ij}, and 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. By doing this at different price levels, we can compute arc elasticities. The mean and median values for

---

[30] Our model also estimates a strong degree of heterogeneity in consumer valuation of product search and that the unobserved component of utility that needs to be resolved through search is economically meaningful. That is, average consumers value one standard deviation of e_{ij} at approximately $203.

**Figure 3** Search Volume Share by Manufacturers and by Products
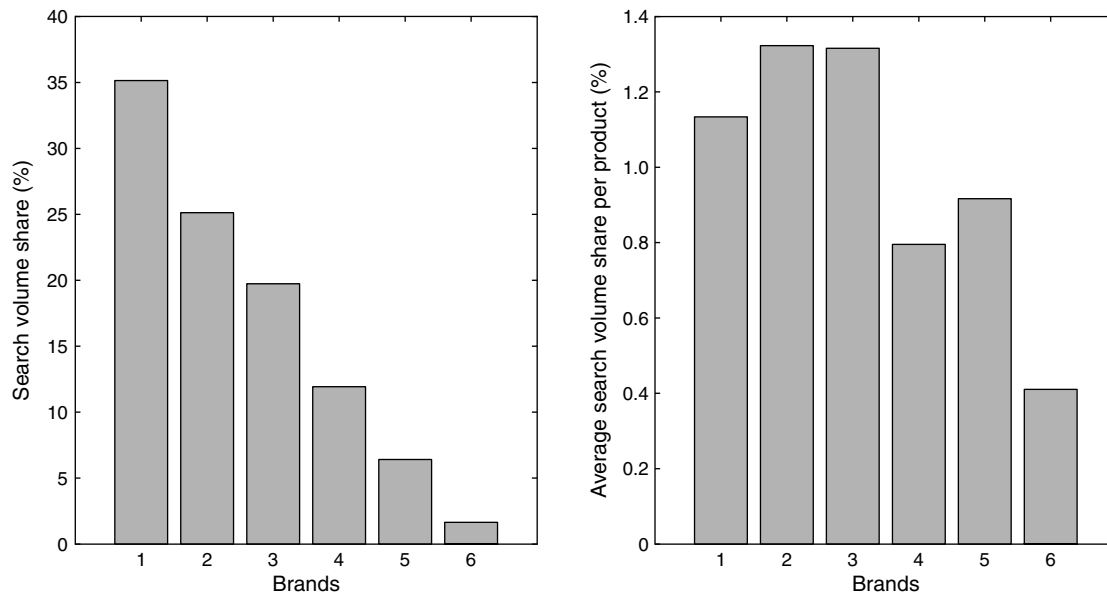
Note. The brands are 1 (Sony), 2 (Panasonic), 3 (Canon), 4 (JVC), 5 (Samsung), and 6 (Sanyo).

own-price elasticities of the products in the empirical analysis are −1.86 and −1.45, respectively.

We summarize the results using Figure 5, which visually shows the cross-price elasticities of demand for a given product (we use the product labeled "product 1," as an illustration). Recall that product 1 is a Sony camcorder with DVD media format, 40× optical zoom, 2.5-inch swivel screen, etc., selling at

$360. Its own elasticity is −2.12. From the top panel, we observe that only 9 out of 90 products have cross elasticities bigger than 0.1, 55 products have cross elasticities smaller than 0.02, and 24 products are computed to have cross-price elasticities substantively equal to 0. From the top panel in Figure 5, we conclude that products offered at similar price ranges are predicted to be, overall, more substitutable to

Figure 4 Joint Search Share Conditional on Search on Sony (Left) and Canon (Right)
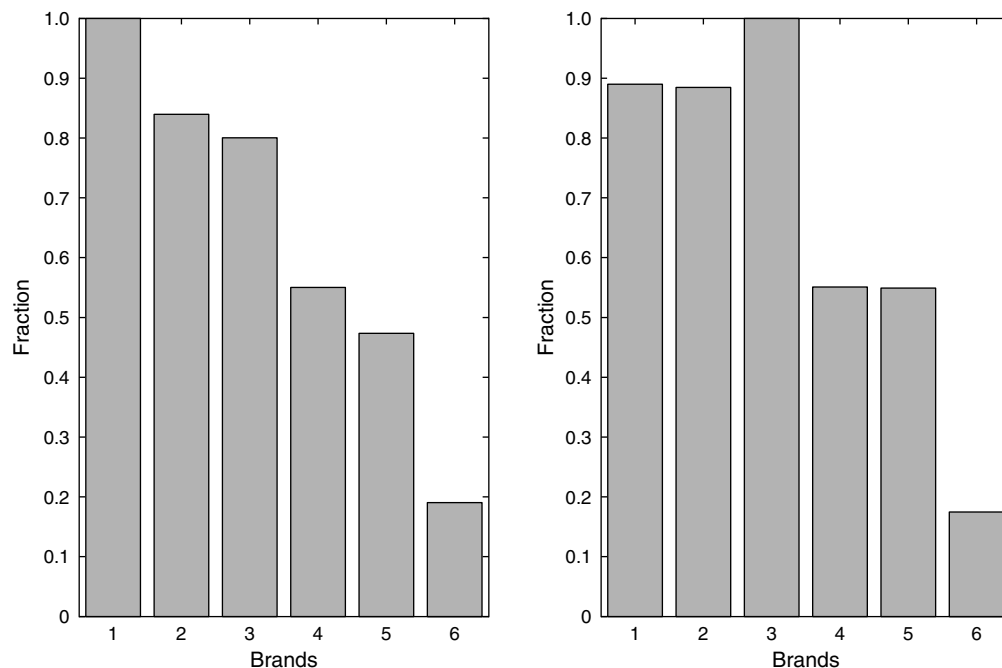
Note. The brands are 1 (Sony), 2 (Panasonic), 3 (Canon), 4 (JVC), 5 (Samsung), and 6 (Sanyo).

Figure 5   Cross Elasticities for Product 1, Sorted by Price (Top) and Media Formats (Bottom) of Competing Products
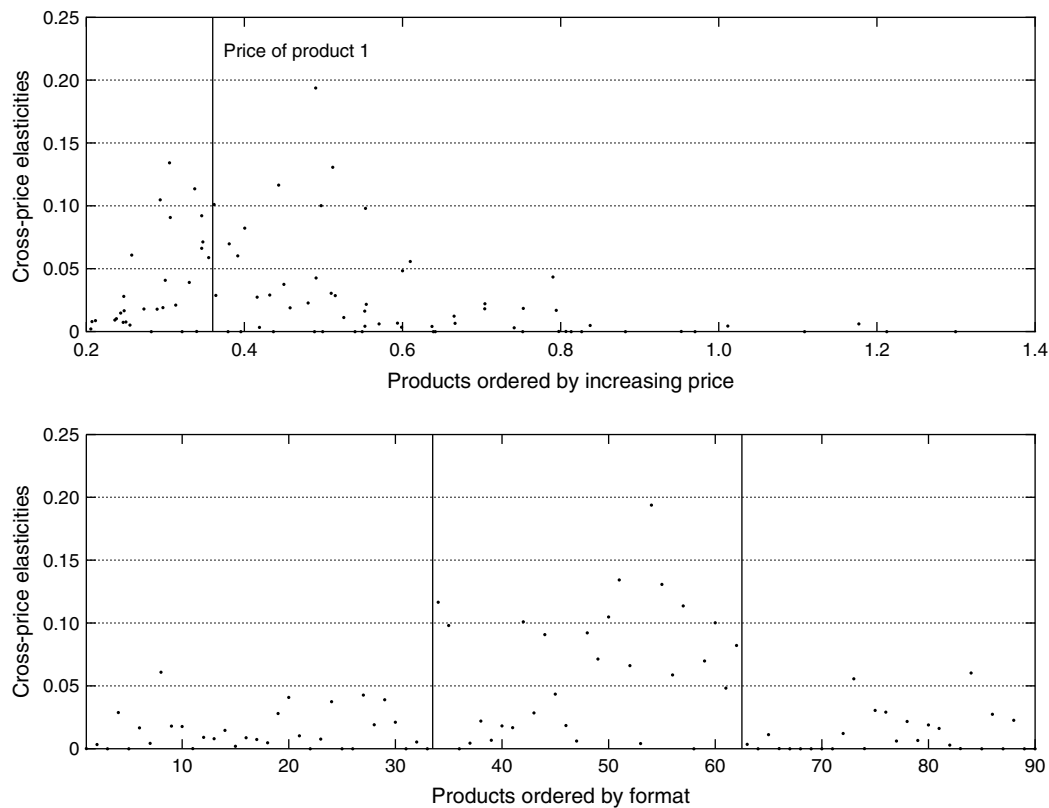
product 1 than the products offered at different price ranges. From the bottom panel, we observe that the same media format products are predicted to be more substitutable to product 1. These two graphs suggest that only a small number of products effectively compete with product 1. We note that this is true for the other products as well.

Table 7 shows, in detail, the closest and the most distant substitutes of product 1, as measured by cross-price elasticities. As previously mentioned, we see that the most elastic top five products share the same media format (DVD) and are in a similar price range. Note that among 14 Sony products based on DVD, product 1 is offered at the lowest price, and most closely substitutable to product 1 are other similarly priced Sony products based on the same media format. Also, other similarly priced, but less expensive, options from different manufacturers are strong substitutes. At the bottom of Table 7, we show that the four least cross-elastic products do not share price levels or media formats with product 1. They are mostly different brands, are based on different media formats, or are in different price brackets.

For completeness, we note that our model allows us to also compute an elasticity measure that captures the impact of price on product search volume. For instance, a natural measure of the impact of price on search would be the elasticity of search volume with

respect to price as

$$\epsilon_{j,k}^{\text{search}} = \frac{\%\ \text{search frequency change in } k}{\%\ \text{price change for } j}. \quad (27)$$

Using this measure, we find that product 1 has an own-search elasticity of -0.58. This means that if we increase the price of product 1 by 1%, its own-search frequency will drop by 0.58%. The price increase of

Table 7   Own- and Cross-Price Elasticities of Demand for Product 1

Product ID	Demand elasticity	Brand	Media format	Zoom		Price ($)
1	−2.116	Sony	DVD	40	⋯	360
73	0.194	Sony	DVD	12	⋯	490
58	0.134	Samsung	DVD	34	⋯	305
75	0.131	Sony	DVD	12	⋯	512
17	0.117	Sony	DVD	25	⋯	433
82	0.113	Panasonic	DVD	30		337
55	0.105	Samsung	DVD	32		294
42	0.101	Canon	DVD	10		362
⋮	⋮	⋮	⋮	⋮	⋮	⋮
35	0.000	JVC	HD	10	⋯	1,404
23	0.000	Panasonic	MiniDV	12	⋯	797
68	0.000	Samsung	FM	10	⋯	340
6	0.000	Sony	HD	10	⋯	807

Note. Product 1 is a Sony camcorder with DVD media format, 40x optical zoom, 2.5-inch swivel screen, etc., selling at $360.

product 1 results in a lower z_{ij}, providing an opportunity for its close competitors to enter the search set at its expense. As is the case with demand, we observe that search for some products is not meaningfully affected by price changes of product 1. About half of all products have cross-search elasticities smaller than 0.01, whereas only the top five products have cross-search elasticities greater than 0.05. This makes sense in our modeling framework because 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.

Using our model and estimates, we find 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. Our approach predicts intuitive substitution patterns even though we estimate the demand parameters using consumer search data only. As was pointed out earlier, this is possible because we believe the view-rank data contain rich information about the consumer substitution patterns, potentially more so than typical demand data. Given that the search data are becoming more available to marketers (e.g., some publicly available data at Amazon.com and Walmart.com), our proposed empirical model has the benefit that we can investigate the substitution patterns in many other product categories, which was heretofore often challenging as a result of a lack of good data. Second, the model is flexible yet parsimonious. Finally, the random coefficient discrete-choice model is a special case of the proposed model with zero search cost. With nonzero search cost in this model, the proposed model is capable of representing a more realistic substitution patterns by effectively focusing on the products that are in direct competition.

8. Counterfactual Simulations

8.1. The Effect of Reduced Search Costs on Consumer Surplus

Providing easy access by means of product references selectively lowers search costs for some products, but not for all. The net effect of such easy product access is a priori not clear; on one hand, lowering search cost may increase consumer surplus if it facilitates the finding of preferred products at a lower cost. 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 how the frequency of the product reference or appearance affects the consumer surplus. We do so by evaluating the effects of Amazon.com's product references

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 references. 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:

$$NS(S_i) = \max_{j \in S_i}\{u_{ij}\} - \sum_{j \in S_i} c_{ij}. \tag{28}$$

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

$$\Delta_{NS} = \sum_{i=1}^{I} \Delta_{NS,i} \tag{29}$$

$$= \sum_{i=1}^{I} NS(S_i^* \mid L = \{L_j\}) - NS(S_i^* \mid L = \varnothing), \tag{30}$$

where $(S_i^* \mid L)$ is i's optimal search set given L. The first term computes the net surplus across consumers under the presence of Amazon.com's product references, and the second term computes the net benefits across consumers in a hypothetical case where Amazon.com does not provide such product references.

From this simulation study, we find that virtually all consumers benefit from the reduced search cost through the product references offered at Amazon.com. The vast majority of consumers (99%) who benefit choose the same quality products, but their total search costs are lower. Although the consumers generally search more products in the presence of Amazon's product references,[31] their total search costs are lower because per-product search cost is much lower under the presence of the product appearances. Among the consumers who benefit from these features, their total net surplus increases by less than 1%. However, we find that their total search costs decrease by 72% on average.

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 (Nelson 1970, Anderson and Renault 1999). One such example is when popular products are being recommended, thereby

---

[31] Without product appearances, the median and mean search set sizes are 7 and 9, respectively. With product appearances, they are 11 and 14, respectively. This makes sense because lower search cost encourages consumers to conduct more search.

Figure 6   **The Impact of Limited Product Search on Market Shares**
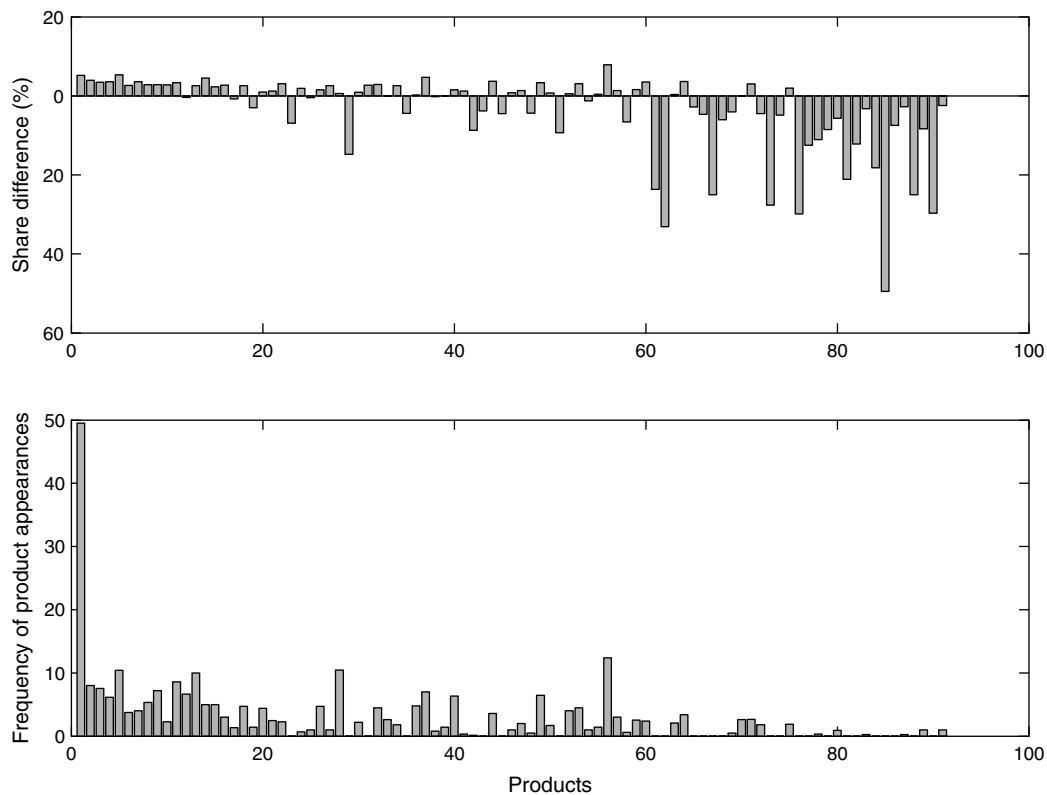
Notes. The top panel shows the forecast of market share under limited search less the share under full search. The bottom panel shows the frequency of appearances for each product.

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.[32] Given the ever-growing presence of variant forms of recommendations at many online stores, this counterfactual exercise helps manufacturers understand the direct impact of search costs on their product's performance.

The top panel in Figure 6 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 that 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 store-level frequency of appearances 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 appear less at the online store. Indeed, references to competing products form a double jeopardy to lower-value products—namely, not only do these products suffer from low preferences, they additionally suffer from high search cost (in the sense of having no appearances or a smaller number of store-level appearances).

We find that the online market for camcorders is more concentrated with than without a mechanism that selectively lowers the search costs. Figure 6 illustrates, in general, shares of popular brands increase with positive search cost and decrease with zero search cost. On the other hand, the demand for many lower-selling camcorders would increase as much as 30% if search cost were absent. In this sense, nonzero search cost tends to concentrate the online market for camcorders into demand for popular items. We note that this "polarization" effect of recommendation

---

[32] 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.

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 the 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 references or appearances.

Our modeling framework has a number of important benefits. First, it constitutes an internally consistent theory of information search and demand, 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." At the market level, the model allows for the occurrence of all possible search sets, 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, clicktrail data, viewing frequencies, coincidence of pairs of brands being viewed in the same session, or observed consideration sets. Last, our empirical model is applicable to other search contexts than shopping for durable goods, such as the analysis of clickstream data.

We used data experiments to show that the model is identified and that the estimation strategy works well. These data experiments suggest that it is possible to estimate heterogeneous demand systems from product-viewing data, which are freely available for many consumer durable products.

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 also consistently find that the cross-price elasticity for the large majority of product pairs is effectively zero. We further find that almost all consumers are better off with tools that selectively reduce search costs.

Finally, we also find that these tools tend to concentrate demand into popular products.

We see three areas for further development of our model. First, the current model uses only the view-rank data. Amazon.com also provides sales-rank data.[33] Recently, Bajari et al. (2008) show how these data can be used in demand estimation. We hope that the combination of both types of 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 of products (see Equation (23)) but should not impact the sales rank (see also Bajari et al. 2008). Thus, whereas view-rank data allow for the identification of the random-effects choice-based demand system, the combination of view-rank 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.[34] 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. We also see an application of our model in research on the effects of advertising, e.g., in lowering search cost and affecting 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.

Appendix. Derivation of the Reservation Utilities

The reservation utilities z_{ij} can be expressed by rewriting the implicit Equation (6) into

$$c_{ij} = \int_{z_{ij}}^{\infty} (u_{ij} - z_{ij}) f(u_{ij}) \, du_{ij}$$

$$= (1 - F(z_{ij}))\left[\int_{z_{ij}}^{\infty} (u_{ij} - z_{ij}) \frac{f(u_{ij})}{1 - F(z_{ij})} \, du_{ij}\right], \quad (31)$$

with $F(z_{ij})$ equal to the cumulative probability distribution of u_{ij} evaluated at z_{ij}. The term in parentheses 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

[33] Note that we currently use the sales-rank data for validation purposes.

[34] Note that we currently test the robustness of our empirical model under different assumptions of partial consumer knowledge.

that u_{ij} larger than z_{ij}. Using the assumption of normality $u_{ij} \sim \mathrm{N}(V_{ij}, \sigma_{ij}^2)$ and substituting the expectation of a truncated normal distribution (e.g., Johnson and Kotz 1970) in Equation (31), we obtain

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

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

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

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

$$x = (1 - \Phi(\zeta))(\lambda(\zeta) - \zeta). \tag{32}$$

This equation is identical to Equation (24) in the main text. Note that if we know ζ, we can compute $z = V + \zeta \times \sigma$ from the definition of $\zeta = (z - V)/\sigma$.

There are a number of properties of this equation that deserve further mentioning. First, and most important, we propose to solve ζ out of Equation (32), 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 (32) with respect to ζ, we obtain that

$$\frac{\partial x}{\partial \zeta} = -(1 - \Phi(\zeta)). \tag{33}$$

Thus the derivative of x with respect to ζ is negative,[35] and x is a decreasing function of ζ. This implies also that ζ is a decreasing function of x. From monotonicity, solutions to Equation (32) yield unique pairs of x and ζ.[36]

Third, from the results of 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

---

[35] Using

$$x = (1 - \Phi(\zeta))(\lambda(\zeta) - \zeta)$$

and differentiating with respect to ζ, we get

$$x' = -\Phi'(\zeta)(\lambda(\zeta) - \zeta) + (1 - \Phi(\zeta))(\lambda(\zeta)(\lambda(\zeta) - \zeta) - 1).$$

The second term on the right-hand side is

$$\Phi'(\zeta)(\lambda(\zeta) - \zeta) - (1 - \Phi(\zeta)),$$

which follows from the definition of the hazard rate; i.e.,

$$\lambda(\zeta) = \frac{\Phi'(\zeta)}{1 - \Phi(\zeta)}.$$

[36] Also see the appendix of Weitzman (1979, p. 651) for a more general way to prove the existence and uniqueness of the reservation utility.

this equation is $x = 0$ and $\zeta = \infty$. Indeed, at zero cost, the attractiveness to search an item is determined by the maximum upside of product utility, which is $+\infty$.

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 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 $x_{ij} = c_{ij}/\sigma_{ij}$. Given x_{ij}, we can look up $\zeta(x_{ij})$ that solves Equation (32) and compute z_{ij} from the definition of z_{ij}^* and the current values for V_{ij} and σ_{ij}; i.e.,

$$z_{ij} = V_{ij} + \zeta\left(\frac{c_{ij}}{\sigma_{ij}} \right) \times \sigma_{ij}. \tag{34}$$

With reference to Equation (23) 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 c_{ij}/σ_{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. Because this table is constructed outside the estimation, the marginal impact of extra precision in computing the zs on estimation time is zero.

References

Anderson, S. P., R. Renault. 1999. Pricing, product diversity, and search costs: A Bertrand-Chamberlin-Diamond model. *RAND J. Econom.* **30**(4) 719–735.

Bajari, P., J. T. Fox, S. P. Ryan. 2007. Linear regression estimation of discrete choice models with nonparametric distributions of random coefficients. *Amer. Econom. Rev.* **97**(2) 459–463.

Bajari, P., J. T. Fox, S. P. Ryan. 2008. Evaluating wireless carrier consolidation using semiparametric demand estimation. *Quant. Marketing Econom.* **6**(4) 299–338.

Barrow, D. F., A. C. Cohen Jr. 1954. On some functions involving Mill's ratio. *Ann. Math. Statist.* **25**(2) 405–408.

Bertolucci, J. 2007. Best places to buy tech products—Now and post-holidays. *PC World* (December 24), http://www.pcworld.com/article/140687/.

Boatwright, P., R. McCulloch, P. L. Rossi. 1999. Account-level modeling for trade promotion: An application of a constrained parameter hierarchical model. *J. Amer. Statist. Assoc.* **94**(448) 1063–1073.

Bresnahan, T. F. 1987. Competition and collusion in the American automobile industry: The 1955 price war. *J. Indust. Econom.* **35**(4) 457–482.

Bruno, H. A., N. J. Vilcassim. 2008. Structural demand estimation with varying product availability. *Marketing Sci.* **27**(6) 1126–1131.

Chiang, J., S. Chib, C. Narasimhan. 1999. Markov chain Monte Carlo and models of consideration set and parameter heterogeneity. *J. Econometrics* **89**(1–2) 223–248.

ComScore. 2007. Retail e-commerce climbs 23 percent in Q2 versus year ago. Press release (July 30), http://www.comscore.com/Press_Releases/2007/07/Retail_E-Commerce_Q2_2007.

Diamond, P. A. 1971. A model of price adjustment. *J. Econom. Theory* **3**(2) 158–168.

Efron, B., R. J. Tibshirani. 1993. *An Introduction to Bootstrap*. Chapman & Hall, New York.

Goeree, M. S. 2008. Limited information and advertising in the U.S. personal computer industry. *Econometrica* **76**(5) 1017–1074.

Gowrisankaran, G., M. Rysman. 2009. Dynamics of consumer demand for new durable goods. Working paper, Washington University in St. Louis, St. Louis.

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

Hong, H., M. Shum. 2006. Using price distributions to estimate search cost. *RAND J. Econom.* **37**(2) 257–275.

Horsky, D., P. Nelson. 2006. Testing the statistical significance of linear programming estimators. *Management Sci.* **52**(1) 128–135.

Hortaçsu, A., C. Syverson. 2004. Product differentiation, search costs, and competition in the mutual fund industry: A case study of S&P 500 index funds. *Quart. J. Econom.* **119**(2) 403–456.

Howard, J. A., J. N. Sheth. 1969. *The Theory of Buyer Behavior*. John Wiley & Sons, New York.

Huang, J.-H., Y.-F. Chen. 2006. Herding in online product choice. *Psych. Marketing* **23**(5) 413–428.

Johnson, N. L., S. Kotz. 1970. *Distributions in Statistics: Continuous Univariate Distributions*, Vol. 1. John Wiley & Sons, New York.

Linden, G. D., B. R. Smith, N. K. Zada. 2001. Use of product viewing histories of users to identify related products. U.S. Patent 6,912,505, filed March 29, 2001, issued June 28, 2005.

McCall, J. J. 1965. The economics of information and optimal stopping rules. *J. Bus.* **38**(3) 300–317.

Mehta, N., S. Rajiv, K. Srinivasan. 2003. Price uncertainty and consumer search: A structural model of consideration set formation. *Marketing Sci.* **22**(1) 58–84.

Meijer, E., J. Rouwendal. 2006. Measuring welfare effects in models with random coefficients. *J. Appl. Econometrics* **21**(2) 227–244.

Moe, W. W. 2003. Buying, searching, or browsing: Differentiating between online shoppers using in-store navigational clickstream. *J. Consumer Psych.* **13**(1–2) 29–39.

Montgomery, A. L., S. Li, K. Srinivasan, J. C. Liechty. 2004. Modeling online browsing and path analysis using clickstream data. *Marketing Sci.* **23**(4) 579–595.

Moorthy, S., B. T. Ratchford, D. Talukdar. 1997. Consumer information search revisited: Theory and empirical analysis. *J. Consumer Res.* **23**(4) 263–277.

Moraga-Gonzáles, J. L. 2006. Estimation of search cost. Working paper, University of Groningen, Groningen, The Netherlands.

Morgan, P., R. Manning. 1985. Optimal search. *Econometrica* **53**(4) 923–944.

Nelson, P. 1970. Information and consumer behavior. *J. Political Econom.* **78**(2) 311–329.

Nelson, P. 1974. Advertising as information. *J. Political Econom.* **82**(4) 729–754.

Petrin, A. 2002. Quantifying the benefits of new products: The case of the minivan. *J. Political Econom.* **110**(4) 705–729.

Price, K. V., R. M. Storn, J. A. Lampinen. 2005. *Differential Evolution: A Practical Approach to Global Optimization*. Springer, Berlin.

Reinganum, J. F. 1982. Strategic search theory. *Internat. Econom. Rev.* **23**(1) 1–15.

Reinganum, J. F. 1983. Nash equilibrium search for the best alternative. *J. Econom. Theory* **30**(1) 139–152.

Roberts, J. H., J. M. Lattin. 1991. Development and testing of a model of consideration set composition. *J. Marketing Res.* **28**(4) 429–440.

Rush, L. 2004. E-commerce growth will impact SMBs. *Internet.com* (January 23), http://www.internetnews.com/stats/print.php/3303241.

Senecal, S., J. Nantel. 2004. The influence of online product recommendations on consumers' online choices. *J. Retailing* **80**(2) 159–169.

Sonnier, G., A. Ainslie, T. Otter. 2007. Heterogeneity distributions of willingness-to-pay in choice models. *Quant. Marketing Econom.* **5**(3) 313–331.

Stigler, G. J. 1961. The economics of information. *J. Political Econom.* **69**(3) 213–225.

Weitzman, M. L. 1979. Optimal search for the best alternative. *Econometrica* **47**(3) 641–654.