Consumer Search and Prices in the Automobile Market*

José Luis Moraga-González†

Zsolt Sándor‡

Matthijs R. Wildenbeest§

First version: December 2009

Current version: February 2011

PRELIMINARY AND INCOMPLETE, COMMENTS WELCOME

Abstract

In many markets consumers only have imprecise information about the alternatives available. Before deciding which alternative to purchase, if any, consumers search to find their preferred products. This paper develops a discrete-choice model with optimal consumer search. Consumer choice sets are endogenous and therefore imperfect substitutability across brands does not only arise from variation in product characteristics but also from variation in the costs of searching alternative brands. We apply the model to the automobile industry using aggregate data on prices, market shares, as well as data on dealership locations and consumer demographics. Our estimate of search cost is highly significant and indicates that consumers conduct a limited amount of search. The paper shows that accounting for search cost and its effect on generating heterogeneity in choice sets is important in explaining variability in purchase patterns.

Keywords: consumer search, choice sets, differentiated products, demand and supply, discrete choice, automobiles

JEL Classification: C14, D83, L13

---

*We thank Sergei Koulayev and Marc Santugini for their useful comments and suggestions. This paper has also benefited from presentations at the University of Illinois at Urbana-Champaign, 2009 Workshop on Search and Switching Costs at the University of Groningen, 2010 IIOC meeting in Vancouver, and 2010 Marketing Science Conference in Cologne. Financial support from Marie Curie Excellence Grant MEXT-CT-2006-042471 is gratefully acknowledged.

†IESE Business School and University of Groningen, E-mail: jlmoraga@iese.edu.

‡University of Groningen, E-mail: z.sandor@rug.nl.

§Indiana University, Kelley School of Business, E-mail: mwildenb@indiana.edu.

1 Introduction

In markets like those for automobiles, electronics, computers, and real estate, finding out an acceptable alternative is time-consuming. Conducting a purchase involves first gathering preliminary information about the various alternatives available in the market. After this, most consumers collect further information about the most promising alternatives and then decide whether or not to buy the most preferred product. Preliminary information is usually easy to obtain either from television, the Internet, newspapers, specialized magazines, or just from neighbors, family, and friends. However, since some product characteristics are difficult to quantify, print, or advertise consumers then proceed by conducting a more or less exhaustive in-store search to find out which product fits them best. In practice, since visiting shops involves significant search costs, it is known most consumers engage in a rather limited amount of search.

Earlier work on estimation of demand and supply by means of random coefficients logit models (Berry, Levinsohn, and Pakes (BLP), 1995; Nevo, 2001) has proceeded by assuming that consumers have perfect information on all the products available in the market. Since information can either be gathered by consumers and/or be advertised by the firms, there are at least two natural ways to interpret the full information assumption of BLP. The first is by assuming that search costs are negligible for all consumers. The second interpretation is that firms' advertisements reach all consumers and convey all relevant information. In many settings the full information assumption seems unrealistic.

In a recent paper Sovinsky Goeree (2008) shows that to obtain realistic estimates of demand and supply parameters it is important to allow for both heterogeneous and limited consumer information. Yet, the settings of BLP and Sovinsky Goeree (2008) have in common that consumers do not need to incur any search costs to evaluate the utility they derive from the various alternatives. That is, in Sovinsky Goeree (2008) not all products are considered by an individual consumer but still consumers have all the information regarding the products they contemplate buying.

Our paper adds to this literature by presenting a discrete choice model with optimal consumer search. In our model consumers first decide which products to inspect. After having incurred search costs to find out all the relevant details they choose which of the inspected products to acquire, if any. Search costs vary across individuals and firms so consumers search distinct subsets of products even if they have similar preferences like in the simple logit model. Similar to the effects of advertising in Sovinsky Goeree (2008), search frictions also generate heterogeneous and limited consumer information.

We apply the model to the automobile market. The automobile market is precisely one where advertisements, reports in specialized magazines, and television programs convey some but not all of the relevant information. As a result, almost no consumer buys a new car without visiting a dealership. In our model consumers first decide which dealers to visit. For this decision, they use preliminary information they freely obtain on car characteristics (design, size, horsepower, fuel efficiency, etc.), equilibrium prices, dealer locations, and search costs. A visit to the dealers is meant to view, inspect and possibly test-drive the car. As a result, buyers incur significant search costs before conducting a purchase. In our model consumer choice sets are endogenous and therefore imperfect substitutability across brands does not only arise from product differentiation but also from costly search.

We estimate the model using macro-level data on prices, market shares, as well as data on dealership locations and consumer demographics. By exploiting variation in distances from consumer households to car dealerships we can identify the magnitude of search costs. Our estimate of the search cost parameter is highly significant. Moreover, search costs turn out to be economically important in the sense that hardly any consumer conduct such an exhaustive search that they can be considered as fully informed.

One advantage of our model is that it yields the BLP setting in the limiting case where search costs go to zero. Our paper shows that relaxing the zero-search-cost assumption leads to more realistic own- and cross-price elasticities. Demand appears not to be too elastic and price-cost margins not to be too low. We conclude that accounting for search cost and its effect on generating heterogeneity in choice sets is important in explaining variability in purchase patterns.

Our paper builds on the theoretical and empirical literature on consumer search behavior. At least since the seminal article of Stigler (1961) on the economics of information a great deal of theoretical and empirical work has revolved around the idea that the existence of search costs has nontrivial effects on market equilibria. Part of the effort has gone into the study of homogeneous product markets (see for instance Burdett and Judd, 1983; Reinganum, 1979; Stahl, 1989). In this literature a fundamental issue has been the existence of price dispersion in market equilibrium. Wolinsky (1984) studies a model with product differentiation and notes that search costs generate market power even in settings with free-entry of firms. More recent contributions investigate how product diversity (Anderson and Renault, 1999) and product quality (Wolinsky, 2005) are affected by search costs.

Some recent empirical research on consumer search behavior has focused on developing techniques to estimate search costs using aggregate market data. Hong and Shum (2006) develop a

structural method to retrieve information on search costs for homogeneous products using only price data. Moraga-González and Wildenbeest (2008) extend the approach of Hong and Shum (2006) to the case of oligopoly and present a maximum likelihood estimator. Hortaçsu and Syverson (2004) study a search model where search frictions coexist with vertical product differentiation. Our model is in this tradition and contributes to the literature by presenting a more general model of demand and supply that allows for heterogeneity in demand parameters as well as in search costs. Koulayev (2010) and Kim, Albuquerque and Bronnenberg (2010) present related models of search and employ micro-level data on search behavior to estimate preferences as well as the costs of searching.

The rest of the paper is organized as follows. In the next section we discuss the economic model. In Section 3 we discuss the estimation procedure. Identification of the model is discussed in Section 4, while the data and estimation results are presented in Section 5. Section 7 concludes.

2 Economic Model

2.1 Utility and demand

We consider a market where there are J different cars (indexed $j = 1, 2, \ldots, J$) sold by F different firms (indexed $f = 1, 2, \ldots, F$). We shall denote the set of cars by $\mathcal{J}$ and the set of firms by $\mathcal{F}$. The utility consumer i derives from car j is given by:

$$u_{ij} = \alpha_i p_j + x_j' \beta_i + \rho d_{if} + \xi_j + \varepsilon_{ij}, \tag{1}$$

where the variable p_j denotes the price of car j and the vector $(x_j, d_{if}, \xi_j, \varepsilon_{ij})$ describes different product attributes from which the consumer derives utility. We assume x_j and ξ_j are product attributes the consumer observes without searching, like horsepower, weight, transmission type, ABS, air-conditioning, number of gears, etc. The variable d_{if} denotes the distance from consumer i's home to firm f selling product j. Information on car characteristics and dealership locations can easily be retrieved from for instance advertisements, the Internet, specialized magazines, and consumer reports. The variable ε_{ij}, which is assumed to be independently and identically type I extreme value distributed across consumers and products, is a match parameter and measures the "fit" between consumer i and product j. We assume that ε_{ij} captures "search-like" product attributes, that is, characteristics that can only be ascertained upon visiting the dealership, inspecting, and possibly test-driving the car, like comfortability, spaciousness, engine noisiness, and

gearbox smoothness. It is assumed that the econometrician also observes the product attributes contained in x_j but cannot observe those in ξ_j and ε_{ij}. Consumers differ in the way they value price and product characteristics. The parameters α_i and β_i, which are assumed to be normally distributed with means α and β and variances σ_α^2 and Σ_β, capture consumer heterogeneity in tastes for price and product attributes. The utility from not buying any of the cars is

$$u_{i0} = \varepsilon_{i0}.$$

Therefore, we regard $j = 0$ as the "outside" option; this includes the utility derived from a non-purchase, or the purchase of a used car.

We shall allow for multi-product firms: firm $f \in \mathcal{F}$ supplies a subset $\mathcal{G}_f \subset \mathcal{J}$ of all products. In the car industry dealers sell disjoint sets of cars, so $\mathcal{G}_f \cap \mathcal{G}_g = \varnothing$ for any $f, g \in \mathcal{F}$.

We assume consumers must search to find out the exact utility they derive from each of the cars available as well as the utility of the outside option. To be more specific, before searching we assume consumers know (i) the location of each car dealership and the subset of makes and models available at each dealership, (ii) car characteristics p_j, x_j and ξ_j for each car j, and (iii) the distribution of ε_{ij}. Therefore, we regard the process of search of a consumer i as a process by which she discovers the exact values of the matching parameter ε_{ij} upon visiting the dealership.[1]

Consumers are assumed to use a non-sequential search strategy, i.e., they choose which subset of dealers to visit in order to maximize their expected utility; once they have visited the chosen dealers and have learned all the attributes of the cars they are interested in, they decide whether to buy any of the inspected cars or else opt for the outside option. Consumers differ in the search costs they incur if they visit a subset of dealers. Let $\mathcal{S}$ be the set of all subsets of dealers in $\mathcal{F}$ and let S be an element of $\mathcal{S}$. We shall denote consumer i's search cost for visiting all the dealerships in S by c_{iS}. An important part of the cost of visiting a set of dealers S is the distance from the consumer's home to the different car dealerships in S; let d_{if} be a vector of such explanatory variables. Then we assume

$$c_{iS} = \sum_{f \in S} c_{if} + \lambda_{iS} = \sum_{f \in S} d_{if}\gamma_i + \lambda_{iS},$$

where $c_{if} = d_{if}\gamma_i$ is the individual i's cost of visiting a firm f in S. The parameter γ_i is a consumer-specific coefficient. To capture heterogeneity in consumers' opportunity cost of time, we assume

[1] In general, one can distinguish between shop search and brand search. In our model consumers search among different brands. The difference with shop search is that the same brand may be sold by several shops, which is not the case in the car market.

it is normally distributed with mean γ and variance σ_γ^2, $\gamma_i \sim N\left(\gamma, \sigma_\gamma^2\right)$. Finally, we introduce a consumer-specific search cost shock λ_{iS} for visiting a set of dealers S; these shocks are assumed to be i.i.d. type I extreme value distributed across consumers and subsets of dealers. To simplify the formulae of the choice probabilities, it is convenient to assume consumers always include the outside good in their choice set. Of course, consumers are allowed to pick a choice set that only includes the outside good, i.e., $S = \varnothing$, for a cost $c_{i\varnothing} = \lambda_{i\varnothing}$.[2]

2.2 Optimal non-sequential search

A consumer i first decides which subset of dealers to visit; then, upon visiting the dealers and inspecting and test-driving the cars that are sold at those dealers, she makes a purchase decision. In order to decide which (subset of) dealers to visit, consumer i must compare the expected gains from searching all the possible subsets of dealers. The expected gains to consumer i from searching the dealerships in a subset S are

$$E\left[\max_{j \in \mathcal{G}_f \cup \{0\},\ f \in S} \{u_{ij}\}\right] - c_{iS},$$

where E denotes the expectation operator, taken in this case over the search characteristics ε_{ij}'s. We now define

$$m_{iS} = E\left[\max_{j \in \mathcal{G}_f \cup \{0\},\ f \in S} \{u_{ij}\}\right] - \sum_{f \in S} d_{if}\gamma_i$$

Letting F denote the CDF of ε_{ij}, the random variable $\max_{j \in \mathcal{G}_f \cup \{0\},\ f \in S} \{u_{ij}\}$ has CDF $\prod_{j \in \mathcal{G}_f \cup \{0\},\ f \in S} F(u - \delta_{ij})$, where δ_{ij} is the mean utility consumer i derives from alternative j, i.e., $\delta_{ij} = \alpha_i p_j + x_j'\beta_i + \xi_j$. Using this, we obtain

$$\begin{aligned} m_{iS} &= \log\left(1 + \sum_{f \in S} \exp[\delta_{if}]\right) - \sum_{f \in S} d_{if}\gamma_i, \\ m_{i\varnothing} &= 0, \end{aligned} \tag{2}$$

[2] An interpretation of this assumption is that if a consumer i does not search then she does not know ε_{i0}; if this consumer searches some firms then she gets to know ε_{i0} at no additional cost.

where $\delta_{if} = \log\left(\sum_{j\in\mathcal{G}_f} \exp[\delta_{ij}]\right)$.[3] Consumer i will pick the subset S_i that maximizes the expected gain $m_{iS} - \lambda_{iS}$, i.e.,

$$\begin{aligned} S_i &= \arg\max_{S\in\mathcal{S}} [m_{iS} - \lambda_{iS}] \\ &= \arg\max_{S\in\mathcal{S}} \left[\log\left(1 + \sum_{f\in S} \exp[\delta_{if}]\right) - \sum_{f\in S} d_{if}\gamma_i - \lambda_{iS} \right]. \end{aligned}$$

Since we assume $(-\lambda_{iS})$ is i.i.d. type I extreme value distributed, the probability that consumer i finds it optimal to sample the set of dealers S_i is P_{iS_i} where

$$P_{iS} = \frac{\exp[m_{iS}]}{\sum_{S'\in\mathcal{S}} \exp[m_{iS'}]}.$$

Given that consumer i searches the set S_i, the probability that consumer i buys j is equal to the probability that j is purchased out of the products of the firms in S_i is $P_{ij|S_i}$ where

$$P_{ij|S} = \frac{\exp[\delta_{ij}]}{1 + \sum_{r\in S} \exp[\delta_{ir}]}.$$

---

[3] Note that

$$\begin{aligned} E\left[\max_{j\in\mathcal{G}_f\cup\{0\},\ f\in S} \{u_{ij}\}\right] &= \int_{-\infty}^{\infty} u \frac{d}{du}\left(\prod_{j\in\mathcal{G}_f\cup\{0\},\ f\in S} F(u-\delta_{ij})\right) du; \\ &= \int u \frac{d}{du}\left(\prod_{j\in\mathcal{G}_f\cup\{0\},\ f\in S} \exp\left[-\exp\left[-(u-\delta_{ij})\right]\right]\right) du; \\ &= c + \log\left(1 + \sum_{j\in\mathcal{G}_f,\ f\in S} \exp[\delta_{ij}]\right). \end{aligned}$$

So

$$m_{iS} = c + \log\left(1 + \sum_{j\in\mathcal{G}_f,\ f\in S} \exp[\delta_{ij}]\right) - \sum_{f\in S} d_{if}\gamma_i.$$

Also,

$$m_{i\varnothing} = E\left[\max_{j\in\{0\}} \{u_{ij}\}\right] - \sum_{f\in\varnothing} d_{if}\gamma_i = E\left[\varepsilon_{i0}\right] = c.$$

In the expression in equation (2) we omit the Euler constant c because it does not affect choices.

In order to obtain the unconditional probability s_{ij} that consumer i purchases product j, we need to 'integrate out' S_i from this probability, i.e.,

$$\begin{aligned} s_{ij} &= \sum_{S \in \mathcal{S}_f} P_{iS} P_{ij|S} \\ &= \sum_{S \in \mathcal{S}_f} \frac{\exp[m_{iS}]}{\sum_{S' \in \mathcal{S}} \exp[m_{iS'}]} \frac{\exp[\delta_{ij}]}{1 + \sum_{r \in S} \exp[\delta_{ir}]}. \end{aligned}$$

where f is the firm producing j and $\mathcal{S}_f$ is the set of all choice sets containing firm f. Let $\tau_i := (\alpha_i, \beta_i, \gamma_i)$ be the vector of all random variables that need to be integrated out of s_{ij}. Then the probability that product j is purchased is the integral

$$s_j = \int s_{ij} f_\tau(\tau_i) d\tau_i. \tag{3}$$

Such an integral is difficult to compute analytically but it can be estimated by Monte Carlo simulations (see Section 3.3).

2.3 Supply side

We include the supply side in order to obtain estimates of price-cost markups. We assume each firm $f \in \{1, \ldots, F\}$ supplies a subset $\mathcal{G}_f$ of the J products. Let M denote the number of consumers and let mc_j denote the marginal cost of producing product j. Then the profit of firm f is given by

$$\Pi_f(p) = \sum_{j \in \mathcal{G}_f} (p_j - mc_j) M s_j(p).$$

Following BLP we assume mc_j depends log-linearly on observed product characteristics affecting cost, w_j, and an unobserved cost characteristic ω_j:

$$\ln(mc_j) = w_j' \eta + \omega_j. \tag{4}$$

We expect the unobserved cost characteristics ω_j to be correlated with the unobserved demand characteristics ξ_j. For instance, if the researcher does not observe whether a car has a navigation system as standard equipment, then cars having this characteristic will have a higher unobserved demand characteristics and, because it is more costly for the firm to include a navigation system, a higher unobserved cost characteristics as well. We will account for this correlation in the estimation procedure.

We assume firms maximize their profits by setting prices, taking into account prices and attributes of competing products. Assuming a Nash equilibrium exists for this game, any product sold should have prices that satisfy the first order conditions

$$s_j(p) + \sum_{r \in \mathcal{G}_f} (p_r - mc_r) \frac{\partial s_r(p)}{\partial p_j} = 0.$$

To obtain the price-cost markups for each product we can rewrite the first order conditions as

$$p - mc = \Delta(p)^{-1} s(p), \tag{5}$$

where the element of $\Delta(p)$ in row j column r is denoted by Δ_{jr} and

$$\Delta_{jr} = \begin{cases} -\dfrac{\partial s_r}{\partial p_j}, & \text{if } r \text{ and } j \text{ are produced by the same firm;} \\ 0, & \text{otherwise.} \end{cases}$$

For the derivation of the partial derivatives of the market shares with respect to price it matters whether deviation prices are assumed to be observable by consumers or not. In our search context, we adopt the standard assumption that consumers know the equilibrium prices but do not observe price deviations before searching (see e.g., Burdett and Judd, 1983; Wolinsky, 1986; Anderson and Renault, 1999).

3 Estimation Procedure

Our estimation procedure closely resembles BLP and Sovinsky Goeree (2008), except that we allow for an endogenous choice set selection stage which is the outcome of an optimal consumer search problem. As shown by BLP the parameters of the demand and supply model without search frictions can be estimated by generalized method of moments (GMM). Their GMM procedure accounts for price endogeneity by solving for the unobservables ξ_j and ω_j in terms of the observed variables and taking these as the econometric error term of the model. As in BLP, we can compute the vector $\xi = (\xi_1, ..., \xi_J)$ of unobserved characteristics as the unique fixed point of a contraction mapping. The fact that this mapping is indeed a contraction follows from the fact that the first order derivatives of the market shares with respect to the unobserved characteristics have the same form as in BLP. In this section we provide a method to estimate the model with search frictions by GMM as well.

3.1 Moments

Following BLP, model j's predicted market share $s_j(\theta)$ should match observed market shares $\hat{s}_j$, or

$$s_j(\delta(\theta), \theta) - \hat{s}_j = 0.$$

We use a contraction mapping suggested by BLP to solve for $\delta(\theta)$. The first moment unobservable follows from $\delta(\theta)$ and is

$$\xi_j = \delta_j(\theta) - x_j\beta.$$

The second moment unobservable follows from the parametric marginal cost specification and the first order conditions—combining equations (4) and (5) and solving for ω_j gives

$$\omega_j = \ln(p - \Delta^{-1}s(\theta)) - w_j'\eta.$$

3.2 GMM estimation

We use GMM to estimate the model. The estimation relies on the assumption that the true values of the demand and cost unobservables are mean independent of observed product characteristics, that is,

$$E[(\xi_j, \omega_j)|(X, W)] = 0.$$

Let Z be a matrix of instruments with $2J$ rows and let $\psi(\theta) = (\xi_1(\theta), ..., \xi_J(\theta), \omega_1(\theta), ..., \omega_J(\theta))'$. Let the sample moments be

$$g_J(\theta) = \frac{1}{J} Z'\psi(\theta).$$

The GMM estimator of θ is

$$\hat{\theta} = \arg\min_{\theta} g_J(\theta)' \Xi g_J(\theta),$$

where Ξ is a weighting matrix. Some parameters enter linearly in the model which means we can concentrate them out of the above GMM minimization. Let $\theta = (\theta_1', \theta_2')'$ and

$$\psi(\theta) = \delta(\theta_2) - X_1\theta_1.$$

By assuming θ_2 known we can obtain θ_1 as the linear IV estimator

$$\widehat{\theta_1} = \left(X_1'Z\Xi Z'X_1\right)^{-1} X_1'Z\Xi Z'\delta(\theta_2),$$

and substituting this in $g_J(\theta)$ we obtain a new sample moment, which is a function of θ_2 only

$$\overline{g}_J(\theta_2) = \frac{1}{J} Z' \overline{\psi}(\theta_2), \quad \text{where} \quad \overline{\psi}(\theta_2) = \delta(\theta_2) - X_1 \left(X_1' Z \Xi Z' X_1 \right)^{-1} X_1' Z \Xi Z' \delta(\theta_2).$$

The GMM estimator of θ based on this is

$$\hat{\theta}_2 = \arg\min_{\theta_2} \overline{g}_J(\theta_2)' \Xi \overline{g}_J(\theta_2).$$

3.3 Simulation of purchase probabilities

We use the empirical distribution of consumer demographics, including distances to dealers to proxy for search costs, which means the predicted market shares given by equation (3) do ni-257(ni33(on)-334(t0ition)10.9

Clearly, for $S \in \mathcal{S}_{-f}$

$$P_{i\{f\}\cup S} = \frac{\exp\left[\log\left(1+\exp[\delta_{if}]+\sum_{g\in S}\exp[\delta_{ig}]\right)-\left(d_{if}+\sum_{g\in S}d_{ig}\right)\gamma_i\right]}{\sum_{S'\in\mathcal{S}}\exp[m_{iS'}]}$$

and

$$P_{ij|\{f\}\cup S} = \frac{\exp(\delta_{ij})}{1+\exp(\delta_{if})+\sum_{g\in S}\exp(\delta_{ig})}.$$

Now, rewrite s_{ij} in the importance sampling form

$$s_{ij} = \sum_{S\in\mathcal{S}_{-f}} Q_{iS}\frac{P_{i\{f\}\cup S}}{Q_{iS}}P_{ij|\{f\}\cup S},$$

where $\sum_{S\in\mathcal{S}_{-f}} Q_{iS} = 1$ holds. A set S drawn randomly from $\mathcal{S}_{-f}$ can be represented as the vector of $[0,1]$ i.i.d. uniform random variables $u_{i,-f} = (u_{i1},\ldots,u_{if-1},u_{if+1},\ldots,u_{iF})$ because according to the importance sampling probabilities we can draw S by drawing $u_{i,-f}$ such that $g \in S$ iff $u_{ig} \le \phi_{ig}$ for all $g \in \mathcal{F}\backslash\{f\}$ (we omit the argument θ_0 from $\phi_{ig}(\theta_0)$). So we can use the argument $u_{i,-f}$ in the expressions involved in s_{ij}. We introduce

$$\begin{aligned}
Q_{if}(u_i) &= \prod_{g\in\mathcal{F}\backslash\{f\}} \phi_{ig}^{\mathbf{1}(u_{ig}\le\phi_{ig})}(1-\phi_{ig})^{\mathbf{1}(u_{ig}>\phi_{ig})} \\
P_{if}(u_i) &= \frac{\exp\left[\log\left(1+\exp[\delta_{if}]+\sum_{g\in\mathcal{F}\backslash\{f\}}\mathbf{1}\left(u_{ig}\le\phi_{ig}\right)\exp[\delta_{ig}]\right)-\left(d_{if}+\sum_{g\in\mathcal{F}\backslash\{f\}}\mathbf{1}(u_{ig}\le\phi_{ig})d_{ig}\right)\gamma_i\right]}{\sum_{S'\in\mathcal{S}}\exp[m_{iS'}]} \\
P_{ij|f}(u_i) &= \frac{\exp(\delta_{ij})}{1+\exp(\delta_{if})+\sum_{g\in\mathcal{F}\backslash\{f\}}\mathbf{1}\left(u_{ig}\le\phi_{ig}\right)\exp(\delta_{ig})},
\end{aligned}$$

where $\mathbf{1}(u_{ig} \le \phi_{ig})$ is the indicator of the event $(u_{ig} \le \phi_{ig})$. These correspond to Q_{iS}, $P_{i\{f\}\cup S}$ and $P_{ij|\{f\}\cup S}$, respectively. Then

$$s_{ij} = \int_{[0,1]^{F-1}} \frac{P_{if}(u_i)}{Q_{if}(u_i)}P_{ij|f}(u_i)du_{i,-f} = \int_{[0,1]^F} \frac{P_{if}(u_i)}{Q_{if}(u_i)}P_{ij|f}(u_i)du_i,$$

which yields

$$s_j = \int_{\mathbb{R}^\Delta}\int_{[0,1]^F} \frac{P_{if}(u_i)}{Q_{if}(u_i)}P_{ij|f}(u_i)f_\tau(\tau_i)du_i d\tau_i, \tag{6}$$

where Δ is the dimension of the random vector τ_i.

This latter formula is convenient because it shows how to estimate s_j by Monte Carlo. We can simply draw a random sample $(u_i, \tau_i)_{i=1}^{ns}$ jointly from their distribution and compute the Monte Carlo estimate of s_j as

$$\widetilde{s}_j = \frac{1}{ns} \sum_{i=1}^{ns} \left[\frac{P_{if}(u_i)}{Q_{if}(u_i)} P_{ij|f}(u_i) \right].$$

Note that although u_i is F-dimensional, we only use the $(F-1)$-dimensional $u_{i,-f}$ to compute $\widetilde{s}_j$, as equation (6) also suggests, so there is a kind of redundancy of draws. The draw u_{if} is used for computing $\widetilde{s}_r$ for products r belonging to firms rival to f.

Algorithm 1 (Importance Sampling) *The algorithm consists of the following steps:*

1. *For each $i = 1, \ldots, ns$ draw $u_i \sim U[0,1]^F$ and $\tau_i \sim f_\tau$;*
2. *For each $f \in F$ compute ϕ_{if} and $Q_{if}(u_i)$; this implicitly determines the choice set $S_{i0} \subset F \backslash \{f\}$ of i as*

   $$S_{i0} = \{g \in \mathcal{F} \backslash \{f\} : u_{ig} \leq \phi_{ig}\}$$

   (note that the choice set for computing s_j for $j \in G_f$ will be $\{f\} \cup S_{i0}$, so always contains f);
3. *For each f compute $P_{if}(u_i)$ assuming for the moment that $\sum_{S' \in \mathcal{S}} \exp[m_{iS'}]$ is known;*
4. *For each j compute $P_{ij|f}(u_i)$;*
5. *For each j compute $\widetilde{s}_j$.*

In order to specify $\phi_{if}(\theta)$, the first idea that comes to mind is to use the criterion that the two sets of probabilities are proportional at the singleton subsets of firms $\{f\},\ f = 1, \ldots, F$, i.e.,

$$\frac{Q_{i\{f\}}}{Q_{i\varnothing}} = \frac{P_{i\{f\}}}{P_{i\varnothing}},$$

which implies that

$$\frac{\phi_{if}}{1 - \phi_{if}} = \exp[m_{i\{f\}}],$$

so

$$\phi_{if} = \frac{\exp[m_{i\{f\}}]}{1 + \exp[m_{i\{f\}}]} = \frac{\exp[\log(1 + \exp[\delta_{if}]) - d_{if}\gamma_i]}{1 + \exp[\log(1 + \exp[\delta_{if}]) - d_{if}\gamma_i]}.$$

Simulation experiments based on these importance sampling probabilities show that they are not satisfactory.

We can exploit more information on the structure of m_{iS} and incorporate it in the ϕ's by using the criterion that the two sets of probabilities are proportional at subsets of firms $\{f, g_1, \ldots, g_H\}$ and $\{g_1, \ldots, g_H\}$ for $f = 1, \ldots, F$. More precisely,

$$\frac{Q_{i\{f,g_1,\ldots,g_H\}}}{Q_{i\{g_1,\ldots,g_H\}}} = \frac{P_{i\{f,g_1,\ldots,g_H\}}}{P_{i\{g_1,\ldots,g_H\}}},$$

which implies

$$\frac{\phi_{if}}{1-\phi_{if}} = \exp[m_{i\{f,g_1,\ldots,g_H\}} - m_{i\{g_1,\ldots,g_H\}}].$$

If we assume that this holds for all $\{f, g_1, \ldots, g_H\} \subset \mathcal{F}$ we get that

$$\frac{\phi_{if}}{1-\phi_{if}} = \exp[\overline{m}_{if,H}],$$

where $\overline{m}_{if,H}$ denotes the mean of $m_{i\{f,g_1,\ldots,g_H\}} - m_{i\{g_1,\ldots,g_H\}}$ over all $\{g_1, \ldots, g_H\} \subset \mathcal{F}\backslash\{f\}$ (this implicitly assumes that $g_h \neq g_k$ for all $h \neq k,\ h,\ k = 1, \ldots, H$). This yields

$$\phi_{if} = \frac{\exp[\overline{m}_{if,H}]}{1+\exp[\overline{m}_{if,H}]},\ f = 1, \ldots, F.$$

The number of terms involved in $\overline{m}_{if,H}$ is the number of subsets of $\mathcal{F}\backslash\{f\}$ having H elements, that is, combinations $\binom{F-1}{H}$.

Intuitively, the larger the subsets $\{g_1, \ldots, g_H\}$ involved, the more information on the structure of m_{iS} is captured by the ϕ's. This has also been found in simulation experiments for the cases $H = 1, 2$, and 3; for $H = 3$ the importance sampling probabilities constructed in the way described above work well in these experiments.The computational burden for computing the ϕ's for large F and H is high, but they only need to be computed for the starting value of the parameters.

We still need to find an estimator for $M_i = \sum_{S\in\mathcal{S}} \exp[m_{iS}]$, the denominator of $P_{if}(u_i)$. We can again use importance sampling based on the probabilities $\{Q_{iS}\}_{S\in\mathcal{S}}$ defined above. For this, write

$$M_i = \sum_{S\in\mathcal{S}} Q_{iS}\frac{\exp[m_{iS}]}{Q_{iS}} = \int\limits_{[0,1]^F} \frac{m_i(w)}{Q_i(w)}dw,$$

where

$$
\begin{aligned}
m_i(w) &= \exp\left[\log\left(1+\sum_{f\in\mathcal{F}}\mathbf{1}(w_f\le\phi_{if})\exp[\delta_{if}]\right)-\sum_{f\in\mathcal{F}}\mathbf{1}(w_f\le\phi_{if})d_{if}\gamma_i\right], \\
Q_i(w) &= \prod_{f\in\mathcal{F}}\phi_{if}^{\mathbf{1}(w_f\le\phi_{if})}(1-\phi_{if})^{\mathbf{1}(w_f>\phi_{if})} \quad \text{with} \quad w=(w_1,\ldots,w_F).
\end{aligned}
$$

The Monte Carlo estimator of M_i is

$$\widetilde{M_i}=\frac{1}{N}\sum_{n=1}^{N}\frac{m_i(w_n)}{Q_i(w_n)},$$

where $\{w_n\}_{n=1}^N$ is a set of i.i.d. draws from the uniform distribution on $[0,1]^F$.

Some simulation experiments show that this Monte Carlo importance sampling estimator based on quasi-random samples works well for $H=3$ and $Q_{iS}=Q_{iS}(\theta)$, that is, if we use the current parameter values. It can fail in some cases for $Q_{iS}=Q_{iS}(\theta_0)$ when the current parameter values θ are far from the starting values θ_0. Due to this it may be worthwhile to use an algorithm for finding the estimates that is adaptive so that it searches in a close neighborhood of the starting value, then changes the starting value to the optimum in this close neighborhood, and so on.

4 Identification

In this section we provide an informal discussion on identification. We start with identification of the parameters in the IV logit setting. In our model, variation in the sales across brands is due to (i) variation in the attributes of a car (ii) variation in the distances from the households to the closest dealers of the brands and (iii) variation in the optimal choice sets chosen by the households. We are interested in the identification of the parameters of the utility function which in the IV logit case are α, β and ρ; in addition we seek to identify the search cost parameter γ.

As usual, the β parameters of the utility function can be identified because the econometrician typically observes different market shares corresponding to different product characteristics. In our model, since we can control for the distances from the households to the dealerships, the parameters β can be identified in the same way. The set of instruments we use to control for possible correlation between unobserved characteristics and price are similar to BLP—in addition to product characteristics, which are exogenous by assumption, we add the number of cars and the sum of characteristics of the cars produced by the same firm, as well as the number of competing

cars and the some of characteristics of all competing cars.

As we have seen above, similar cars, like for example Saab 9-5 and Volvo V70, have quite distinct networks of dealerships. If consumers were fully informed, variation in sales across similar brands could only be due to differences in the distances to the closest dealers. This would identify the parameter ρ that weighs distance to the closest dealer of a brand in the utility function. Since we can control for the variation in choice sets across households we can identify this parameter in the same way.

Finally, the parameter γ, which measure search costs, can be identified if there is enough variation in the subsets of firms sampled by the different households. For this it is important that not all households visit just one dealer for otherwise the parameter ρ would be confounded with the parameter γ. When consumers happen to search more than one brand, then variation in market shares due to variation across choice sets allows for the identification of the γ parameters. Suppose we have two cars with similar attributes and similar distances to the closest dealers. Under full information, these two cars should have similar market shares. If this is not the case, then it is because there must be variation in the cost of search these two cars along with other cars. This enables us to identify the γ parameters. The idea is that after controlling for car characteristics and distances to the closest dealer, a small change in γ will induce variation in consumer choice sets that will ultimately be reflected in market shares.

5 Data and Results

5.1 Data

Our data set consists of prices, sales, physical characteristics, and locations of dealers of virtually all cars sold in the Netherlands between 2003 and 2008. We include a model in a given year if more than fifty cars have been sold during that year; this means "exotic" car brands like Rolls-Royce, Bentley, Ferrari, and Maserati are excluded. This leaves us with a total of 323 different models that were sold during this period—in any given year about 232 different models. We treat each model-year combination as one observation, which results in a total of 1391 observations.

The data on product characteristics are obtained from *Autoweek Carbase*, which is an online database of prices and specifications of all cars sold in the Netherlands from the early eighties until now.[4] Characteristics include horsepower, number of cylinders, maximum speed, fuel efficiency, weight, size, and dummy variables for whether the car's standard equipment includes

[4]See http://www.autoweek.nl/carbase.php.

air-conditioning, power steering, cruise control, ABS, and a board computer. Unfortunately transaction prices are not available, so all prices are listed (post-tax) prices. We have used the Consumer Price Index to normalize all prices to 2006 euros.

We have supplemented the data set with several macroeconomic variables including number of households and average gasoline prices, as reported by Statistics Netherlands.[5] The total number of households allows us to construct market shares, while average gasoline prices are used to construct our kilometers per euro (KM€) variable, which is calculated as kilometers per liter (KPL) divided by the price of gasoline per liter.

Like BLP we define a firm as all brands owned by the same company. We use information on the ownership structures from 2007 to determine which car brands are part of the same parent company—the 40 different brands in our sample are owned by 17 different companies. For instance, in 2007 Ford Motor Company owned Ford, Jaguar, Land Rover, Mazda, and Volvo.

Year	No. of Models	Sales	Price	American	Asian	European	HP/Wt	Size	Cruise Control	KPL	KP€
2003	214	481,969	19,523	0.026	0.212	0.762	0.784	7.153	0.227	14.512	12.524
2004	229	476,876	19,923	0.027	0.224	0.749	0.785	7.185	0.307	14.735	11.767
2005	234	458,433	20,459	0.029	0.245	0.726	0.791	7.270	0.297	14.891	11.009
2006	233	476,266	20,240	0.029	0.257	0.714	0.797	7.272	0.306	15.204	10.766
2007	239	495,555	20,379	0.028	0.261	0.711	0.803	7.330	0.279	15.171	10.396
2008	242	489,658	18,502	0.022	0.264	0.714	0.807	7.270	0.291	15.867	10.326
All	1391	479,793	19,831	0.027	0.244	0.729	0.794	7.247	0.284	15.068	11.126

Notes: Prices are in 2006 euros. All variables are sales weighted means, except for number of models and sales.

Table 1: Summary Statistics

Table 1 gives the sales weighted means for the main variables we use in our analysis. The number of models has increased from 214 in 2003 to 242 in 2008. Sales were lowest in 2005 and peaked in 2007. Prices have been going up mostly in real terms, although 2008 saw a sharp decrease, possibly as a result of the onset of the recession. The share of European cars sold shows an downward trend, mainly to the benefit of cars that originate from East Asia. The ratio of horsepower to weight has been increasing steadily. The share of cars with cruise contral as standard equipment increased in the first half of the sampling period, but then decreased somewhat again. Cars have become more fuel efficient during the sampling period. Nevertheless, as shown in the last column of Table 1 fuel efficiency has not increased enough to offset rising gasoline prices—the number of kilometers that can be traveled for one euro has decreased over the sampling period.

In addition to car characteristics we use data on the distribution of household characteristics to

[5]See http://www.cbs.nl.

	Mean	Std.dev.
Number of inhabitants	1,471	2,048
Household size	2.32	0.44
Single person households (%)	32.77	14.28
Households with children (%)	37.12	12.70
Households without children (%)	30.08	7.89
Number of cars per household	0.97	0.30
Disposable income per inhabitant	13,249	2,527

Notes: Except for number of inhabitants, mean is weighted by number of households.

Table 2: Descriptive statistics household characteristics

model the distribution of consumer taste parameters. We also combine information on the location of car dealerships and geographic data on where people reside to construct a matrix of distances between households and the different car dealerships. These distances are later used to proxy the cost of visiting a dealership to learn all product characteristics of a vehicle.

Our demographic and socioeconomic data on households are obtained from Statistics Netherlands. These data are available at various levels of regional disaggregation (neighborhoods, districts, city councils, counties, provinces, etc.). Since the purpose of our study is to estimate the importance of search costs, we choose to work at the highest level of regional disaggregation, that is, at the neighborhood level. This permits us to proxy the costs of traveling to the different car dealers rather accurately. Statistics Netherlands provides a considerable amount of useful demographic and socioeconomic data at this level of disaggregation.

For every neighborhood, the demographic data include the number of inhabitants and their distribution by age groups, the number of households, the average household size, the proportion of single-person households, and the proportion of households with children. The socioeconomic data include the average home value, the average income per inhabitant and income earner, as well as the total number of cars and their ownership status (company-leased versus privately-owned). We only include neighborhoods with a strictly positive number of inhabitants, which leaves us with a total of 11,122 neighborhoods for 2007.[6] Table 2 provides some summary descriptive statistics of several of the household characteristic variables we use in the specifications we discuss below.

In addition to demographic data we have information on the exact location of each neighborhood on the map of the Netherlands. Using a geographical software package we use this information to construct a proxy for the costs incurred when visiting a car dealership. To be able to do this,

[6]There are 284 neighborhoods for which the number of inhabitants is zero. These are neighborhoods that are typically located in industrial areas, ports, or remote agricultural areas. There are some neighborhoods for which we miss some of the relevant variables. To complete the data set we proceed by using information obtained at lower levels of disaggregation (districts or city councils).

for every brand we have first obtained the addresses of all its dealerships in the Netherlands. For instance, Saab has a total of twenty dealers in the Netherlands, which are spread over the country as shown in Figure 1(a). Since we have the exact addresses of the twenty dealerships of Saab, for every neighborhood, we can compute the Euclidean distance from the center of the neighborhood to the closest Saab dealer. We do this for all car manufacturers and obtain a matrix of 11,122 by 38 containing the minimum distances from the center of a neighborhood to a car dealer.

(a) Saab dealerships (b) Volvo dealerships

Figure 1: Locations selected dealers

There is a lot of variation in the distances to the closest dealer of each brand across neighborhoods. Figure 1(b) gives the spread of Volvo dealerships across the country—clearly on average the minimum distance to a Volvo dealer is much smaller than the minimum distance to a Saab dealer. A similar picture arises for other brands. Table 3 gives some descriptive summary statistics for the distances to the nearest dealer for all the car brands in our data. Opel is the most accessible: almost 79% of all households live within 5 kilometer from an Opel dealer. MG/Rover has the lowest percentage of households within 5 kilometer: only 3.5% of households is within easy reach.

5.2 Estimation results logit demand

Table 4 gives the parameter estimates for a logit demand model. As a benchmark case, in the first two columns we present the demand estimates for a model without search frictions. The OLS logit demand results are obtained by regressing $\ln(s_j) - \ln(s_0)$ on product characteristics and price using ordinary least squares. The results in the second column are obtained by using an instrumental variables (IV) approach to control for possible correlation between unobserved characteristics and price.[7] Except for horsepower per weight all parameters are of the expected sign and significant in both specifications—horsepower per weight has an unexpected negative impact in the OLS specification, while it has a positive coefficient and is statistically significant in the IV specification in the second column. The price coefficient is much larger in the IV specifications, which is to be expected given that higher unobserved quality components should lead to higher prices. The results indicate that cars produced by non-European firms yield negative marginal utility, which means cars produced by European firms (e.g., Peugeot/Citroën, Fiat, Volkswagen, etc.) have a higher mean consumer valuation than cars produced by non-European firms (e.g., Toyota, Honda, etc.). Size, a higher mileage per euro, and cruise control as standard equipment all affect the consumers' mean utility in a positive way.

In the last two columns of Table 4 we present the demand estimates using the model of optimal consumer search (models (iii) and (iv)). Since there is no closed form solution for the market share equations, we simulate the buying probabilities by randomly drawing 529 neighborhoods, where each neighborhood is weighted by the number of inhabitants. We use some household characteristics in the search cost specification, namely, 'distance,' 'income,' 'kids,' 'senior' (see Table 4). The variable 'distance' is the distance from the centroid of a neighborhood to the nearest dealers, 'income' is just the variable described in the previous subsection, 'kids' and 'senior' stand for the percentage of households with children and of age above 65 years, respectively, in the given neighborhood. The motivation for including distance in the search cost is clear since it is more costly to visit farther dealers. The motivation for including the other variables is the opportunity cost of time since households with high income or children value more their leisure time, so it is more costly for them to visit car dealers. Retired people are expected to have a lower valuation of leisure time.

Model (iii) in Table 4 contains only the distance variable in the search cost, while model (iv)

[7] As instruments we use own product characteristics, the number of other cars produced by the firm, the number of cars produced by rival firms, the sum of non-dummy product characteristics of other cars produced by the firm, as well as the sum of non-dummy product characteristics of cars produced by rival firms.

	No search		Search	
	OLS Logit Demand	IV Logit Demand	GMM/IV Logit Demand	GMM/IV Logit Demand
Variable	(i)	(ii)	(iii)	(iv)
Constant	-12.117	-15.954	-12.907	-8.245
	(0.640)	(1.067)	(1.019)	(1.057)
HP/Weight	-0.527	2.161	0.926	0.455
	(0.199)	(0.584)	(0.553)	(0.637)
Non-European	-0.546	-1.005	-0.740	-0.636
	(0.074)	(0.126)	(0.120)	(0.138)
Cruise control	0.117	0.289	0.113	0.059
	(0.087)	(0.106)	(0.098)	(0.104)
Fuel efficiency	0.238	0.213	0.222	0.248
	(0.024)	(0.028)	(0.024)	(0.029)
Size	0.291	0.790	0.490	0.419
	(0.057)	(0.119)	(0.117)	(0.127)
Price	-0.028	-0.100	-0.057	-0.042
	(0.003)	(0.015)	(0.015)	(0.018)
Search costs				
distance	—	—	0.072	0.080
			(0.015)	(0.026)
income	—	—	—	0.372
				(0.013)
kids	—	—	—	1.695
				(0.300)
senior	—	—	—	-3.674
				(0.486)
R^2	0.330	n.a.	n.a.	n.a.
Objective function	n.a.	n.a.	46.84	31.69

Notes: The number of observations is 1,382. Standard errors are in parenthesis. The number of simulated consumers is 529. Starting values are generated by using a grid search.

Table 4: Results with Logit Demand

contains all the four variables. A comparison of the parameter estimates with search to those without search shows that the price coefficient goes down in absolute value, which means that the cars become more inelastic. We can say roughly the same thing about the other utility parameters when we compare the last two columns. Further, in these two models the utility estimates have the expected signs and are statistically significant, with the exception of horsepower per weight and cruise control. The signs of the search cost parameter estimates are according to our expectations explained in the previous paragraph.

Figure 2 provides some more details on consumers' search behavior for 2008. We have estimated the probability mass function of the number of searches by the Metropolis-Hastings algorithm using 5000 simulated households. On the figure we can see the frequencies of households that search 1, 2, ..., 36 times. The percentage of households that do not search at all (not presented in the figure) is 87.76%. The percentages of households that search one dealer, 2–5 dealers and 6–36 dealers

are 5.48%, 4.58% and 2.18%, respectively. These results are according to our expectations. They reflect that search costs are relatively high, since only a very small fraction of consumers search more than 5 dealers.

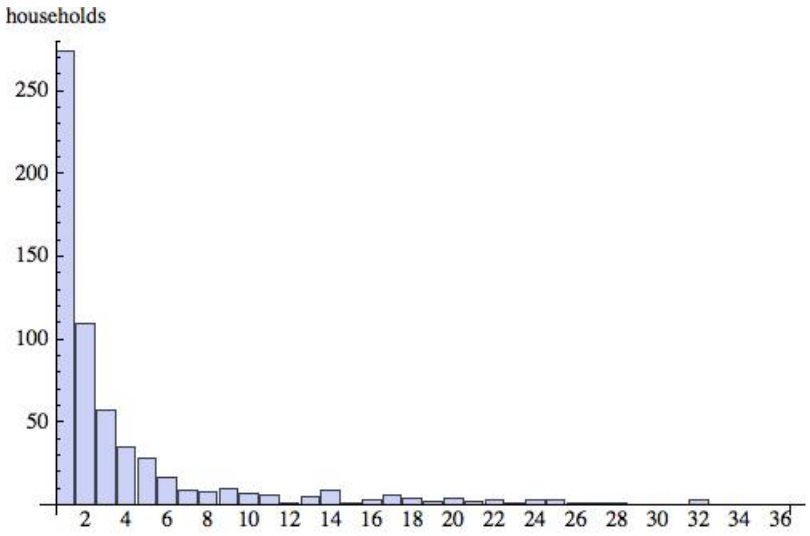

Figure 2: Distribution number of searches

6 Conclusions

In our analysis we have investigated how relaxing the assumption that consumers know all relevant product characteristics affects demand estimates in a discrete choice model of product differentiation. In our economic model consumers are initially unaware of whether a specific product is a good match—in order to find out consumers have to search non-sequentially among firms for the product that provides them with the highest utility. The model consists of an initial choice set selection stage, where consumers optimally determine the choice that gives them the highest expected utility taking into account cost of searching each choice set, and a buying stage where consumers pick the good with the highest realized utility, after the matching parameter of all products in their choice sets is revealed.

We have provided a way to estimate the model and have applied the model to the Dutch market for automobiles. We use distances from consumers to the nearest dealer of a specific brands well as household characteristics reflecting the opportunity cost of time to specify search cost. Our estimation results indicate that search costs are both significant and economically meaningful. According to our estimates consumers conduct a rather limited amount of search before buying.

References

[1] Anderson, Simon P. and Régis Renault: "Pricing, Product Diversity, and Search Costs: a Bertrand-Chamberlin-Diamond Model," *RAND Journal of Economics* 30, 719-35, 1999.

[2] Berry, Steven, James Levinsohn, and Ariel Pakes: "Automobile Prices in Market Equilibrium," *Econometrica* 63, 841-90, 1995.

[3] Burdett, Kenneth and Kenneth L. Judd: "Equilibrium Price Dispersion," *Econometrica* 51, 955-69, 1983.

[4] Hong, Han and Matthew Shum: "Using Price Distributions to Estimate Search Costs," *RAND Journal of Economics* 37, 257-75, 2006.

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

[6] Kim, Jun, Paulo Albuquerque, and Bart J. Bronnenberg: "Online Demand Under Limited Consumer Search," *Marketing Science* 29, 1001-23 2010.

[7] Konovalov, Alexander and Zsolt Sándor: "On Price Equilibrium with Multi-Product Firms," forthcoming in *Economic Theory*.

[8] Koulayev, Sergei: "Estimating Demand in Search Markets: the Case of Online Hotel Bookings," Mimeo, 2010.

[9] Moraga-González, José Luis and Matthijs R. Wildenbeest: "Maximum Likelihood Estimation of Search Costs," *European Economic Review* 52, 820-48, 2008.

[10] Nevo, Aviv: "Measuring Market Power in the Ready-to-Eat Cereal Industry," *Econometrica* 69, 307-42, 2001.

[11] Reinganum, Jennifer F.: "A Simple Model of Equilibrium Price Dispersion," *Journal of Political Economy* 87, 851-58, 1979.

[12] Sovinsky Goeree, Michelle: "Limited Information and Advertising in the U.S. Personal Computer Industry," *Econometrica* 76, 1017-74, 2008.

[13] Stahl, Dale O.: "Oligopolistic Pricing with Sequential Consumer Search," *American Economic Review* 79, 700-12, 1989.

[14] Stigler, George: "The Economics of Information," *Journal of Political Economy* 69, 213-25, 1961.

[15] Wolinsky, Asher: "Product Differentiation with Imperfect Information," *Review of Economic Studies* 51, 53-61, 1984.

[16] Wolinsky, Asher: "Procurement via Sequential Search," *Journal of Political Economy* 113, 785-810, 2005.