Semi-Nonparametric Estimation of Consumer Search Costs *

José Luis Moraga-González[a,b], Zsolt Sándor †[c], and Matthijs R. Wildenbeest[d]

[a]Faculty of Economics and Business Administration, VU University, Amsterdam, The Netherlands
[b]Faculty of Economics and Business, University of Groningen, Groningen, The Netherlands
[c]Faculty of Economic and Human Sciences, Sapientia University, Miercurea Ciuc, Romania
[d]Kelley School of Business, Indiana University, Bloomington, IN, USA

Forthcoming in the *Journal of Applied Econometrics*

Abstract

This paper studies the estimation of the distribution of non-sequential search costs. We show that the search cost distribution is identified by combining data from multiple markets with common search technology but varying consumer valuations, firms' costs, and numbers of competitors. To exploit such data optimally, we provide a new method based on semi-nonparametric (SNP) estimation. We apply our method to a dataset of online prices for memory chips and find that the search cost density is essentially bimodal such that a large fraction of consumers searches very little, whereas a smaller fraction searches a relatively large number of stores.

Keywords: consumer search, oligopoly, search costs, semi-nonparametric estimation
JEL Classification: C14, D43, D83, L13

*We are indebted to the Editor, Thierry Magnac, and three anonymous referees for comments and suggestions that have substantially improved this paper. In addition, we wish to thank Vladimir Karamychev, Allard van der Made, Paulo K. Monteiro, Martin Pesendorfer, Michael Rauh, and Matthew Shum for their useful comments. The paper has also benefited from presentations at Universidad Carlos III de Madrid, University of Chicago, CORE, University of Essex, Tinbergen Institute, the ESEM Meetings 2006 (Vienna), the EEA Meetings 2007 (Budapest), and the IIOC 2009 (Boston). Financial support from the Netherlands Organization for Scientific Research (NWO) and from Marie Curie Excellence Grant MEXT-CT-2006-042471 is gratefully acknowledged. Moraga-González is also affiliated with the Tinbergen Institute, the CEPR, and the Public-Private Sector Research Center (IESE, Barcelona). Earlier drafts were circulated under the title "Nonparametric Estimation of the Costs of Non-Sequential Search."

†Correspondence to: Zsolt Sándor, Faculty of Economic and Human Sciences, Sapientia University, Piaţa Libertăţii 1, 530104 Miercurea Ciuc, Romania. E-mail: zsosan@gmail.com

1 Introduction

A significant body of work in economics has shown that search costs have far-reaching effects in economic activity. Well-known facts are that search costs alone can lead to price dispersion (Burdett and Judd, 1983; Stahl, 1989; Varian, 1980) as well as to wage and technology dispersion (Burdett and Mortensen, 1998; Acemoglu and Shimer, 2000). Search costs can also generate excessive product diversity in differentiated product markets (Wolinsky, 1984; Anderson and Renault, 2000) as well as inefficient quality investments (Wolinsky, 2005).

As a result, the estimation of consumer search costs has become an important area of empirical research. Hong and Shum (2006) were the first to develop a structural method to retrieve information on consumer search costs using market data. They focus on markets for homogeneous goods and present various approaches to estimate non-sequential and sequential consumer search models using only price data. Moraga-González and Wildenbeest (2008) present an alternative estimator based on maximum likelihood for non-sequential consumer search models. Hortaçsu and Syverson (2004) and Wildenbeest (2011) study search models where search frictions coexist with vertical product differentiation. In all these models, consumer search costs are found to be sizable and therefore an important source of market power.

The present paper studies the non-parametric identification and estimation of the costs of non-sequential search in markets for homogeneous products. It adds to the literature in three ways. First, we provide a proof that the critical search costs estimated in earlier work (cf. Hong and Shum, 2006; Moraga-González and Wildenbeest, 2008) are indeed non-parametrically identified. Second, we provide a new method based on semi-nonparametric (SNP) estimation that allows us to pool price data from different consumer markets with the same underlying search cost distribution but different valuations, selling costs, and numbers of competitors. Pooling data from different markets increases the number of estimated critical search cost cutoffs at all quantiles of the search cost distribution, which increases the precision of the estimate of the search cost CDF. Third, we provide sufficient conditions under which this type of data allows for the distribution of search costs to be identified on its full support.

The new method outperforms the spline approximation methods employed earlier in the literature (Hortaçsu and Syverson, 2004; Hong and Shum, 2006; Moraga-González and Wildenbeest, 2008) with this type of data. Instead of estimating the parameters of the price distribution market by market, which ignores the link between the different datasets, our semi-nonparametric approach takes the search cost density as a parameter of the likelihood function and exploits all the data at once when estimating the model. SNP density estimators use a flexible polynomial-type parametric function that can approximate arbitrarily closely a large class of sufficiently smooth density functions (Gallant and Nychka, 1987), which means we obtain an essentially nonparametric estimator of the search cost distribution common to all markets.

To illustrate how our method works with real-world data we apply the SNP estimation procedure to a dataset of online prices for ten notebook memory chips. Our estimates indicate that median search costs are close to \$5. Search costs are quite dispersed; the majority of consumers visits at most three online stores before buying, while only a small fraction of consumers searches more than four times. Similar findings have been reported in several other empirical studies (Moraga-González and Wildenbeest, 2008; De los Santos, 2008; De los Santos *et al.*, 2011). Consumers with high search costs do not search much and therefore do not compare many prices, which gives substantial market power to the firms; as a result, estimated price-cost margins are significantly larger than what one would expect on the basis of the observed large number of firms operating in each market.

The next section reviews the non-sequential consumer search model. The identification result and the SNP estimation method are presented in Section 3. Section 4 is devoted to the empirical application. Finally, Section 5 concludes. Proofs are relegated to an Appendix.

2 The model

The model we study was proposed by Hong and Shum (2006) and generalizes the non-sequential consumer search model of Burdett and Judd (1983) by adding consumer search cost heterogeneity.[1] There is a finite number of firms K producing a good at constant returns to scale. The unit cost of the firms is equal to r. There is a unit mass of buyers. Each consumer wishes to purchase a single unit of the good at most. The maximum price any buyer is willing to pay

[1] Janssen and Moraga-González (2004) study the same model assuming the search cost distribution has a two-point support.

for the good is v. Consumers must engage in costly search to observe prices. We assume they search non-sequentially. In addition we assume that the first price quotation is observed at no cost.[2] Once a consumer has observed the desired number of prices, she chooses to buy from the store charging the lowest price. Consumers differ in their costs of search. A buyer's search cost is drawn independently from a common atomless distribution G with support $(0, \infty)$ and positive density g everywhere. A consumer with search cost c who searches k firms incurs a total search cost of $(k-1)c$. The maximum number of prices a consumer can observe is K.

Firms and buyers play a simultaneous moves game. The solution concept is Nash equilibrium. An individual firm chooses its price strategy taking the price strategies of the rivals as well as consumers' search behavior as given. To allow for both pure and mixed pricing strategies, a firm i's strategy is denoted by a probability distribution of prices F_i. Let F_{-i} denote the vector of pricing strategies used by firms other than i. The (expected) profit to firm i from charging a price p_i given the rivals' pricing strategies F_{-i} is denoted as $\Pi_i(p_i, F_{-i})$.

Likewise, an individual buyer takes as given the firms' pricing strategies and decides on her optimal search strategy to maximize her expected utility. The strategy of a consumer with search cost c is then a number k of prices to search for. Let the fraction of consumers searching k firms be denoted by μ_k.

We shall concentrate on symmetric Nash equilibria, i.e., equilibria where $F_i = F$ for all i. A symmetric equilibrium is a distribution of prices F and a collection $\{\mu_k\}_{k=1}^{K}$ such that: (a) $\Pi_i(p; F) \leq \overline{\Pi}$ for all p outside the support of F for all i; (b) $\Pi_i(p; F)$ is equal to a constant $\overline{\Pi}$ for all p in the support of F, for all i; (c) a consumer searching for the prices of k firms obtains no lower utility than by searching for any other number of prices; (d) $\sum_{k=1}^{K} \mu_k = 1$.

Condition (a) is the standard Nash requirement that a firm must play a best-response to the strategies of the other players; condition (b) says that if the firms use a mixed strategy in equilibrium, then they must be indifferent among all the prices in the support of F; finally, conditions (c) and (d) require that consumers search in such a way that their (expected) utility is maximized. Let us denote the equilibrium density of prices by f, with maximum price $\overline{p}$ and

[2] In our setting with search cost heterogeneity this assumption is inconsequential and can easily be relaxed at the cost of some additional notation. Earlier literature has assumed the first search to be costless in order to avoid problems of existence of equilibrium (cf. Diamond's (1971) paradox). To keep with the earlier literature we also maintain such an assumption here.

minimum price $\underline{p}$.

2.1 Nash equilibria

We start by observing that in our game there may be two types of equilibria. There may be an equilibrium in pure strategies in which all firms charge a price equal to v and consumers optimally respond by not searching at all and visiting just one firm. This equilibrium is rejected right away in most empirical settings since we typically observe firms charging different prices while consumers are actively searching the market. Moreover, this equilibrium is non-robust in the sense that it heavily relies on the assumption that the first search is conducted at no cost; in fact, when the first search is costly this pure-strategy equilibrium fails to exist.

There may also be an equilibrium in mixed pricing strategies. In this equilibrium firm prices are dispersed and consumers respond by searching optimally to maximize their expected utility. Since both are common to most empirical settings this will be the equilibrium we will focus on. For this equilibrium to exist, there must be some consumers who search for one price only and others who search for two prices or more, i.e., $1 > \mu_1 > 0$ and $\mu_k > 0$ for some $k = 2, 3, \ldots, K$. The intuition behind this observation is as follows. Suppose all consumers did search at least for two prices. If this were so, all firms would then be subject to price comparisons with rival firms. As a result, the firms would encounter themselves in a situation identical to the so-called Bertrand paradox (see e.g Tirole, 1988). In such a situation, firms would have an incentive to undercut one another and thus all prices would be equal to marginal cost. Suppose now that no consumer did search at all. Then, since consumers would not be able to compare the prices of different firms, they would charge the monopoly price v and again there would not be price dispersion in the market.

Our second observation is that, given that in equilibrium some consumers must search just once and others more than once, it must be the case that in symmetric equilibrium firms play mixed strategies with atomless price distributions.[3] The intuition behind this remark is as follows. Suppose firms used distributions with an atom at a price $p \in (r, v]$. Since a price-tie at p would occur with strictly positive probability, an individual firm would gain by undercutting

[3] That is, discrete distributions or continuous distributions with "jumps" can be ruled out.

the tied price p, thereby attracting a larger share of the consumers who search for various prices and so obtaining greater profits. If there is an atom at $p = r$, then an individual firm would obtain zero profits; because some consumers do not search at all and therefore accept any price below or at v, the firm would have an incentive to deviate by increasing its price.

A third remark is that the upper bound of the symmetric equilibrium price distribution F must be equal to v. The reason for this is as follows. Suppose the upper bound were $\overline{p} < v$ and consider a firm charging $\overline{p}$. Since this firm would not sell to any of the consumers who search for various prices, its payoff would simply be equal to $(\overline{p} - r)\mu_1/K$, which is strictly increasing in $\overline{p}$; as a result the firm would gain by deviating and charging v instead of $\overline{p}$.

Now that we have presented the basic properties of the mixed pricing strategy of the firms, let us consider the problem faced by the consumers. A consumer with search cost c must choose a number of prices to maximize her expected utility, where expected utility is equal to the difference between the consumer's valuation and the price she expects to pay, minus the cost of searching. If the consumer picks k prices to be searched, her expected utility is therefore given by $v - Ep_{1:k} - (k-1)c$, where $Ep_{1:k}$ is short-hand notation for $E[\min\{p_1, p_2, \ldots, p_k\}]$ and E indicates the expectation operator. Since every price is a random draw from F, the distribution of the minimum of k prices is equal to $(1 - F(p))^k$. Therefore, the number of prices that maximizes the utility of a consumer with search cost c is given by

$$k(c) = \arg\min_{k} \left[(k-1)c + \int_{\underline{p}}^{v} pk(1 - F(p))^{k-1} f(p) dp \right]. \tag{1}$$

Since $k(c)$ must be an integer, the problem in equation (1) induces a partition of the set of consumers into groups μ_k of consumers searching for k prices, with the property that $\sum_{k=1}^{K} \mu_k = 1$. We now describe such a partition. First we define the search cost cutoff c_k as the search cost of a consumer indifferent between searching for k prices, which gives her a utility equal to $v - Ep_{1:k} - (k-1)c$, and searching for $k+1$ prices, which allows her to obtain a utility level equal to $v - Ep_{1:k+1} - k \cdot c$. Solving for c_k gives

$$c_k = Ep_{1:k} - Ep_{1:k+1}, \ k = 1, 2, \ldots, K-1. \tag{2}$$

Since c_k decreases in k,[4] the fractions of consumers searching for k prices, μ_k, are given by

$$\mu_1 = 1 - G(c_1); \quad (3a)$$

$$\mu_k = G(c_{k-1}) - G(c_k), \ k = 2, 3, \ldots, K. \quad (3b)$$

To complete the equilibrium characterization, we now look at how firms should choose the distribution of prices F to maximize their profits given consumers' search behavior. The expected profit to a firm i charging p_i when rivals are setting prices according to strategy F is given by

$$\Pi_i(p_i; F) = (p_i - r)\left[\sum_{k=1}^{K} \frac{k}{K}\mu_k(1 - F(p_i))^{k-1}\right].$$

To understand this equation, note that firm i obtains a per consumer profit of $p_i - r$ and sells to a consumer who searches for k prices whenever the prices of the other $k-1$ firms observed by this consumer are all higher than the price of firm i, which occurs with probability $(1 - F(p_i))^{k-1}$.

In a mixed strategy equilibrium, all the prices in the support of F must give the firm the same level of profits. Thus, for any price p in the support of F it must be the case that $\Pi_i(p; F) = \Pi_i(v; F)$. As a result, equilibrium requires

$$(p - r)\left[\sum_{k=1}^{K} k\mu_k(1 - F(p))^{k-1}\right] = \mu_1(v - r) \quad (4)$$

to hold for all prices p in the support of F. Setting $F = 0$ in this equation and solving for p gives the minimum price charged in the market:

$$\underline{p} = \frac{\mu_1(v - r)}{\sum_{k=1}^{K} k\mu_k} + r. \quad (5)$$

In Moraga-González *et al.* (2010) we show that an equilibrium always exists.

[4] The cutoffs $c_k = Ep_{1:k} - Ep_{1:k+1}$ are in fact strictly monotonically decreasing in k because $Ep_{1:k}$ is strictly convex in k. A proof of this is available from the authors upon request. See also Stigler (1961), who mentions this property.

3 Econometric analysis

The econometric problem is to estimate the search cost distribution G using price data. Hong and Shum (2006) and Moraga-González and Wildenbeest (2008) propose different methods that exploit equations (2) to (5) to estimate the search cost CDF. In what follows, we briefly explain the two methods proposed so far (for details we refer the reader to their original contributions).

Hong and Shum (2006) formulate the estimation of the unknown search cost distribution as a two-step procedure. They propose to estimate first the parameters $\{\mu_k\}_{k=1}^K$ of the equilibrium price distribution obtained from equation (4) by maximum empirical likelihood (MEL), and then to recover the collection of cutoffs in equation (2) using the empirical CDF of prices. Suppose the researcher has a (large) dataset with n prices and suppose $K(\leq n-1)$ is the maximum number of prices a consumer may observe in the market. Let us assume each price p_j has probability π_j. Using equilibrium condition (4), for each price p_i we have the approximate equality

$$(p_i - r)\left[\sum_{k=1}^{K} k\mu_k \left(1 - \left[\sum_{j=1}^{n} \pi_j \mathbf{1}(p_j \leq p_i)\right]\right)^{k-1}\right] \simeq (v-r)\mu_1, \tag{6}$$

which can be transformed into a number $Q \geq K$ of population quantile restrictions:

$$\sum_{j=1}^{n} \pi_j \left[\mathbf{1}\left(p_j \leq r + \frac{(v-r)\mu_1}{\sum_{k=1}^{K} k\mu_k \left(1 - s_\ell\right)^{k-1}}\right) - s_\ell\right] \simeq 0 \tag{7}$$

for $s_\ell \in [0,1]$, $\ell = 1, 2, \ldots, Q$. Using the lower bound defined in equation (5) one can eliminate marginal cost r from these constraints. Then, using MEL based on these constraints, one can obtain estimates of the parameters $\{\mu_k\}_{k=1}^K$. Finally, by combining these estimates with the cutoff points in equation (2) obtained directly from the empirical CDF of prices, one gets K points $\{(c_k, G(c_k))\}_{k=1}^K$ on the search cost distribution. These points serve to construct an estimate of the search cost CDF by interpolation.

Moraga-González and Wildenbeest (2008) put forward an alternative maximum likelihood (ML) method. There are two differences with respect to Hong and Shum's method. First, they compute the likelihood of a price as a function of F and exploit the equilibrium constancy-of-profits condition (4) to numerically calculate the value of the price CDF. In this way they obtain ML estimates of the parameters $\{\mu_k\}_{k=1}^K$. The second difference is that they introduce a method to compute ML estimates of the cutoffs by rewriting equation (2) as

$$c_k = \int_0^1 p(z)[(k+1)z - 1](1-z)^{k-1}dz, \quad k = 1, 2, \ldots, K-1. \tag{8}$$

where $p(z)$ is the inverse of the price distribution obtained from equation (4):

$$p(z) = \frac{\mu_1(v-r)}{\sum_{k=1}^K k\mu_k(1-z)^{k-1}} + r. \tag{9}$$

These two methods yield estimates of the points $\{(c_k, G(c_k))\}_{k=1}^K$ of the search cost distribution. Under the standard regularity conditions, these points are estimated consistently. These two papers base their asymptotics on the number of prices n going to infinity. Although one of the regularity conditions is identification of the points $\{(c_k, G(c_k))\}_{k=1}^K$ of the search cost CDF, none of the earlier papers studied the identification issue. In the next subsection, we show that the sequence of points $\{(c_k, G(c_k))\}_{k=1}^K$ is identified.

3.1 Identification of search costs

Here we ask whether the search cost distribution G can be non-parametrically identified when the price distribution F is known by the researcher. This treatment of the identification problem is common in nonparametric estimation and is in the spirit of Koopmans and Reiersøl (1950).

The analysis in Section 2 shows how the model maps G into F. A feature of the model is that the entire set of consumers is partitioned into K groups of them. As a result, using the price equilibrium mapping, one can only hope to recover the (countable) sequence of points $\{(c_k, G(c_k))\}_{k=1}^K$ of the search cost CDF. The proposition below, proved in the Appendix, shows indeed that if we know the price distribution F, then we can identify the values of the search cost CDF corresponding to the cutoffs $\{c_k\}_{k=1}^K$.

Proposition 1 *Suppose that the econometrician observes the equilibrium price distribution F with support $(\underline{p}, v)$, which is continuous and is generated by the vector of variables (G, v, r, K) through equations (2), (3a), (3b), and (4). Then the points of the search cost distribution G*

corresponding to the sequence $\{c_k\}_{k=1}^K$ *are identified.*

Obviously, when K is small, the sequence of points $\{(c_k, G(c_k))\}_{k=1}^K$ will be insufficient to obtain a precise estimate of the search cost distribution. The question that arises is whether such sequence allows us to recover the search cost CDF when $K \to \infty$. The answer is negative because, as shown above, the sequence $\{c_k\}_{k=1}^K$ is decreasing and therefore convergent. As a result, data from a single market does not provide the econometrician with sufficient information to recover search costs at relatively high quantiles. The purpose of the remainder of this subsection is to deal with the problem of identification of the search cost distribution in the relevant support. Our proposal consists of bringing additional information to be able to estimate search costs more fully. Pooling data from various markets with similar search technology but different valuations, selling costs or numbers of competitors naturally lends itself as a feasible strategy to solve this identification problem. Implementing this idea in practice is not straightforward and in Section 3.2 we propose a new estimation method to do it.

The next proposition presents sufficient conditions under which the search cost distribution is identified on its full support using price data from multiple markets.

Proposition 2 *Suppose that there are infinitely (countably) many markets, indexed by* m*, all of them with the same underlying search cost distribution* G*. In every market* m*, the econometrician observes the price distribution function* F^m *with support* $(\underline{p}^m, v^m)$*, which is continuous and is generated by the vector of variables* (G, v^m, r^m, K^m) *through equations (2), (3a), (3b), and (4). In addition, assume that the difference between valuations and marginal costs* $\{v^m - r^m\}_{m\geq 1}$ *are random variables drawn independently from a distribution with support* $(0, \infty)$*. Then* G *is identified on the interval* $[0, \sup_m c_1^m]$*, where* $\sup_m c_1^m = \sup\{c_1^m : m = 1, 2, \ldots\}$ *is the supremum of the set of* c_1*-cutoff points from all the markets.*

This result says that one can solve the identification problem mentioned above by combining price data from various markets for which the search technology is similar but there is variation in valuations, selling costs, and numbers of competitors. Gathering the appropriate data is then relatively easy for the researcher. One just needs to take markets for different products in which consumers search for low prices in a similar fashion. To provide an example, if one aims to estimate the costs of search in the market for carpentry, one could pool data from the various professional services needed to refurbish a house: a carpenter, an electrician, a painter, a plumber, a bricklayer, a tiler, etc. The search technology to find acceptable prices for these professional services is basically the same; however, valuations, costs and the number of available professionals can be quite different across these services. This sort of data will do.[5]

Intuitively, pooling data from various markets solves the problem of identification because every market generates a distinctive sequence of cutoff points, $\{c_k^m\}_{k=1}^{Km}$, and this forces the search cost distribution function to be uniquely determined for a larger set of points, $\{\{c_k^m\}_{k=1}^{Km}\}_{m\geq 1}$. Under the (large support) condition of the proposition, this set of points can be shown to be dense in the interval $[0, \sup_m c_1^m]$. If $\sup_m c_1^m = \infty$, then this proposition establishes identification of the search cost distribution in the full support.

We note that the (large support) condition in Proposition 2 is a sufficient condition also used in related nonparametric identification problems (see e.g. Matzkin, 1992; Matzkin, 1993; Ichimura and Thompson, 1998; Berry and Haile, 2009). In our case, we have adopted it to simplify the proof of identification. The assumption allows us to rely only on the cutoff points c_1^m to show identification.[6]

3.2 Estimation of search costs

As mentioned above, the previous studies on estimation of search cost distributions proceed in three steps: first, the parameters $\{\mu_k\}_{k=1}^K$ of the price distribution are estimated; second, the search cost cutoff points $\{c_k\}_{k=1}^K$ are obtained using the parameters of the price distribution; finally, a spline approximation of the search cost distribution is constructed by interpolating the sequence of estimated points $\{(c_k, G(c_k))\}_{k=1}^K$. As shown in the previous section, identification requires to pool data from many markets and in such a framework this earlier procedure presents

[5]If alternatively one considers the costs of search for prices on the Internet, one could take markets for different books, CDs, or DVD movies; in our application in Section 4, we use data from multiple markets for memory chips.

[6]Since the large support condition is a sufficient condition, the identification result could be true under a weaker assumption. The proof would however be much more difficult because one would need to use the additional variation obtained from the other cutoff points c_k^m's. The problem is that the mathematical relationship among all the c_k's in a market, which is given by system of equations (8), is highly nonlinear and therefore using the other cutoffs in the proof of Proposition 2 is quite difficult. We would also like to note that the variation in K^m across markets is another source of variation that is useful in applications.

some problems. In fact, it has to be applied market by market, in which case the researcher obtains multiple search cost estimates, one for each market. Interpolation is no longer feasible and one would have to fit a curve through the different estimated search cost distributions. It is not clear how one should proceed in such a case. For example, because the number of competitors K^m varies from market to market, whether all the markets should be allocated the same weight when fitting the curve is open to question. These difficulties lead us to propose a new estimation method that addresses these issues naturally.

We propose to employ semi-nonparametric (SNP) maximum likelihood estimation (Gallant and Nychka, 1987). The advantage of this method is that it is not applied market by market but designed to maximize the likelihood from all the markets jointly. In this way, the SNP procedure exploits the link between the prices not only within a market but also across markets because they all have the same underlying search cost distribution. We note that this method is different in essence because it takes the search cost density directly as the parameter of the likelihood. In this sense, it exploits the data more efficiently than the previous spline methods, since those rely on estimating the parameters of the price distribution in every market separately and, therefore, ignore the link between the different datasets.

The idea behind SNP estimation is to use a flexible functional approximation of the search cost density. This functional approximation depends on a finite set of parameters to be estimated and this set can be made arbitrarily large as the number of observations goes to infinity. We construct our estimator of the search cost density by employing a flexible polynomial-type approximation, following the SNP estimation technique of Gallant and Nychka (1987).

The likelihood function can be constructed by deriving the density of prices in each market m as a function of the search cost density g. Let $f^m(p_i|g)$ denote the density of price p_i observed in a market m given the search cost distribution g. Since the prices in a market m are independent draws from F^m, the log-likelihood function in market m is $L^m(g|\mathbf{p}_m) = \sum_{i=1}^{K^m} \log f^m(p_i|g)$ where $\mathbf{p}_m$ is the K^m-dimensional vector of prices in market m. In order to compute this, first we apply the implicit function theorem to equation (4), which yields:

$$f^m(p_i|g) = \frac{\sum_{k=1}^{K^m} k\mu_k^m (1 - F^m(p_i|g))^{k-1}}{(p_i - r^m)\sum_{k=1}^{K^m} k(k-1)\mu_k^m (1 - F^m(p_i|g))^{k-2}}. \tag{10}$$

The quantities μ_k^m and r^m in this expression need to be computed in terms of g. By solving equation (5) for r^m we obtain an expression for the marginal cost in market m

$$r^m = \frac{\underline{p}^m \sum_{k=1}^{K^m} k\mu_k^m - \mu_1^m v^m}{\sum_{k=2}^{K^m} k\mu_k^m}.$$

We can (superconsistently) estimate the lower and upper bounds $\underline{p}^m$ and v^m of the price distribution in a market m by taking the minimum and maximum prices, respectively, observed in the data.[7] Then, for every market m, we compute $\{c_k^m\}_{k=1}^{K^m}$ from equation (8) in terms of g,[8] and then use equations (3a), (3b), (4), and (10) to find the values of $F^m(p_i|g)$ and $f^m(p_i|g)$. In this way we obtain the joint log-likelihood of all markets as a function of g:

$$L_M(g|\mathbf{p_1}, \mathbf{p_2}, \ldots, \mathbf{p_M}) = \frac{1}{M} \sum_{m=1}^{M} \left(\frac{1}{K^m} \sum_{i=1}^{K^m} \log f\left(p_i^m|g\right) \right)$$

For the polynomial-type parametric function that estimates the search cost density we employ the SNP density estimator of Gallant and Nychka (1987). This SNP estimator is based upon a Hermite polynomial expansion. The idea behind their SNP procedure is that any reasonable density can be mimicked by such a Hermite polynomial series. SNP density estimators are essentially nonparametric because the set of all Hermite polynomial expansions is dense in the set of density functions that are relevant (Gallant and Nychka, 1987).[9]

To apply the SNP estimation to our problem, we specify the search cost density as follows:

$$g(c; \gamma, \sigma, \theta) = \frac{\left[\sum_{i=0}^{N} \theta_i u_i(c)\right]^2}{\sum_{i=0}^{N} \theta_i^2}, \;\; \theta \in \Theta_N \tag{11}$$

[7]In a similar fashion, order statistics are also used to estimate the lower and upper bound of distributions of bids (see e.g. Donald and Paarsch, 1993).

[8]In markets with many cutoff points solving this nonlinear system of equations may be time consuming. One alternative (used in the application) is to estimate the cutoffs directly by the empirical price CDF. The trade-off is precision of the estimates against computational time.

[9]SNP has recently been applied to the estimation of labor search frictions (Koning *et al.*, 2000), labor supply (Van Soest *et al.*, 2002), travel demand (Van der Klauw and Koning, 2003), and auctions (Brendstrup and Paarsch, 2006).

where $\Theta_N = \{\theta : \theta = (\theta_0, \theta_1, \ldots, \theta_N), \theta_0 = 1\}$, N is the number of polynomial terms and

$$
\begin{aligned}
u_0(c) &= (c\sigma\sqrt{2\pi})^{-1/2}\, \mathrm{e}^{-((\log c - \gamma)/\sigma)^2/4}, \\
u_1(c) &= (c\sigma\sqrt{2\pi})^{-1/2}((\log c - \gamma)/\sigma)\, \mathrm{e}^{-((\log c - \gamma)/\sigma)^2/4}, \\
u_i(c) &= \left[((\log c - \gamma)/\sigma)u_{i-1}(c) - \sqrt{i-1}u_{i-2}(c)\right] / \sqrt{i}, \text{ for } i \geq 2.
\end{aligned}
$$

This parametric form corresponds to the univariate SNP estimator studied extensively by Fenton and Gallant (1996). Our expressions are obtained by transforming their random variable x with the density defined in their Section 4.3 into $c = \exp(\gamma + \sigma x)$. This transformation is useful in our case since search costs are positive. The vector of parameters to be estimated by maximum likelihood is $\{\gamma, \sigma, \theta_1, \ldots, \theta_N\}$ and N can be made arbitrarily large as the number of observations increases to infinity. In practice the number of polynomial terms N has to be chosen in an optimal way. For this, we can build on the cross-validation method of Coppejans and Gallant (2002). The essence of their cross-validation is to determine N for the data at hand by minimizing some loss function. In the empirical example the prices from different markets have different distributions, so we take the approximation of the average ISE (i.e., integrated squared error) across all markets, i.e.,

$$
\frac{1}{M}\sum_m \int_{\underline{p}}^{\overline{p}} \left(\widehat{f_N^m}(p) - f^m(p)\right)^2 dp,
$$

where f^m denotes the true price density function in market m and where $\widehat{f_N^m}$ is market m's price density function estimate using N polynomial terms.[10]

Gallant and Nychka (1987) provide conditions on the unknown density (e.g. differentiability and restricted tail behavior) under which their estimator is consistent using i.i.d. observations.[11] In Moraga-González *et al.* (2012) we present a Monte Carlo exercise that explores the small sample properties of the estimator. Using a similar number of observations and number of markets as in our application, we find that the estimates mimic the true shape of the search

[10] See Moraga-González *et al.* (2012) for details concerning how to approximate the ISE in case the true price density function is not known, as in empirical applications.

[11] In Moraga-González *et al.* (2012) we provide details on how these conditions can be adapted to our search cost density. In addition we discuss some primitive conditions specific to our model.

cost CDF as well as PDF relatively well at most quantiles. In addition we find that in our setting cross-validation works well in picking the number of SNP parameters. The simulation results also indicate that our estimator works better than an estimator that does not directly link different markets but instead estimates search costs market-by-market (based on Moraga-González and Wildenbeest, 2008).

4 Application

In this section we use the SNP estimation method described above to quantify search costs in real-world markets for memory chips. We focus on computer memory chips for notebooks (so called SO-DIMM, or Small Outline Dual In-line Memory Module). Since we need products from different markets, we select memory chips produced for different brands and types of notebooks. Table 1 gives the details of the ten products we include in our dataset. There are several reasons for choosing these memory chip data for the analysis. First, since all the chips are sold online, we expect search costs to be similar across markets. Second, even though all memory chips are manufactured by Kingston, the largest producer in the sector, each memory chip in our sample is meant to be used in a particular notebook brand only—including Toshiba, Dell, Acer, IBM (now Lenovo), and HP Compaq. Given that substitutability across products is somewhat limited due to technical reasons, we shall assume that different microchips belong in separate markets so the use of a search model with homogeneous products is reasonable.[12] All the memory chips we consider were somewhat at the top of the product line at the time of data collection. In particular, they exhibit relatively large storage capacity (1 gigabyte) and fast speed of operation (most of them above 400 MHz). Given the large storage capacity of the memory chips in the dataset, most consumers would only consider to buy one memory chip, so the single-unit inelastic demand assumption of the theoretical model seems reasonable.

For all the memory chips in the dataset we collected online prices charged in the US, in February 2006. To obtain a sufficiently representative sample, we gathered product and price information from several sources at the same time. We proceeded as follows. We first visited

[12] Note that even though (within a market) the memory chips are exactly the same, the stores selling the chips might differ in terms of offered service, speed and quality of shipment, payment methods, etc. We come back to this issue at the end of this section.

Table 1: List of products

Part number	Manufacturer	Compatibility	Size	Speed	Form factor
KTT3311A	Kingston	Toshiba	1GB	333MHz DDR333/PC2700	200-pin SoDIMM
KTT533D2	Kingston	Toshiba	1GB	533MHz DDR2-533/PC2-4200	200-pin SoDIMM
KTD-INSP8200	Kingston	Dell	1GB	266MHz DDR266/PC2100	200-pin SoDIMM
KTD-INSP5150	Kingston	Dell	1GB	333MHz DDR333/PC2700	200-pin SoDIMM
KTD-INSP6000	Kingston	Dell	1GB	533MHz DDR2-533/PC2-4200	240-pin SoDIMM
KTD-INSP6000A	Kingston	Dell	1GB	533MHz DDR2-533/PC2-4200	200-pin SoDIMM
KAC-MEME	Kingston	Acer	1GB	533MHz DDR2-533/PC2-4200	200-pin SoDIMM
KTD-INSP9100	Kingston	Dell	1GB	400MHz DDR400/PC3200	200-pin SoDIMM
KTM-TP3840	Kingston	IBM	1GB	533MHz DDR2-533/PC2-4200	200-pin SoDIMM
KTH-ZD8000A	Kingston	HP Compaq	1GB	533MHz DDR2-533/PC2-4200	200-pin SoDIMM

the price comparison sites *shopper.com* and *pricegrabber.com* and collected the names of all the shops that were seen active in markets for memory chips; in total we found 49 stores. If for a particular product we saw a shop quoting its price on shopper.com and/or pricegrabber.com, we took the price directly from the price comparison site; otherwise we visited the web-address of the vendor to check if the product was available and at what price it was offered.

Table 2 gives some summary statistics of the dataset. The number of firms quoting prices in each market is relatively large, ranging from 24 to 41. In our study we estimate the maximum number of prices consumers can search for in each market K^m by the number of firms that were observed to be quoting prices in that market. Almost all memory chips are priced above \$100. For all products we observe significant price dispersion as measured by the price range (difference between the maximum and the minimum prices) and by the coefficient of variation. We note that the (gross) benefits to a consumer from searching are significant; in particular, the (gross) gains from searching all the firms thereby becoming fully informed relative to searching for one price only in these markets range from \$16.32 to \$33.05. As mentioned above, we estimate the valuation of a memory chip by the maximum price observed in the market.

Our model assumes consumers search non-sequentially. Consumers obviously visit stores sequentially in the real world, so what truly distinguishes non-sequential search from sequential search is how consumers select the stores they visit—if they search non-sequentially, the number of stores searched is determined before searching, while if they search sequentially, the number of searches depends on what has been observed. Although non-sequential search is often thought of as a constrained version of sequential search, Morgan and Manning (1985) have shown that the optimal search rule is hybrid in nature: it includes decisions on the sample size as well as

Table 2: Summary statistics

Part number	No. of Stores	Mean Price (Std)	Min. Price	Max. Price	CV (as %)
KTT3311A	32	181.67 (24.62)	148.62	235.00	13.55
KTT533D2	33	123.33 (15.62)	100.45	161.40	12.66
KTD-INSP8200	39	173.59 (21.31)	148.62	249.54	12.28
KTD-INSP5150	39	179.09 (19.84)	148.62	222.35	11.08
KTD-INSP6000	35	120.29 (13.48)	100.45	151.05	11.21
KTD-INSP6000A	38	116.33 (13.43)	94.99	154.50	11.54
KAC-MEME	24	123.58 (17.47)	101.92	161.64	14.14
KTD-INSP9100	33	175.84 (24.38)	148.62	249.54	13.87
KTM-TP3840	37	122.83 (14.32)	104.55	161.94	11.65
KTH-ZD8000A	41	116.77 (12.25)	100.45	154.50	10.49

Notes: Prices are in US dollars. Std stands for standard deviation. CV (coefficient of variation) is calculated as the ratio of the standard deviation to the mean.

whether to continue searching or not. This means both non-sequential and sequential search are special cases—non-sequential search is typically optimal when the search outcome is observed with delay, for instance when applying for a job or college, or when obtaining estimates from contractors. Even though a typical online shopper does not face such a delay when searching online, according to Manning and Morgan (1982) sufficiently large economies of scale when searching can make it optimal to search multiple firms at once without using the option to continue searching. This is typically the case when searching online: once having found the correct memory chip and price at one web store, simply copying and pasting the chip's model number to another online store is all it takes to obtain an additional price quote. This might also explain why, using individual specific observations on browsing history, De los Santos *et al.* (2011) find that observed search patterns for online books are more consistent with non-sequential search than sequential search.

Because we only observe the stores' prices at one moment in time, we cannot check whether stores indeed use mixed pricing strategies, as predicted by our search model. However, using a different dataset Moraga-González and Wildenbeest (2008) show that firms indeed seem to mix prices in the online market for memory chips; at the same time, other studies find evidence for mixed strategies in other markets (e.g., Lach (2002) for chicken, refrigerators, coffee, and flour in Israel; Lewis (2008) for gasoline; and Wildenbeest (2011) for groceries in the UK).

We follow the procedure explained in Section 3.2 and use cross-validation for choosing the number of polynomial terms N. Table 3 gives the SNP estimation results for different values of N. These results are obtained using the empirical price CDF in each market to calculate the

Table 3: Fit SNP estimates for different values of N

N	LL	ISE	Conditional Kolmogorov statistic	p-value
1	40.047	-0.0416	1.143	0.00
2	40.011	-0.0414	1.140	0.00
3	39.532	-0.0449	0.855	0.01
4	39.527	-0.0451	0.870	0.01
5	39.525	-0.0451	0.877	0.00
10	39.258	-0.0484	0.639	0.14
15	39.207	-0.0487	0.559	0.23
20	39.180	-0.0492	0.562	0.24
25	39.158	-0.0491	0.736	0.08
30	39.146	-0.0492	0.560	0.24
35	39.141	-0.0493	0.587	0.19
40	39.133	-0.0492	0.556	0.25
45	39.131	-0.0493	0.572	0.22
50	39.127	-0.0493	0.555	0.25

Notes: N is the number of SNP parameters. LL is the log-likelihood value. ISE is the integrated squared error.

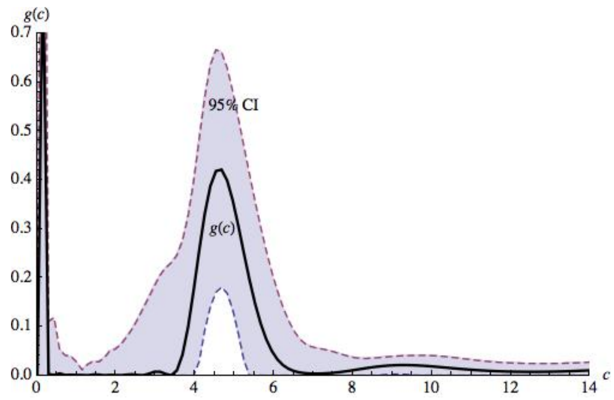
c_k's.[13] As can be seen in the table, up to $N = 20$ there is a steady improvement in ISE, while for larger N the improvement is very small. We therefore use the estimated parameters for $N = 20$ to derive our estimate of the search cost distribution. The solid curve in Figure 1(a) denotes the estimated search cost PDF, while the shaded area indicates the 95 percent confidence interval.[14] Similarly, Figure 1(b) gives the estimated search cost CDF, with the shaded band depicting the 95 percent confidence interval. The graph also shows how the estimated search cost cutoffs (gray dots) cover the support of the search cost distribution.

To test whether the estimated model explains observed prices well, we use the Conditional Kolmogorov (CK) test developed by Andrews (1997). The null hypothesis of the test is that the model, conditional on the estimated SNP parameters and market covariates, is correctly specified; the alternative hypothesis is that the model is incorrectly specified. The fourth column of Table 4 gives the CK test results for various values of N, whereas the last column of the table gives corresponding p-values.[15] For p-values that exceed 0.05 we cannot reject the null

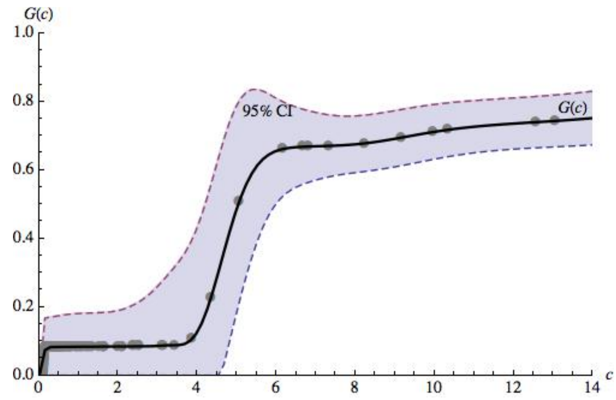
[13] In cases when there are sufficiently many observations, as is the case in our dataset, we can use the empirical distribution of prices in each market directly to estimate the c_k's. The gain in computing time is huge and the results for our data are very similar. See also Figure 2(a).

[14] Since standard errors of the parameter estimates are only meaningful in the case where the presented model is the true parametric model, we have obtained the confidence interval using bootstrapping. The 95 percent confidence interval is obtained using $\left(g_N(c) - 1.96\sqrt{\mathrm{var}(g_N(c))}, g_N(c) + 1.96\sqrt{\mathrm{var}(g_N(c))}\right)$, where $\mathrm{var}(g_N(c))$ is estimated by bootstrap resampling. In Moraga-González *et al.* (2012) we discuss the algorithm for constructing the confidence band in more detail. In addition we provide the results of a small Monte Carlo study on the coverage probability of the confidence interval.

[15] The test statistic is calculated as $CK = \max_{1\leq j\leq n} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[(p_i \leq p_j) - F\left(p_j | X_i, \widehat{\theta}\right) \right] (X_i \leq X_j) \right|$, where



(a) Search cost PDF



(b) Search cost CDF

Figure 1: Estimated search cost distribution

hypothesis, which is the case for $N = 20$. Only for specifications with a relatively low number of SNP parameters we can reject the equality of the empirical price CDFs and the estimated price CDFs.

Using the estimates of the parameters of the SNP specification we can compute the mean, the median, and the standard deviation of the search cost distribution. The median consumer has a search cost equal to \$5.05. On average a consumer has a search cost value equal to \$8.70 and the standard deviation is \$7.35. It is also interesting to investigate the distribution of search intensities in these markets. Since each market has specific parameters, even though search costs are assumed to be similar, it is unlikely that consumer search behavior will be the same across markets. Table 4 shows that it is indeed the case that search intensities are different across markets. For example, in the market for the KTT3311A memory chip, 26 percent of consumers searches for one price only while in the market for the KTH-ZD8000A memory chip the share of consumers who searches once is 34 percent. Similarly, in the market for the KTT3311A chip, 24 percent of consumers searches for two prices, while in the market for the KTD-INSP8200 memory chip the share of consumers who searches for two prices is 62 percent. However, the share of consumers searching at most three times is more or less similar across markets; approximately 91 percent of the consumers have search cost above \$3.70 and

n is the total number of price observations, $\widehat{\theta}$ is a $\sqrt{n}$-consistent estimator of the parameters of the SNP density θ, $(p_i \leq p_j)$ is the indicator of the event $p_i \leq p_j$, $F\left(p_j|X_i, \widehat{\theta}\right)$ is the parametric conditional price CDF (the estimated price CDF), and X_i is a vector of market-specific covariates (r^m, v^m, and K^m) for observation i. The p-value of the CK test is obtained by bootstrapping, following the procedure described in Andrews (1997).

search for at most three prices. Table 4 also illustrates that the group of consumers searching for the prices of between 4 and 15 firms is with percentages between 0 and 4 relatively small. About 8 percent of consumers search for prices thoroughly; they have search costs less than 43 dollar cents and search for the prices of more than 15 stores. Figures 1(a) and 1(b) show that the consumers can roughly be divided into three groups: buyers who do not search, buyers who search for at most three prices and buyers who search for many prices in the market.

Our findings are in line with several other empirical studies; Moraga-González and Wildenbeest (2008) report similar results using a different estimation method and dataset, while Wildenbeest (2011) finds that very few consumers visit an intermediate number of stores when searching for grocery products, even if quality differentiation is taken into account. Moreover, our result that consumers search very little is supported by the consumer-specific web browsing data for online bookstores used in De los Santos (2008) and De los Santos *et al.* (2011).

Table 4: Parameter estimates products

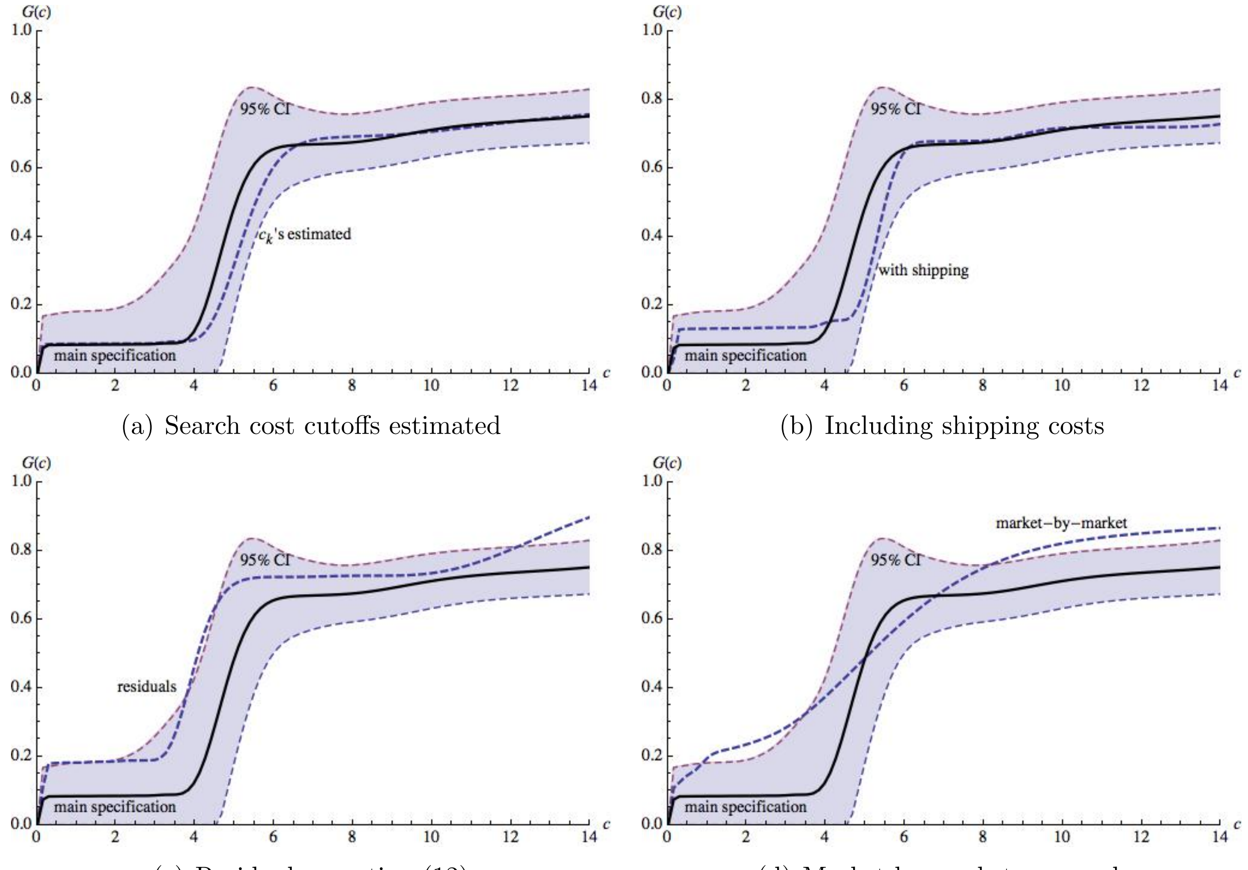
Part number	K	$\underline{p}$	v	r	μ_1	μ_2	μ_3	μ_4	$\mu_{5...15}$	$\mu_{16...K}$
KTT3311A	32	148.62	235.00	142.73	0.26	0.24	0.42	0.00	0.00	0.08
KTT533D2	33	100.45	161.40	93.93	0.32	0.59	0.00	0.00	0.00	0.08
KTD-INSP8200	39	148.62	249.54	138.75	0.29	0.62	0.00	0.00	0.00	0.08
KTD-INSP5150G	39	148.62	222.35	142.52	0.28	0.61	0.02	0.00	0.00	0.08
KTD-INSP6000	35	100.45	151.05	94.74	0.33	0.58	0.00	0.00	0.00	0.08
KTD-INSP6000A	38	94.99	154.50	87.40	0.33	0.58	0.00	0.00	0.01	0.08
KAC-MEME	24	101.92	161.64	96.28	0.26	0.61	0.00	0.00	0.01	0.07
KTD-INSP9100	33	148.62	249.54	139.11	0.26	0.51	0.14	0.00	0.04	0.04
KTM-TP3840	37	104.55	161.94	97.09	0.33	0.59	0.00	0.00	0.02	0.06
KTH-ZD8000A	41	100.45	154.50	93.30	0.34	0.58	0.00	0.00	0.02	0.07

Notes: K is the number of firms, $\underline{p}$ the lower bound and v the upper bound of the price distribution (in US$), r the firms' unit cost (in US$), and μ_k the share of consumers searching k times.

The fact that a significant proportion of consumers does not search for many prices confers substantial market power to the firms. Using the estimates of the SNP specification, we can retrieve the marginal cost r in each market, which is also reported in Table 4. Marginal costs range between 56 and 64 percent of the value of the product, while average price-cost margins range between 19 and 24 percent across markets.

We have estimated the model using the empirical price CDF in each market to calculate the search cost cutoffs. The main reason for doing so is the gain in computing time: this avoids having to solve the system of equations (8) in each function evaluation. Figure 2(a) shows that the search cost CDF when the c_k's are estimated (dashed curve) is not very different from the one obtained when using the empirical CDF to get the cutoffs (solid curve). The average absolute distance between the two CDFs, calculated as the average absolute difference between the search cost CDFs evaluated at 100 search cost values between zero and \$14, is only 0.022, which is in line with the eyeball comparison of the estimated curves.

The prices used for our estimations include neither shipping costs nor sales taxes. The main reason for leaving these out is that shipping costs and sales taxes depend on the state in which the consumer resides, which makes it difficult to compare total prices. However, for robustness purposes, we also estimate the model neglecting sales taxes but including shipping costs for residents of New York. The average absolute distance between the two CDFs is with 0.036 relatively small, which can also be seen from the comparison of the estimated search cost CDF (dashed) and the search cost CDF obtained when ignoring shipping costs (solid) in Figure 2(b).



(a) Search cost cutoffs estimated

(b) Including shipping costs

(c) Residuals equation (12)

(d) Market-by-market approach

Figure 2: Estimated search cost CDF alternative specifications

Although the memory chips themselves are completely homogeneous, the price differences across vendors for a given chip may be due to store differentiation. Consumers might prefer one shop over another on the basis of observable store characteristics like quality ratings, return policies, stock availability, order fulfillment, payment methods, etc. To see the impact of observable shop characteristics on prices, we regress prices on indicators that are readily available from the price comparison sites. More precisely, we estimate the following model:

$$PRICE_j = \beta_0 + \beta_1 \cdot RATING_j + \beta_2 \cdot DISCLOSE_j + \beta_3 \cdot STOCK_j + \beta_4 \cdot LOGO_j + \varepsilon_j, \quad (12)$$

where, for each product, $PRICE_j$ is the list price of store j, $RATING_j$ is an average of the ranking of store j on shopper.com and pricegrabber.com, $DISCLOSE_j$ is a dummy for whether shop j disclosed shipping cost on either shopper.com or pricegrabber.com, $STOCK_j$ is a dummy for whether shop j had the item in stock, and $LOGO_j$ is a dummy for whether shop j had its logo on either shopper.com or pricegrabber.com. We estimate this equation by OLS. The resulting R-squared values indicate that only between 3 and 27 percent of the total variation in prices can be attributed to observable differences in store characteristics.[16] Although this does not rule out that there are unobservable differences between stores (e.g., cost differences or branding), this does suggest that the observable characteristics cannot explain the vast majority of variation in prices and that something else must cause such variability. In spite of this, for robustness purposes, we also estimate the model using the residuals of the regression above. This is standard practice in many structural auction models (e.g. Haile *et al.*, 2003; Bajari *et al.*, 2006; An *et al.*, 2010); Wildenbeest (2011) shows that if stores obtain quality input factors in perfectly competitive markets, the quality production function exhibits constant returns to scale, and consumers have the same preferences towards quality, this procedure is theoretically correct (moreover, all our results on identification and consistency of the estimator hold for such a specification as well). As shown in Figure 2(c), estimated search costs are uniformly lower (average absolute distance between CDFs is 0.099). This result is intuitive: when taking store heterogeneity into account, the (gross) gains from searching will be lower, which means that in order to explain observed prices consumers should have lower search costs and search more than in the model without store heterogeneity.

[16] For all memory chips, all the OLS coefficient estimates were not significant except the coefficient for $LOGO_j$, which was positive and significant at a 5 percent level for the KTM-TP3840 and KTH-ZD8000A chips.

Finally, Figure 2(d) shows the estimated search cost CDF when estimating search costs market-by-market using the approach put forward by Moraga-González and Wildenbeest (2008). This estimate is obtained by fitting an SNP density function to the estimated search costs in each market and taking the average. The average absolute distance between the two CDFs is substantial: 0.107. As can be seen from the graph, while our SNP procedure predicts only 9 percent of consumers have search costs less than \$3.70, this would be 34 percent according to the market-by-market approach.

5 Conclusions

Since the seminal contribution of Stigler (1961), economists have dedicated a significant amount of effort to understand the nature of competition in markets where price information is not readily available to consumers. One of the lessons learnt is that consumer search models may lead to price dispersion, a prediction quite different from the 'law of one price' obtained from conventional economic theory. Another is that the particular direction of the effects of public policy measures such as the introduction of taxes or the dismantling of barriers to entry depends on the shape of the search cost distribution. These observations motivate the development of methods to estimate search costs to be used in the simulation of counterfactual scenarios. The estimation of consumer search costs is nowadays an important area of empirical research.

This paper has studied the non-parametric identification and estimation of the costs of simultaneous search in markets for homogeneous products. We have argued that in order to increase the precision of the estimate of the search cost distribution one needs to increase the number of estimated critical search cost cutoffs in all quantiles of the search cost CDF. We have shown this can be done by pooling price data from various markets with similar search technology but different valuations, firms' costs and numbers of competitors. To take advantage of the relationship between the distinct markets we have proposed a new method to estimate the search cost density function by a semi-nonparametric density estimator whose parameters maximize the joint likelihood corresponding to all the markets. The paper has also illustrated the potential of our method by applying it to a dataset of online prices for ten notebook memory chips. The estimates obtained suggest that the search cost density is essentially bimodal such that a large fraction of consumers searches for very few prices and a small fraction of consumers searches for a relatively large number of prices.

Along the way we have made several simplifying assumptions. One of the assumptions has been that, within a market, consumers have the same valuation. In future work, we would like to relax this assumption and study a framework where there is heterogeneity both in consumer valuations and search costs. One of the advantages of developing such a framework is that it would enable the econometrician to estimate the correlation between consumer valuations and search costs. Another simplifying assumption has been that firms have complete information about the costs of one another. Our model could be extended to a setting with private information about the marginal costs of production. Estimation of such a model would enable us to distinguish price dispersion due to marginal cost heterogeneity from price dispersion due to search costs. Finally, an important restriction of our model has been that we treat the different markets as completely separated. In more general settings, one would like to develop a model of product differentiation with search costs. We strongly believe the ideas developed in this paper can be applied to such markets. The first step is to develop a tractable model that incorporates strategic price dispersion together with product heterogeneity. This is work we will pursue in the future.

References

Acemoglu D, Shimer R. 2000. Wage and technology dispersion. *Review of Economic Studies* **67**: 585–607.

An Y, Hu Y, Shum M. 2010. Estimating first-price auction models with unknown number of bidders: a misclassification approach. *Journal of Econometrics* **157**: 328–341.

Anderson SP, Renault R. 2000. Consumer information and firm pricing: negative externalities from improved information. *International Economic Review* **41**: 721–742.

Andrews D. 1997. A conditional Kolmogorov test. *Econometrica* **65**: 1097–1128.

Bajari P, Houghton S, Tadelis, S. 2006. *Bidding for incomplete contracts: an empirical analysis.* NBER Working Paper 12051.

Berry ST, Haile PA. 2009. *Nonparametric identification of multinomial choice demand models with heterogeneous consumers.* NBER Working Paper 15276.

Brendstrup B, Paarsch HJ. 2006. Identification and estimation in sequential, asymmetric, English auctions. *Journal of Econometrics* **134**: 69–94.

Burdett K, Judd KL. 1983. Equilibrium price dispersion. *Econometrica* **51**: 955–969.

Burdett K, Mortensen DT. 1998. Wage differentials, employer size, and unemployment. *International Economic Review* **39**: 257–273.

Coppejans M, Gallant AR. 2002. Cross-validated SNP density estimates. *Journal of Econometrics* **110**: 27–65.

De los Santos B. 2008. *Consumer search on the Internet.* NET Institute Working Paper 08-15.

De los Santos B, Hortaçsu A, Wildenbeest MR. 2011. Testing models of consumer search using data on web browsing and purchasing behavior. *American Economic Review*, forthcoming.

Diamond PA. 1971. A model of price adjustment. *Journal of Economic Theory* **3**: 156–168.

Donald SG, Paarsch HJ. 1993. Piecewise pseudo-maximum likelihood estimation in empirical models of auctions. *International Economic Review* **34**: 121–148.

Fenton VM, Gallant AR. 1996. Qualitative and asymptotic performance of SNP density estimators. *Journal of Econometrics* **74**: 77–118.

Gallant AR, Nychka DW. 1987. Semi-nonparametric maximum likelihood estimation. *Econometrica* **55**: 363–390.

Haile PA, Hong H, Shum M. 2003. *Nonparametric tests for common values in first-price sealed-bid auctions.* NBER Working Paper 10105.

Hong H, Shum M. 2006. Using price distributions to estimate search costs. *RAND Journal of Economics* **37**: 257–275.

Hortaçsu A, Syverson C. 2004. Product differentiation, search costs, and competition in the mutual fund industry: a case study of S&P 500 index funds. *Quarterly Journal of Economics* **119**: 403–456.

Ichimura H, Thompson TS. 1998. Maximum likelihood estimation of a binary choice model with random coefficients of unknown distribution. *Journal of Econometrics* **86**: 269–295.

Janssen MCW, Moraga-González JL. 2004. Strategic pricing, consumer search and the number of firms. *Review of Economic Studies* **71**: 1089–1118.

van der Klaauw B, Koning RH. 2003. Testing the normality assumption in the sample selection model with an application to travel demand. *Journal of Business & Economic Statistics* **21**: 31–42.

Koning P, van den Berg GJ, Ridder G. 2000. Semi-nonparametric estimation of an equilibrium search model. *Oxford Bulletin of Economics and Statistics* **62**: 327–356.

Koopmans TC, Reiersøl O. 1950. The identification of structural characteristics. *Annals of Mathematical Statistics* **21**: 165–181.

Lach S. 2002. Existence and persistence of price dispersion: an empirical analysis. *Review of Economics and Statistics* **84**: 433–444.

Lewis MS. 2008. Price dispersion and competition with differentiated sellers. *Journal of Industrial Economics* **56**: 654–678.

Manning R, Morgan P. 1982. Search and consumer theory. *Review of Economic Studies* **49**: 203–216.

Matzkin RL. 1992. Nonparametric and distribution-free estimation of the binary choice and threshold crossing models. *Econometrica* **60**: 239–270.

Matzkin RL. 1993. Nonparametric identification and estimation of polychotomous choice models. *Journal of Econometrics* **58**: 137–168.

Moraga-González JL, Wildenbeest MR. 2008. Maximum likelihood estimation of search costs. *European Economic Review* **52**: 820–48.

Moraga-González JL, Sándor Z, Wildenbeest MR. 2010. *Nonsequential search equilibrium with search cost heterogeneity*. IESE Business School Working Paper 869.

Moraga-González JL, Sándor Z, Wildenbeest MR. 2012. *Supplement to 'Semi-nonparametric estimation of consumer search costs'*. http://www.kelley.iu.edu/mwildenb/msw-suppl.pdf

Morgan P, Manning R. 1985. Optimal search. *Econometrica* **53**: 923–944.

van Soest A, Das M, Gong X. 2002. A structural labour supply model with flexible preferences. *Journal of Econometrics* **107**: 345–374.

Stahl DO. 1989. Oligopolistic pricing with sequential consumer search. *American Economic Review* **79**: 700–712.

Stigler G. 1961. The economics of information. *Journal of Political Economy* **69**: 213–225.

Tirole, J. 1988. *The theory of industrial organization*. The MIT Press.

Varian HR. 1980. A model of sales. *American Economic Review* **70**: 651–659.

Wildenbeest MR. 2011. An empirical model of search with vertically differentiated products. *RAND Journal of Economics* **42**: 729–757.

Wolinsky A. 1984. Product differentiation with imperfect information. *Review of Economic Studies* **51**: 53–61.

Wolinsky A. 2005. Procurement via sequential search. *Journal of Political Economy* **113**: 785–810.

APPENDIX

Proof of Proposition 1. Consider the triplets of variables $(F, \{\mu_k\}_{k=1}^K, \{c_k\}_{k=1}^K)$ and $(F, \{\mu'_k\}_{k=1}^K, \{c'_k\}_{k=1}^K)$ that are generated by the quadruplets of variables (G, v, r, K) and (G', v', r', K'), respectively, where G' is another distribution function with support $(0, \infty)$ and positive density on this support. Then we prove the result by showing that $\mu'_k = \mu_k$, $c'_k = c_k$ and $G'(c_k) = G(c_k)$ for any $k \in \{1, 2, \ldots, K\}$.

First we note that neither μ_1 nor μ'_1 can be equal to zero. Indeed, if $\mu_1 = 0$ then by equation (4) $\sum_{k=2}^K k\mu_k (1 - F(p))^{k-1} = 0$ for any $p \in [\underline{p}, v]$, which, due to the fact that F is increasing and continuous, can only happen if $\mu_k = 0$ for any $k \geq 2$. This is in contradiction with $\sum_{k=1}^K \mu_k = 1$, so $\mu_1 > 0$. Since exactly the same arguments apply to μ'_1, we have shown that μ_1 and μ'_1 are strictly positive.

Next we prove that $r' = r$. Since $\mu_1 > 0$, equation (4) implies that F is strictly increasing on its support and hence invertible. By putting $p = F^{-1}(1 - z)$ in (4) for $\{\mu_k\}_{k=1}^K, r$ and $\{\mu'_k\}_{k=1}^K, r'$, we obtain that

$$\frac{\mu_1(v - r)}{\sum_{k=1}^K k\mu_k z^{k-1}} + r = \frac{\mu'_1(v - r')}{\sum_{k=1}^K k\mu'_k z^{k-1}} + r' \text{ for any } z \in [0, 1].$$

This implies that

$$\mu_1(v-r)\left(\sum_{k=1}^K k\mu'_k z^{k-1}\right) - \mu'_1(v-r')\left(\sum_{k=1}^K k\mu_k z^{k-1}\right) - (r' - r)\left(\sum_{k=1}^K k\mu_k z^{k-1}\right)\left(\sum_{k=1}^K k\mu'_k z^{k-1}\right) = 0$$

for any $z \in [0, 1]$. Since the LHS is a polynomial in z, all its coefficients must be equal to 0.

Suppose by contradiction that $r' \neq r$. This implies that

$$\text{either } \mu_K = \mu_{K-1} = ... = \mu_2 = 0 \text{ (so } \mu_1 = 1) \text{ or } \mu'_K = \mu'_{K-1} = ... = \mu'_2 = 0 \text{ (so } \mu'_1 = 1). \quad \text{(A13)}$$

Indeed, by contradiction assume that (A13) does not hold; then let $M, M' \geq 2$ denote the maxima of k, ℓ such that $\mu_k > 0$ and $\mu'_\ell > 0$. The coefficient of $z^{M+M'-2}$ is $-(r' - r) M\mu_M M'\mu'_{M'}$, which must be equal to 0, a contradiction with our assumptions. Therefore, (A13) must hold. In either case we have a contradiction because equation (5) implies that $\underline{p} = v$, which means that the price distribution F is degenerated. This establishes that $r' = r$.

Next we show that $\mu_k' = \mu_k$ for any k. From equation (4) we obtain

$$\sum_{k=1}^{K} k \frac{\mu_k}{\mu_1} (1 - F(p))^{k-1} = \frac{v-r}{p-r} = \sum_{k=1}^{K} k \frac{\mu_k'}{\mu_1'} (1 - F(p))^{k-1} \quad \text{for any } p \in \left[\underline{p}, v\right].$$

This is equivalent to

$$\sum_{k=2}^{K} k \left(\frac{\mu_k}{\mu_1} - \frac{\mu_k'}{\mu_1'} \right) z^{k-1} = 0 \quad \text{for any } z \in [0,1].$$

Since the LHS is a polynomial in z, all its coefficients must be equal to 0. Therefore, $\frac{\mu_k}{\mu_1} = \frac{\mu_k'}{\mu_1'}$ for $k = 2, \ldots, K$. On the other hand, $\mu_1 + \sum_{k \geq 2} \mu_k = \mu_1' + \sum_{k \geq 2} \mu_k' = 1$. These equalities together imply $\frac{1}{\mu_1} = \frac{1}{\mu_1'}$ and therefore $\mu_k' = \mu_k$ for any $k \geq 1$.

The equalities $c_k' = c_k$ follow from equation (2). It remains to show that $G'(c_k) = G(c_k)$ for any $k \geq 1$. We do so by showing that $\{G(c_k)\}_{k \geq 1}$ is uniquely determined by the series $\{\mu_k\}_{k \geq 1}$. By equations $(3a)$ and $(3b)$, $G(c_{k-1}) - G(c_k) = \mu_k$ for any $k \geq 1$. This implies that $G(c_k) = 1 - \sum_{h=1}^{k} \mu_h$ for any $k \geq 1$. The result then follows from the equality $\mu_k' = \mu_k$ for any $k \geq 1$ established above. ■

Proof of Proposition 2. In the proof we write $c_1(\theta)$ to make explicit the dependence of c_1 on $\theta \equiv v - r$. Note that due to the continuity of $c_1(\theta)$, $\sup_{\theta \in (0,\infty)} c_1(\theta) = \sup_m c_1^m$. Take an arbitrary interval $(a,b) \subset \left(0, \sup_{\theta \in (0,\infty)} c_1(\theta)\right)$. Then the pre-image set defined as $c_1^{-1}(a,b) = \{\theta : c_1(\theta) \in (a,b)\}$ is a nonempty set, open in $(0,\infty)$ because $\lim_{\theta \to 0+} c_1(\theta) = 0$ (by equation (4), if $\theta = 0$ then $\mu_k = 0$ for $k \geq 2$, so $\mu_1 = 1$ and thus $G(c_1) = 0$) and c_1 is, by assumption, a continuous function of θ. Therefore, with probability 1 there exists an m such that $\theta_m = v^m - r^m \in c_1^{-1}(a,b)$, which means that $c_1(\theta_m) \in (a,b)$.[17] Because the interval (a,b) has been chosen arbitrarily, we have proven that for any interval, we can find an m such that the corresponding cutoff point $c_1(\theta_m)$ is included in the interval with probability 1. Since $G(\theta_m) = G(c_1(v^m - r^m))$, $m \geq 1$, are identified, this establishes that in an arbitrary interval

[17] The argument for this statement is the following. Suppose that we have i.i.d. random variables $x_1, x_2, \ldots x_n$ drawn from a distribution with support $(0,\infty)$ and let $(c,d) \subset (0,\infty)$. Then the probability that at least one of these random variables is in (c,d) is equal to $1 - P(x_i \notin (c,d))^n = 1 - [1 - P(x_i \in (c,d))]^n$. Since $P(x_i \in (c,d)) > 0$, the above probability goes to 1 when $n \to \infty$. So when we have a countably infinite sequence of random variables, the probability that at least one of these random variables is in (c,d) is 1.

$(a, b) \subset \left(0, \sup_{\theta \in (0,\infty)} c_1(\theta)\right)$ we can find a point at which the search cost distribution is identified with probability 1. Therefore, since it is continuous, G is identified on $\left[0, \sup_{\theta \in (0,\infty)} c_1(\theta)\right]$. ■