# Learning Product Characteristics and Consumer Preferences from Search Data*

Luis Armona | Greg Lewis | Georgios Zervas

Harvard University | | Boston University

June 27, 2024

**Abstract**

A key idea in demand estimation is to model products as bundles of characteristics. In this paper, we offer an approach for jointly learning latent product characteristics and consumer preferences from search data, in order to predict demand more accurately. We combine data on consumers' web browsing histories and hotel price/quantity data to test this method in the hotel market. In two distinct applications, we show that closeness in latent characteristic space predicts competition, and parameters learned from search data substantially improve post-merger demand predictions.

**JEL Codes:** C13, C38, C51, C52, L1, L22, L81.

**Keywords:** E-Commerce; Search; Demand Estimation; Transfer Learning; Embeddings

*This research was started while Luis was an intern at Microsoft Research. We are grateful for comments from Dean Eckles, Liran Einav, Matt Gentzkow, the participants at the Marketplace Innovation Workshop, our reviewers from the 22nd ACM Conference on Economics and Computation, the two anonymous reviewers along with the Associate Editor at Marketing Science. We are also grateful to Duane Vinson and STR for providing the hotel data for this paper.

# 1 Introduction

Demand estimation is a widely used tool in empirical industrial organization and beyond, with applications to many industries, including transportation, healthcare and media. In many of these industries, the number of products is large, and modeling demand over all the products—as in the AIDS demand system (Deaton and Muellbauer 1980)—is impractical, because it requires estimating a substitution matrix that is quadratic in the number of products. Building on the ideas of Lancaster and McFadden (Lancaster 1966, McFadden 1973), a literature has emerged that assumes consumer preferences are over product characteristics rather than products. When the number of characteristics is much smaller than the number of products, this reduces data requirement. It also facilitates interpretation: products that are close in characteristic space are competitors because they vie for consumers who value the characteristics they share.

There are two important practical problems with this approach. The first is that in some markets—such as books (De los Santos and Wildenbeest 2017), movies (Einav 2007) and cereal (Nevo 2001)—the observable characteristics, such as genre, length or mushiness, are coarse. *"Ben-Hur"* and *"The Lord of the Rings: The Return of the King"* are both long movies, but they may not be competing for the same viewers. As a result, demand systems built from these observables are often unable to capture the relevant substitution patterns for choice. The second is that it is notoriously difficult to estimate consumer preferences accurately from market-level choice data alone. As Berry, Levinsohn, and Pakes (2004) and others have shown, the presence of second choice data helps considerably in learning about the distribution of consumer tastes.

In this paper, we propose a method that uses a set of revealed preference inequalities constructed from an auxiliary search dataset to improve demand estimation on a primary choice dataset. By search dataset, we mean records of options that consumers considered prior to making a choice. Search data reveals what products are considered together exists in many online markets (e.g., browsing data from an e-commerce platform), though nothing in our approach is limited to this case.[1] Given that search data often originates from a single search engine or spans different time periods, the set of participants may not fully represent market demand patterns. For this reason, we are interested in the case where the search data is used to inform demand estimation on a traditional market-level dataset of price and quantity. For example, in our application to the hotel market, we use search data from Internet Explorer (IE) users, and combine this dataset with a choice dataset from STR, a company whose data has high coverage of the entire hotel market.

Our method builds off the observation that consumers will generally only consider options that they believe they might choose i.e. products that, for them, offer high expected utility. It follows that products which are often searched together are either attractive to everyone (high utility products), or share some common characteristic that makes them attractive to a subset of consumers with high taste for that characteristic. This characteristic need not be observed to the econometrician; the search co-occurrence reveals it.

Reversing this logic, sets of consumers who all search a given product are likely to share a strong taste for some characteristic of that product. This too can be learned from the data, after accounting for the possibility of globally preferred products with fixed effects.

Formally, we use a set of revealed preference inequalities to learn these latent tastes and characteristics: any product searched by a consumer must have higher expected utility for that consumer than any product they did not search. To estimate these latent parameters, we apply the Bayesian Personalized Ranking (BPR) method (Rendle, Freudenthaler, Gantner, and Schmidt-Thieme 2009) to these inequalities. This method has the advantage of being computationally scalable to estimate millions of parameters on large datasets, in addition to being implementable using easily accessible software. We then map these parameters to a distinct population of consumers relevant for estimating market-level demand.

The inequalities we use can be micro-founded by a sequential search model (Honka, Seiler, and Ursu 2023). Consumers see some subset of product characteristics, and click on options

[1] For example, it could instead be companies the consumer called looking for service.

to learn more. Under some assumptions, the model implies that the products a consumer chooses to search on are the ones which are expected to offer the largest gains to utility. This in turn implies a number of pairwise inequalities: if options $A$ and $B$ were clicked, but $C$ and $D$ were not, the consumer expects more utility from $A$ or $B$ than $C$ or $D$, resulting in four inequalities. Compared to purchase data in discrete choice settings, where a single product purchase decision is observed, this substantially expands the number of consumer revealed-choice inequalities seen by the researcher. We then show this is consistent with a downstream model of discrete choice using the results of (Moraga-González, Sándor, and Wildenbeest 2023). Our proposed estimation method could also be used in other settings where the researcher has access to supplementary data on consumer preferences in the form of inequalities (e.g. from consumer surveys). The main contribution of the paper is showing that this set of inequalities makes learning latent characteristics and consumer preferences tractable, and these latent parameter improve the predictions of demand estimation.

To show this, we offer two applications. Our first application tests whether the characteristics learned from search data are most similar for products which are closer competitors. We estimate an event study design based on the entry of new hotels, and predict which hotels will lose market share post-entry on the basis of how close each hotel is to the new entrant, where the distance is either in latent, observable, or geographic space. We find that the latent model predicts better than a model based on standard observables (such as hotel class and amenities), but less well than one based purely on geographic distance. We interpret this result as showing that the search data allows estimation of latent characteristics that can predict substitution patterns, but if the observables are rich—e.g. geography plus class and amenities—the value added from estimating latent characteristics may be smaller. As noted earlier, there are markets (e.g. books, movies, cereal) in which observable data are not rich, and this method would be even more useful in those settings.

Our second application concerns merger analysis. We consider the 2016 merger of Marriott International and Starwood Hotels & Resorts Worldwide, which created the world's largest hotel company. After showing that this merger has substantial price effects, we apply our method to predicting out-of-sample post-merger market shares using only pre-merger data. Here we run a horse-race between the canonical demand model of BLP (Berry, Levinsohn, and Pakes 1995), which uses only observables and choice data, and an augmented BLP model with latent characteristics and/or latent consumer preferences learned from search data. We find that adding these parameters yields substantial improvements: while BLP fits better out-of-sample (in the sense of mean squared error) than a standard logit by 15%, when we include consumer preferences learned from search data, the corresponding improvement is 48% in our preferred specification. We also find that specifications using latent characteristics in place of observables yield improvements over the standard BLP model.

Our results suggest two ways in which the methods developed in the paper may be valuable in applications. First it may allow discovery of latent characteristic representations of markets with limited observables, which may facilitate discussions around market structure and the nature of competition. Second it may improve demand estimation by allowing consumer preferences to be learned (up to mean utility and scale parameters) from search data, where they are more plausibly identified.

Our paper relates to the literature on structural demand estimation, the literature on consumer search, and the literature on incorporating machine learning for demand estimation. Since Berry, Levinsohn, and Pakes (1995), structural demand estimation in economics has often relied on modeling consumer preferences over a "characteristic space", which is taken as given, to determine how close products are in a market for discrete choices. Berry, Levinsohn, and Pakes (2004) uses "second choice" data as supplemental micro-data that aids the identification of substitution patterns across cars in the auto market. In our framework, search data allows us to observe multiple products a consumer expects to yield high utility, enabling better identification of substitution patterns.

Methodologically, this paper is most similar in spirit to a strand of the marketing literature, beginning with Elrod (1988), that has concerned itself with estimating the structure of markets via factor models using revealed choices (Elrod 1988, Elrod and Keane 1995, Keane et al. 2013). These papers use observations of consumers in a panel structure to estimate latent characteristics of products, identifying the market structure from consumers who switch from one product to another. Our identification strategy follows a similar procedure, instead relying on products that are searched together within a single purchase decision to identify products as close substitutes.

Our paper also relates to a more recent strand of the marketing and economics literature using search data to identify substitution patterns over the product characteristic space, with a focus on explicitly modeling the choice set formation process. Kim, Albuquerque, and Bronnenberg (2011) use aggregate search and sales data to estimate the latent characteristics of products, and consumer preferences over these characteristics. Koulayev (2014) uses clickstream data in a similar empirical setting to our own, along with revealed preference inequalities implied by sequential search, to estimate search costs in the presence of unobserved consumer heterogeneity. Amano, Rhodes, and Seiler (2019) uses the sparsity of choice sets in a market with many products, to feasibly estimate consumer demand and substitution patterns in online settings. Our paper similarly uses search data to better explain consumer preferences at the purchase decision stage in a market with a large number of products, though our applications do not model the search process explicitly.

The search model used for our estimation is built on the literature using search models (Honka, Seiler, and Ursu 2023) to explain consumer purchase decisions. These papers rely on an assumption of either simultaneous (Honka 2014) or sequential (Moraga-González, Sándor, and Wildenbeest 2023) search to quantify the role of search in consumer welfare and substitution patterns. Our paper is less ambitious than many of these papers, in the sense that we choose to estimate latent parameters in a prior step to demand estimation, rather than jointly estimating a model of search and choice. This makes our estimation procedure more tractable and versatile across empirical settings, where search and choice data may come from different datasets, but also unable to estimate richer counterfactuals, such as altering the search environment and evaluating its implications for downstream demand.

Our paper also relates to a new literature that seeks to employ machine learning methods to augment demand estimation (Bajari, Nekipelov, Ryan, and Yang 2015). Most closely related to our own paper in this literature are those papers that attempt to directly estimate the characteristic space of products using *embeddings* methods popularized in the machine learning literature for representing a large number of products in a common domain (Salakhutdinov and Mnih 2008, Koren, Bell, and Volinsky 2009, Johnson 2014, Rudolph, Ruiz, Mandt, and Blei 2016). A number of recent papers use these methods to estimate latent characteristics, latent preferences, or both in economic settings (Ruiz, Athey, and Blei 2017, Athey, Blei, Donnelly, Ruiz, and Schmidt 2018, Sams 2019, Donnelly, Kanodia, and Morozov 2020, Magnolfi, McClure, and Sorensen 2022). Our paper similarly uses established machine learning methods, specifically the Bayesian Personalized Ranking method of Rendle, Freudenthaler, Gantner, and Schmidt-Thieme (2009), to estimate the latent characteristics and preferences over hotels from search data.

# 2 Data

We draw on 3 datasets to estimate latent parameters and conduct our empirical analysis.

**Hotel Demand and Price Data.** Our first dataset, which we obtained from STR, spans January 2001 to March 2019 and contains monthly financial performance data for 5,358 hotels in 5 western U.S. States (Arizona, California, Nevada, Oregon, and Washington). Specifically, for each month-year, STR records an anonymized ID uniquely associated with each hotel, the total number of rooms sold, and the total revenue received by the hotel from room sales. From this, we infer the average price per room sold, also known as the average daily rate (ADR) in the hotel industry. We deflate the nominal ADR recorded each month to real March 2019 U.S. dollars using the CPI for All Urban Consumers time series. Reporting of financial performance data is voluntary on the part of hotels, however coverage is quite high. Of the universe of hotels in these states, only 10% of hotels reported no data for the entire sample, and we have data for 83% of all hotel-by-month-years.

**Census of Hotel Characteristics.** The second dataset we use from STR links anonymous hotel IDs to certain observable hotel characteristics. This dataset contains the market, according to STR definitions, that each hotel competes in, the sub-market a hotel is located (e.g. Hollywood/Beverly Hills within the Los Angeles market), an anonymized ID to link to the transaction data, the total meeting space in square feet, the month-year the hotel was first opened/closed (where applicable), as well as categorical variables representing the size (total capacity in rooms) of each hotel, the location type of each hotel (near airport, urban, suburban, near highway, a resort area, or a small metro / town), the operation structure of the hotel (franchise, owned by chain, or an independently owned/branded hotel), and the price segment the hotel belongs to, which is a categorization based on their typical ADR (higher segment meaning more expensive hotels). This dataset also includes anonymized IDs describing the affiliation brand of each hotel, the company owning the hotel, the management company managing day-to-day operations of the hotel, and the parent company of hotel. For example, although we never observe the hotel, affiliation, or company name in the data, a Ritz Carlton hotel in the dataset would contain a numeric affiliation ID corresponding to all Ritz Carlton hotels in the dataset, and a numeric parent company ID corresponding to Marriott International, Inc., the parent corporation that owns the Ritz Carlton brand. The dataset also contains the zipcode associated with each anonymized hotel ID, which we map to the average latitude-longitude within each zipcode, using a cross-walk originally based on 2013 U.S. census data.[2] We use latitude and longitude to capture the spatial component of product differentiation in the hotel industry. Affiliation IDs are provided at the annual level, to account for mergers and acquisitions, while all other characteristics are measured in 2019, and do not change over time.

**Internet Explorer Click-Stream Data.** Finally, our third dataset consists of the

[2]Source: `https://gist.github.com/erichurst/7882666`

web browsing histories of a sample of 29,936 Internet Explorer (IE) users from June 1st, 2014 to May 31st, 2015 visiting the website `expedia.com`. We make use of a session ID variable recorded by IE that captures the URLs visited by a single user during one continuous "session" defined as continuous usage of their computer where the time between clicks is no more than 30 minutes. The URLs in the click-stream data, contain the Expedia ID of each hotel, the check-in date selected by the user, the price displayed to the user for each hotel that appears on screen, as well as whether the user clicked on a particular hotel in `expedia.com`. We classify a click on a hotel's Expedia page that is displayed to a user on their web browser as constituting a "search" of the hotel by a user. This allows us to group together hotels that were jointly considered within each session. We use this search data to estimate consumer heterogeneity and latent characteristics of hotels.

Tables 1 and 2 displays the summary statistics describing the observable characteristics of the 4,218 hotels that appear in both datasets. While the median consumer searches one hotel per search session, there is a substantial mass of consumers (44%) that search multiple hotels. These consumers provide information on hotel substitutability, which we will use to identify the latent parameters governing demand from search data. For our search data, we remove consumers who search more than 35 hotels during a single session, which are likely web-scrapers and not actual consumers. We retain those visitors to the website who searched at least one of the 4,218 hotels included in both datasets. Our final sample of searchers consists of 18,492 unique visitors who engage in 23,986 unique search sessions. Table 3 contains summary statistics on the search sessions in the Expedia dataset.

# 3 Methodology

## 3.1 Demand Estimation and Pitfalls

A standard model for consumer demand in discrete choice settings, popularized by Berry, Levinsohn, and Pakes (1995), and known as "BLP", is the mixed logit choice model. One of

| | Mean | Std. Dev | Minimum | 25th Pctile | Median | 75th Pctile | Maximum |
|---|---|---|---|---|---|---|---|
| **Transaction Data** | | | | | | | |
| Monthly Price ($) | 129.34 | 81.95 | 18.06 | 80.91 | 110.66 | 151.47 | 1,576.09 |
| Monthly Hotel Occupancy | 2,951.06 | 3,381.11 | 7 | 1,224 | 2,101 | 3,360 | 77,305 |
| **Continuous Variables** | | | | | | | |
| Latitude | 37.50 | 4.77 | 31.39 | 33.81 | 36.13 | 38.65 | 48.95 |
| Longitude | -118.83 | 3.66 | -124.19 | -122.14 | -119.02 | -117.21 | -108.95 |
| Meeting Space (Sq. Ft.) | 3,981.22 | 11,326.00 | 0 | 0 | 560 | 2,500 | 220,000 |
| Year Hotel Opened | 1984.22 | 27.24 | 1798 | 1977 | 1989 | 1999 | 2014 |
| # of Hotels | 4,218 | | | | | | |
| # of Geographical Markets | 21 | | | | | | |
| # of Month-Years | 87 | | | | | | |

Table 1: Summary Statistics: STR Hotel Dataset (Continuous Data)

| **Ownership Structure** | | **Size Category** | | **Price Segment** | | **Location Type** | |
|---|---|---|---|---|---|---|---|
| Value | Mean | Value | Mean | Value | Mean | Value | Mean |
| Chain Management | 0.16 | Less Than 75 Rooms | 0.32 | Economy Class | 0.24 | Airport | 0.08 |
| Franchise | 0.70 | 75 - 149 Rooms | 0.43 | Midscale Class | 0.16 | Interstate | 0.07 |
| Independent | 0.13 | 150 - 299 Rooms | 0.18 | Upper Midscale Class | 0.25 | Resort | 0.10 |
| | | 300 - 500 Rooms | 0.05 | Upscale Class | 0.17 | Small Metro/Town | 0.16 |
| | | Greater Than 500 Rooms | 0.02 | Upper Upscale Class | 0.11 | Suburban | 0.45 |
| | | | | Luxury Class | 0.06 | Urban | 0.14 |

Table 2: Summary Statistics: STR Hotel Dataset (Categorical Variables)

| | Mean | Std. Dev | Minimum | 25th Pctile | Median | 75th Pctile | Maximum |
|---|---|---|---|---|---|---|---|
| Price of Hotels on Platform ($) | 118.02 | 77.86 | 14.00 | 72.49 | 100.11 | 138.37 | 19,374.00 |
| Price of Searched Hotels | 207.88 | 253.92 | 16.95 | 99.99 | 145.00 | 221.88 | 19,374.00 |
| # of Hotels Searched per Session | 1.85 | 1.80 | 1 | 1 | 1 | 2 | 35 |
| Pr(Hotel Stay Purchased) | 0.05 | 0.21 | 0 | 0 | 0 | 0 | 1 |
| # of Consumers | 18,492 | | | | | | |
| # of Search Sessions | 23,986 | | | | | | |

Table 3: Summary Statistics: Expedia Search Dataset Hotel Dataset

the key features of this model is that preferences are expressed over characteristics observed to the researcher and associated with each product. The second key feature of this model is consumer heterogeneity. Typically, researchers allow for random unobserved heterogeneity following some parametric distribution (e.g. a normal distribution) in preferences along observed product characteristics. By specifying preferences over characteristics associated with a product, rather than products directly, the model is able to capture rich substitution patterns in a tractable manner for demand estimation.

Three cases stand out as situations where the above model may be limiting in its ability to capture substitution patterns. First, the econometrician may lack access to data on characteristics which consumers have preferences over. Given this scenario, the random preferences over observed characteristics provide very little bite in capturing substitution patterns. Second, the characteristics over which consumers have preferences may be high dimensional. For example, in the hotel market, hotels are differentiated by the local amenities they are close to, which is difficult to summarize with a low-dimensional collection of observable characteristics. Similarly, in the market for books, textual descriptions may provide rich measures of product differentiation, but are not easily summarized. This will make it infeasible to estimate substitution patterns over these characteristics, unless the econometrician has access to market-level data over a proportionally large number of markets. Third, the parametric assumptions on the distribution of random coefficients may be restrictive and unable to capture taste heterogeneity across consumers. Allowing even random, parametric unobserved tastes for product characteristics to correlate with each other can make the preference specification in mixed logit models very rich in capturing underlying consumer demand. In practice, however, it is often difficult to identify the correlation structure in unobserved preferences using only market level price and quantity data, so researchers typically assume unobserved tastes for product characteristics are independent.

This paper seeks to address the above limitations by augmenting the mixed logit demand model with latent parameters learned from search data. First, suppose in addition to observed characteristics there are a set of low dimensional latent characteristics associated with each product. These latent characteristics allow consumers in the same market to have a heterogeneous match value with product $j$ that is not constrained to depend on observable product characteristics. These latent characteristics address the first two limitations of the BLP model.

Second, suppose that instead of following a parametric probability distribution, consumer preferences are non-parametrically estimated in the search data, and are allowed to follow the same distribution (up to scale) in the choice data.. This additional flexibility may allow for non-linearities in preference heterogeneity, along with correlated preferences along particular dimensions of the characteristic space. For example, consumers may have high preferences for hotels that are close to the city center only when they also have high preferences for hotels near bars, but when the consumer is interested in visiting museums, the correlation between preferences for nearby bars and the city center is eliminated.

With only market-level price and quantity data, the estimation of latent characteristics and a non-parametric distribution of consumer heterogeneity is largely hopeless. A natural reason this is challenging in a discrete choice setting is that consumers select at most one good for purchase when they arrive to the market, which limits our ability to observe what the consumer would have chosen in the absence of the purchased good. To estimate unobserved characteristics with market-level data, along with heterogeneous preferences, would require us to observe systemic correlations in demand between hotels that are not similar on observable characteristics, but still explainable by exogenous factors such as entries, exits, and mergers. We would require, at the very least, data on preferences a consumer has over other products besides those they purchase.

In order to address the estimation limitations of this proposed, more flexible demand model, we make use of individual-level search data to supplement the traditional market-level price and quantity data used to estimate discrete choice models. With the advent of online commerce, search data is increasingly available to researchers interested in studying consumer behavior. The primary advantage of search micro-data is it allows us to observe substitution patterns *within* an individual consumer and purchase decision. In this way, we view our usage of search data for demand estimation as an extension of the incorporation of "second choice" data in (Berry, Levinsohn, and Pakes 2004) and micromoments used in (Petrin 2002). The intuition for our approach is that learning all available information for each product in a market is costly, particularly in an online setting where there can be thousands of products. Thus, consumers must form choice sets of a subset of all goods in the market before making a purchase decision. Much like discrete choice demand models use a "revealed preference" approach to infer purchased products yield high utility, we use the choice sets formed by consumers as a revealed preference signal that consumers expect products they search to yield high utility. The key difference is that we can observe multiple products entering a consumer's choice set, whereas purchase decisions in discrete choice settings are limited to one good per person / purchase decision. For example, we may observe 2 hotels are systemically searched together that do not share any observable characteristics, which would suggest they are similar along a latent characteristic.

One advantage of our approach is that we do not require purchase decisions to appear in the search microdata in order to estimate demand. That is, we use the search data to estimate latent characteristics and consumer preferences in a prior estimation step, without any information on purchases. This summarizes the information learned from search data into a set of latent parameters. Then, assuming an affine mapping from the search and market population of consumers, researchers can transfer these latent parameters into a standard demand framework. This distinction is important because search data, while widely available, is typically from one specific platform. Even when there is selection into use of that platform—i.e. the population is not fully representative—co-occurrence in search may still help in estimating the demand model for the full market. Then, using aggregate quantity and price data, we can still answer questions of interest that affect an entire market, not just a single platform, such as the impact of a merger on welfare or alternative policies set by a social planner.

## 3.2 A Model of Search and Choice

In this section, we describe how we use search data to construct additional revealed preference inequalities to help identify both consumer heterogeneity and unobserved product characteristics relevant for consumer demand. We do so through a micro-founded model of sequential search, which we then show is internally consistent with a second choice stage demand model one can estimate using market-level data. While we could in principle jointly estimate this two-stage choice model, we treat the estimation of latent demand parameters from search data as a standalone method, since our use case is to augment a traditional demand system using the rich search microdata that informs us about consumer preferences.

**Preferences** Assume that consumers have the following utility over $J$ products:

$$u_{i,j,t} = \delta_{j,t} - \alpha_i p_{j,t} + \vec{\beta}_i \vec{X}_j + \epsilon_{i,j,t} \tag{1}$$

$$\delta_{j,t} = \delta_j + \beta X_{j,t} + \xi_{j,t} \tag{2}$$

$$\vec{X}_{j,t} = [\vec{X}^o_{j,t}, \vec{\gamma}_j] \tag{3}$$

$$[\alpha_i, \vec{\beta}_i] = \vec{v}_i, \quad \vec{v}_i \sim V \tag{4}$$

$$\epsilon_{i,j,t} = \zeta_{i,j,t} + (1 - \sigma_\zeta)\tilde{\epsilon}_{i,j,t} \tag{5}$$

$\delta_{j,t}$ is the mean utility of product $j$ in period $t$, composed of homogenous preferences for product characteristics $X_{j,t}$ ($\beta$), and a vertical demand shock ($\xi_{j,t}$). $\alpha_i$ and $\vec{\beta}_i$ denote the consumer's heterogeneous preferences over price and product characteristics, respectively. $\epsilon_{i,j,t}$ denotes an idiosyncratic match value, decomposed into a component known to consumers prior to search, $\zeta_{i,j,t}$, and a component revealed after search, $\tilde{\epsilon}_{i,j,t}$. Product characteristics $\vec{X}_{j,t}$ are a $(K + L) \times 1$ vector of product characteristics, $K$ of which are observable to the researcher ($\vec{X}^o_{j,t}$), and $L$ of which are unobservable to the researcher ($\vec{\gamma}_j$). $V \in \Delta\mathbb{R}^{L+K+1}$ is the distribution of consumer preferences $\vec{v}_i$ over prices and product characteristics. We normalize the utility of the outside option $u_{i,0}$, involving no purchase, to be zero. We assume the econometrician is able to observe both $\vec{X}^o_{j,t}$ and $p_{j,t}$, but unable to observe the remaining parameters governing preferences.

In terms of the consumer's information structure, we assume that $\tilde{\epsilon}_{i,j,t}$ is the only unknown component of utility. Both the cumulative match value $\epsilon_{i,j,t}$ and the post-search component $\tilde{\epsilon}_{i,j,t}$ are distributed i.i.d. according to an extreme value type 1 (Gumbel) distribution with normalized mean of zero 0 and scale parameter 1, while $\zeta_{i,j,t}$, the pre-search idiosyncratic shock, is distributed i.i.d. according to the conjugate Gumbel distribution derived in Cardell

(1997) with scale parameter $\sigma_\zeta$.[3] As a result, expected utility is as follows:

$$\bar{\delta}_{i,j,t} \equiv E[u_{i,j,t}] = \delta_{j,t} - \alpha_i p_{j,t} + \vec{\beta}_i \vec{X}_j + \zeta_{i,j,t} \tag{6}$$

**Search Decision** We assume that consumers engage in sequential search across products (Weitzman 1978, Honka, Seiler, and Ursu 2023). Search is costly, with the cost for consumer $i$ to search product $j$ being $c_{i,j,t}$. Weitzman (1978) characterizes the optimal search behavior in this environment. Each product is associated with a reservation value $r_{ijt}$ which solves the equation:

$$c_{i,j,t} = \int_{r_{i,j,t}}^{\infty} (z - r_{i,j,t}) dF(z|\bar{\delta}_{i,j,t}) \tag{7}$$

This reservation value is the utility that would make a consumer indifferent between searching product $j$ — and potentially finding a higher utility, but having to pay the search cost — and stopping search now. The distribution $F$ is type 1 extreme-value (TIEV) with location $(1-\sigma_\zeta)\bar{\delta}_{i,j,t}$ and scale $(1-\sigma_\zeta)$. Consumers will optimally search products in descending order of reservation value, stopping if at any point the utility offered by the best option they have found so far exceeds the reservation of the next object to be searched (which, by definition of the reservation value, is then not worth searching).

An implication of the model is that any product that was searched must have had a higher reservation value than any product that was not searched:

$$r_{i,j,t} \geq r_{i,k,t}, \quad \forall j \in J_{i,s}, \quad \forall k \notin J_{i,s} \tag{8}$$

Since the distributions of utility form a location family, a product that was searched either had a lower search cost than one that was not searched, offered higher expected utility, or both. Notice that in comparison to choice data, which offers the same logic (the product

[3] This is analogous to the structure of the nested logit model, where consumers choose a group of products followed by a product within the group, except that the idiosyncratic terms $\tilde{\epsilon}_{i,j,t}$ and $\zeta_{i,j,t}$ apply to different stages of search and choice.

chosen offered either higher utility or was easier to find), search data serves as a means to "expand" the set of revealed preference inequalities observed in the data (when consumers search more than one product). In particular, we use the logic of the search model to imply that all searched products $j$ have reservation values than all unsearched products $k$. By observing a larger set of inequalities the problem of identifying preferences as well as latent characteristics becomes more feasible. We exploit this in our estimation of consumer preferences and unobserved characteristics.

**Choice Decision** Following Moraga-González, Sándor, and Wildenbeest (2023) we assume search costs $c_{i,j,t}$ are distributed i.i.d. across consumers and products according to the following distribution:

$$S(c) = Pr(c_{i,j,t} \leq c) = \frac{1 - \exp(-\exp(-H_0^{-1}(c) - \mu))}{1 - \exp(-\exp(-H_0^{-1}(c)))} \tag{9}$$

where we define $H_0(r) = \int_r^{\infty}(z-r)dF_0(z)$ for $F_0$ a type 1 extreme value distribution with normalized mean of 0 and scale parameter 1, and $\mu > 0$ a location shifter of the search cost distribution.[4] As shown in that paper, if the component of utility that is unknown to consumers is distributed extreme value type 1 , then integrating over consumer heterogeneity $V$, the search cost distribution $S$, and the idiosyncratic match value $\epsilon_{i,j,t}$ yields the following purchase probabilities/market shares:

$$s_{j,t} \equiv Pr(j \text{ is purchased}) = \int_{\beta_i,\alpha_i} \frac{\exp(\delta_{j,t} - \alpha_i p_{j,t} + \vec{\beta}_i \vec{X}_j - \mu)}{1 + \sum_{k \in J_t} \exp(\delta_{k,t} - \alpha_i p_{k,t} + \vec{\beta}_i \vec{X}_k - \mu)} dV(i) \tag{10}$$

That is, these are the mixed traditional logit choice probabilities, up to a location shifter $\mu$ representing the costly search. These conditions hold in our setting - Appendix C provides the derivation.

The model presented here for recovering latent product characteristics and consumer

[4] $H_0$ is strictly decreasing in $r$ and therefore invertible.

preferences from search data serves as a vehicle to better estimate demand. This search model is not without limitations. We discuss two of the most noteworthy ones below, and how a researcher might address them:

**Product Rankings.** We assume here that search costs are idiosyncratically distributed across consumers and products. This may be violated in practice because of the important role that search rankings play on online platforms (Ursu 2018). Higher ranked products are much more prominent, easier to find, and more often clicked. Moreover, because of the large number of products in our market, it is likely that some products are not even viewed by consumers before clicking if they are ranked too low, which could lead to some products not being searched simply because the consumer is unaware of these choices. We could accommodate this situation by augmenting the location of search costs to be a function of the product's ranking $R_j$, e.g. specifying a location parameter $\mu_j = \log(1 + \exp(\mu_0 + \mu_R R_j))$, as recommended in Moraga-González, Sándor, and Wildenbeest (2023). Since we do not observe search rankings in our search data, we have not chosen to go this route, which may lead to some bias in estimation. We expect that these rankings would be most important in impacting the hotel fixed effect we estimate during the search stage, which is not used in later analysis.

**New Products.** Because we estimate the latent characteristic vector $\vec{\gamma}_j$ based on consumer search behavior, we are unable to use this model to characterize market outcomes if a new product became available. This is a well-known issue with embeddings methods known as the "cold-start problem" (Schein, Popescul, Ungar, and Pennock 2002). In Section 4, we show that, at least in this market, our estimated latent characteristics can predict observable characteristics with reasonable accuracy. This suggests that if necessary, researchers may be able to predict the latent characteristics of new products with success using a set of observed characteristics, particularly if the researcher has access to high-dimensional observable data, such as textual product descriptions. One could follow the approach of Cortes (2018) to estimate these latent characteristics for new products.

## 3.3 Estimation

### 3.3.1 Search Model

Given an observed choice set of size $|J_{i,s}|$, without information on search order, the sequential search model of Section 3.2 implies $|J_{i,s}| \cdot (J - |J_{i,s}|)$ inequalities of the kind in equation 8 for each consumer $i$, comparing the reservation value of searched products with unsearched products.[5] A natural way to proceed would be to choose parameters that maximize the likelihood of the observed inequalities; that is to maximize the probability of events of the form $\Pr(r_{ijt} \geq r_{ikt})$ for $j$ searched and $k$ not searched. In Appendix D we build on Moraga-González, Sándor, and Wildenbeest (2023) and show this is equivalent to maximizing:

$$Pr\left(-\alpha_i \Delta p_{j,k,t} + \vec{\beta}_i \Delta \vec{X}_{j,k,t} + \Delta\delta_{j,k,t} > -\Delta\zeta_{i,j,k,t} - (1-\sigma_\zeta)(H_0^{-1}(c_{i,j,t}) - H_0^{-1}(c_{i,k,t}))\right) \quad (11)$$

Where $\Delta x_{j,k,t} = x_{j,t} - x_{k,t}$. The left hand side of the inequality is the difference in expected utilities offered by products $j$ and $k$ *excluding* the pre-search shock $\zeta$, and the right hand side is the sum of the difference in pre-search and search cost components of reservation values.

This is a computationally unwieldy loss function, due to the presence of the term $H_0^{-1}(\cdot)$, which is an inverse of a non-linear function. Since search costs are i.i.d. across products, this term acts mostly to change the shape of the likelihood function. If we were to ignore the term related to search costs, the likelihood of Equation 11 would reduce to:

$$Pr\left(E[u_{i,j,t}] > E[u_{i,k,t}]\right) = \sigma\Big((-\alpha_i(p_{j,t} - p_{k,t}) + \vec{\beta}_i \cdot (\vec{X}_j - \vec{X}_k) + (\delta_{j,t} - \delta_{k,t}))/\sigma_\zeta\Big) \quad (12)$$

[5] This number of inequalities is large in magnitude, particularly compared to those from a purchase decision, which only reveals that the chosen product is preferred to $J-1$ other available products when we abstract from choice set formation. On average, each consumer in our sample has $10,267$ inequalities implied by their search patterns, compared to the 4,192 inequalities on average that would be implied by single purchase decisions. In particular, the *increased* number of revealed preference inequalities revealed from search data relative to purchase data is $(|J_i| - 1) \times (J_t - |J_i| - 1)$. When $J$ is large and $|J_i|$ is even moderately sized, this can result in significantly more data on consumer preferences.

where $\sigma$ denotes the sigmoid, or logit function, $\sigma(x) = \frac{1}{1+\exp(-x)}$, using the fact that the difference of the two TIEV shocks $\zeta$ is logistic with scale $\sigma_\zeta$ (Cardell 1997). It is identical to the posterior likelihood formed in Rendle, Freudenthaler, Gantner, and Schmidt-Thieme (2009) for the Bayesian Personalized Ranking model, a machine learning model used for large-scale embedding problems (under priors that we detail below). Because we want to apply the scalable machine learning techniques from that paper, which are Bayesian, we maximize this posterior likelihood rather than applying maximum likelihood estimation to the likelihood of Equation 11. We are therefore misspecified for the true likelihood, and this is best thought of as a quasi-likelihood estimation procedure. In Appendix D we show via simulations that the quasi-likelihood and likelihood are very similar, as the difference in (true) reservation values closely follows a logistic distribution with a higher scale attributable to the additional variation from $H_0^{-1}(c_{i,j,t})$. This is equivalent to an increased value of $\sigma_\zeta$, which we do not estimate (so that all parameters are learned up to a scaling factor.) Since the objective functions are very similar, assuming we correctly specify utility, we expect the quasi-MLE estimates will be similar to the MLE estimates. Still, we may not consistently or efficiently extract latent product characteristics and preferences from the search data using this approach. We are willing to make this statistical compromise in order to gain access to well established machine learning approaches, because we know that ultimately the learned features will be "plugged-in" into a completely separate demand estimation step, which uses GMM and has much stronger statistical guarantees. Insofar as this estimation is misspecified, our latent parameters will be less informative. Nonetheless, we show later in the paper that these estimates are useful for predicting demand after a significant change to market structure.

Since our search data is a single snapshot over the course of 12 months, we assume that $\delta_{j,t}$ is constant over time, e.g. $\delta_{j,t} = \delta_j$, and estimate a single "hotel fixed effect" for each product.[6] Recall that we make no distributional assumptions on the preferences $\vec{\beta}_i$ or $\alpha_i$. As

[6]Since our characteristics are time-invariant, this is equivalent to assuming the unobservable (to the researcher) quality $\xi_{j,t}$ is constant within product during our sample period. In practice, when we estimate

a result, we will treat them as individual-specific preference parameters to be learned, based on the search data we observe on each individual. This will give us an empirical estimate of the distribution of consumer preferences $V$.

Thus the parameters of the search model consist of the individual consumer preferences, $\alpha_i, \vec{\beta}_i$, the unobserved characteristics of products, $\vec{\gamma}_j$, and the mean utilities, $\delta_j$. In some cases, we may observe a single consumer engage in multiple search sessions, which we treat as independent purchase decisions (but with common preference parameters across sessions). Let $S_i$ denote the set of search sessions we observe for a single consumer $i$, and $J_{i,s}$ denote the choice set (clicked products) formed by consumer $i$ during search session $s$. Given observed choice sets $J_{s,i}$ , the objective we use to estimate the parameters of the model, $\Theta = \{\alpha_i, \beta_i, \gamma_j, \delta_j\}$ is as follows:

$$\log\Big(Pr(\Theta|\{J_{s,i}\})\Big) - \lambda_\Theta ||\Theta||^2 = \sum_i \sum_{s\in S_i} \sum_{j\in J_{i,s}} \sum_{k\notin J_{i,s}} \log\Big(Pr(E[u_{i,j,t}] \geq E[u_{i,k,t}]|\Theta)\Big) - \lambda_\Theta \sum_{k=1}^{N_\Theta} \theta_k^2 \tag{13}$$

where $\lambda_\theta$ denotes an L2 regularization hyperparameter to prevent overfitting, and $\theta_k$ denotes the $k$th element of the vector $\Theta$. This is consistent with the Bayesian interpretation of the model, where we place a normal prior on the parameters $\Theta$ with mean 0 and variance $\lambda_\Theta^{-1}$. It is identical to the posterior likelihood formed in Rendle, Freudenthaler, Gantner, and Schmidt-Thieme (2009) for the Bayesian Personalized Ranking model, a machine learning model used for large-scale embedding problems. This allows us to rely on tools from the machine learning literature to obtain estimates in our high-dimensional parameter space.

We make use of all the inequalities available for each consumer. In our application, the hotel market, this will mean learning from the fact that a consumer searching for a hotel in San Diego does not click on any hotels in New York City. While this decision not to restrict the comparisons — for example, to hotels within a market — may seem odd, it is exactly these comparisons that allows the model to learn that hotels in San Diego should be far from

demand in Section 6, we re-estimate the $\delta_{j,t}$ based on aggregate market shares.

hotels in New York City in the latent space.

**Identification.** The search model estimation amounts to simultaneously learning consumer preferences ($\{\alpha_i, \vec{\beta}_i\}$) and latent product characteristics ($\gamma_j, \delta_j$). Due to the scale of the pre-search shock ($\sigma_\zeta$) being unknown, we cannot identify the scale of the parameters in utility space. We estimate the scale in our choice model. The intuition for why these parameters are identified (up to scale) is as follows. Pairs of products that are often searched together must have similar locations in product space, otherwise they wouldn't both be utility maximizing for a consumer with preferences $\beta_i$.[7] Similarly, consumers who all search some product $j$ must have similar preferences. And higher order relationships provide more information: if a pair of consumers both search a particular pair of products, this is further evidence that both consumers and products are close in their respective spaces. Since prices may be endogeneously correlated with unobservables, the location of our price preferences $\alpha_i$ will be biased. However, as long as suppliers do not price discriminate across consumers, the distribution of preferences $\alpha_i$ across consumers will be properly identified. When we estimate our demand system, we calibrate the mean price coefficient using a prior study from the same dataset (Lewis and Zervas 2016) to overcome this endogeneity concern.

**Specifications.** To test how well this approach works, we consider specifications in which we have only observable characteristics (and learn consumer preferences), and those where we only have latent product characteristics and preferences, with no observables:[8]

1. **Observable Characteristics**: $K = 13, L = 0$. There are no latent characteristics $\vec{\gamma}_j$, but we use the search data to recover the distribution of consumer preferences $V$ over observable characteristics from STR. These include logged meeting space, size category, price segment, a dummy for the hotel being independent, the 6 location

[7]For consumers who search only one product, preferences are learned to maximize the match value with the searched hotel, as would occur with discrete choice purchase data.

[8]We could in principle include both observable and latent characteristics in the same model, but this would result in a high-dimensional characteristic space that would make downstream demand estimation computationally difficult.

type dummies, and latitude/longitude, which capture geographic preferences beyond location type. This specification evaluates the usefulness of our embeddings algorithm solely for recovering consumer preferences.

2. **Latent Characteristics**: $K = 0, L = 12$. We assume that we do not have access to observable characteristics of hotels from the STR dataset. This is done to evaluate the ability for search data to non-parametrically recover a characteristic space that captures the relevant product differentiation over hotels. This will allow us to evaluate how well our model might transfer to settings where we do not observe rich observable characteristics (e.g. books, movies), and one would need to rely on search data to recover the characterstic space. Note that we still allow consumers to have preferences over price, which is observed directly in the search microdata.

**Limitations of Specification Choices** There are two limitations to how we choose to estimate our model of consumer preferences. The first is our choice to include latitude and longitude linearly as a proxy for spatial preferences in the observable characteristics model. In practice, more appropriate measures of spatial preferences would include transformations of these coordinates to capture distances to various amenities consumers value, such as the city center, museums, and airports. We chose to not include these distance metrics because (a) it would require extensive institutional knowledge to construct these measures in each market, which is typically unavailable when working with large-scale search datasets, and (b) adding these distances would dramatically increase the dimensionality of the characteristic space, making downstream demand estimation computationally difficult. At the same time, it is likely that adding a small set of auxiliary measures based on geography, such as distance to city center, to the observable characteristics model would improve the quality of our estimated preferences. This is particularly important given that in Section 6, we will benchmark our search-based preference specifications against a traditional demand model with these observables, which may be unable to capture spatial differentiation.

The second is that we assume that each consumer considered all products in our dataset when deciding which hotels to search. As mentioned in our discussion of product rankings, this assumption is likely to deviate from reality. In practice, a consumer typically searches a subset of hotels within a specific geographic market, for example when planning a vacation to a particular location but unsure what hotel to book. While our demand estimation in Section 6 does restrict choices to be within a geographic market, we do not impose this restriction during the estimation of preferences from search data. In part, this is because we do not want to restrict certain geographic preferences a priori – it may be that consumers co-search hotels in different geographic markets, and thus are actually close substitutes, possibly because of similar amenities at each hotel, or because consumers view these geographically distinct markets as similar. In principle, our model allows for imposing this restriction ex-ante, if one had data on which markets a consumer considered (say, by using their query search words that generated the search options). Such a restriction may improve the search-based preferences in this setting. This choice is likely to be consequential as we show later in the empirical sections that much of what the latent preference capture is spatial differentiation.

**Computational Details.** We provide an implementation of the search model written using the `keras` machine learning package in python.[9] This expedites the time it takes to optimize the model's parameters due to (1) automatic gradient computation and (2) use of GPUs for optimization. We maximize the posterior likelihood of the model using batch optimization[10] and the ADAM stochastic gradient optimizer (Kingma and Ba 2014).

In order to avoid overfitting the model parameters to the search data, we select hyper-parameters (# of dimensions,# of training iterations, and regularization parameter) using a leave-one-out approach for searched hotels.[11] For each session with two or more searched ho-

[9] Link: `https://github.com/luisarmona/learning-mkt-struct`.

[10] We use a batch size of 10,000 inequalities per gradient evaluation

[11] Because of the differing scale and time-varying nature of prices relative to observables, we do not regularize the price coefficients $\alpha_i$ during estimation.

tels, we exclude all inequalities that involve one randomly selected hotel being preferred, as in (Rendle, Freudenthaler, Gantner, and Schmidt-Thieme 2009).[12] This removes 36% of all inequalities from these multi-hotel search sessions. After each iteration through the training set of inequalities, we evaluate the likelihood of the model on the held-out validation set, and iterate until we do not improve the likelihood of the held-out validation set. We then re-run the model on the full set of inequalities using the hyperparameters that maximize the likelihood of the validation set. Appendix Figure A1 displays the out-of-sample log-likelihood as we vary the number of latent dimensions, by regularization hyperparameter.[13] Across dimensions, the best out-of-sample performance occurs with a regularization parameter of $\lambda_\Theta = 10^{-6}$.[14] The returns to additional latent dimensions are negligible after 12 latent dimensions, which is comparable to the dimensionality of our observed characteristics, so we choose the model with 12 dimensions for our analysis.

### 3.3.2 Demand Model

Given that the choice probabilities follow a mixed logit we with endogenous prices, we can use the methods developed in Berry, Levinsohn, and Pakes (2004) to estimate the demand for products, conditional on latent parameter estimates $\hat{\Theta}$. Given choice data from the same sample of search data, the parameters governing demand are already learned from the search data, up to a scale parameter $\sigma_\zeta$ and a location shifter $\mu$ of utilities due to search costs, greatly simplifying demand estimation. In our setting, we lack purchase data for our sample of consumer searches, but have traditional market-level data on price and quantities. As a result, our set of search and purchase consumers differ, which may imply

[12] For example, if A and B are the only searched hotels in a session $i$, and $A$ is the chosen excluded hotel, we remove all inequalities of the form $A \succ_i x$ for unsearched hotels $x$ and only rely on inequalities of the form $B \succ_i x$ during training. Sessions with only one searched hotel are not used during training.

[13] We use a grid of $10^{-x}$ from $x = 4$ to $x = 7$ for $\lambda_\theta$, as this is the region where regularization leads to a non-degenerate number of training iterations.

[14] We normalize the regularization parameter during training by $|\Theta|$, the number of parameters. For the specification with only observable characteristics, we found the hyperparameter $\lambda_\Theta = 10^{-1}$ performed best, and use the learned consumer preferences from this model for all analysis with latent preferences on solely observables characteristics. For this specification, we normalized observables to have mean zero and standard deviation one, so that regularization treats latent preferences equally along each dimension.

a different distribution of preferences across these population. To accommodate such a scenario, we assume the distribution of preferences $\alpha_i, \beta_i$ for our search sample $V$ and the full market population of hotel consumers $G$ belong to the same location-scale family. Explicitly, preferences from search relate to the market-level preference distribution as follows:

$$\vec{v}_i \sim G, \qquad v_i = [\bar{\alpha}, \bar{\beta}] + \Sigma \hat{v}_i \qquad \hat{v}_i \sim \hat{V} \tag{14}$$

where $\hat{V}$ denotes the estimated distribution of preferences from search data, $\Sigma$ is a $(K+L) \times (K+L)$ rescaling matrix, and $[\bar{\alpha}, \bar{\beta}]$ denotes an unknown mean preference vector. In our exercises, we assume $\Sigma$ is a diagonal matrix, so this is equivalent to assuming the demand population distribution is equal to the search population up to an affine transformation. We interpret this assumption as the demand population having the same underlying preference correlation structure as the search population, but allow the importance/scale of consumer heterogeneity to differ by characteristic. This also allows us to estimate our demand system using traditional methods, simply substituting $\hat{V}$ for the typical assumption of random (e.g. normal) unobserved heterogeneity. Since the mean utilities may also differ across populations, we discard the $\delta_j$'s learned from search data and re-estimate $\delta_{j,t}$ using our market-level demand data.

Following Berry, Levinsohn, and Pakes (2004) we estimate the model by generalized method of moments (GMM), using moments conditions of the form $E[\xi_{j,t}|Z_{j,t}]$ where $\xi_{j,t}$ is the product-market specific unobservable defined in the consumer preferences above, and $Z_t$ are a set of market-specific instruments that include own and rival product characteristics.

**Identification** Notice that because hotel characteristics do not vary over time, we absorb the term $\bar{\beta}X_j$ (in addition to the search cost parameter $\mu$) into the fixed effect $\delta_j$, so that $\bar{\beta}, \mu$ are not separately identified. In our empirical application, we choose to calibrate the price coefficient from a paper using the same demand dataset, though in principle if one had access to exogeneous price shifting instruments, the moment restrictions on these instruments would identify $\alpha$. In our context, the moments from GMM instead identify $\Sigma$, the parameter governing the importance of unobserved heterogeneity. This requires a set of instruments that properly capture changes to substitutability across products with similar characteristics. One candidate set of instruments, and the ones used in this paper, are the product differentiation instruments of Gandhi and Houde (2016), which have been shown to be helpful for identification and reduce bias in the estimation of random coefficients. We provide more details on estimation in our empirical application in Section 6.

# 4 Information Content of Latent Characteristics

We begin by visually assessing whether characteristics $\gamma_j$ obtained from search data capture the spatial distribution of hotels. We do so by projecting the 12-dimensional latent characteristics obtained from the latent characteristics search model to a two-dimensional space, using the TSNE method (Maaten and Hinton 2008)[15], to see if clusters formed on the 2-d space correspond to STR's geographical market definitions.

Figure 1 plots in the left panel the geographical locations of hotels in our STR dataset, along with the 2-D projection of their locations in their latent characteristic space learned from the search data in the right panel. Hotels are colored according to the geographical market they reside in, according to STR. We see that hotel clusters formed in the latent characteristic space largely correspond to the geographical markets defined by STR.[16] This suggests that much of what is learned in the latent characteristic space is spatial differentation. It also suggests that we may be able improve estimation by limiting choice sets during search to reside within geographical markets. However, this would come at the cost of dropping consumers who search across markets, and would require a priori knowledge of market boundaries, which we would rather not impose on the model. Indeed, the imperfect corre-

[15]Our perplexity hyper-parameter is set to the square root of the number of hotels in each TSNE embedding estimation routine

[16]The multi-colored cluster of hotels in the center corresponds to hotels which appear in Expedia t but are not actually searched by any of the consumers in our sample.

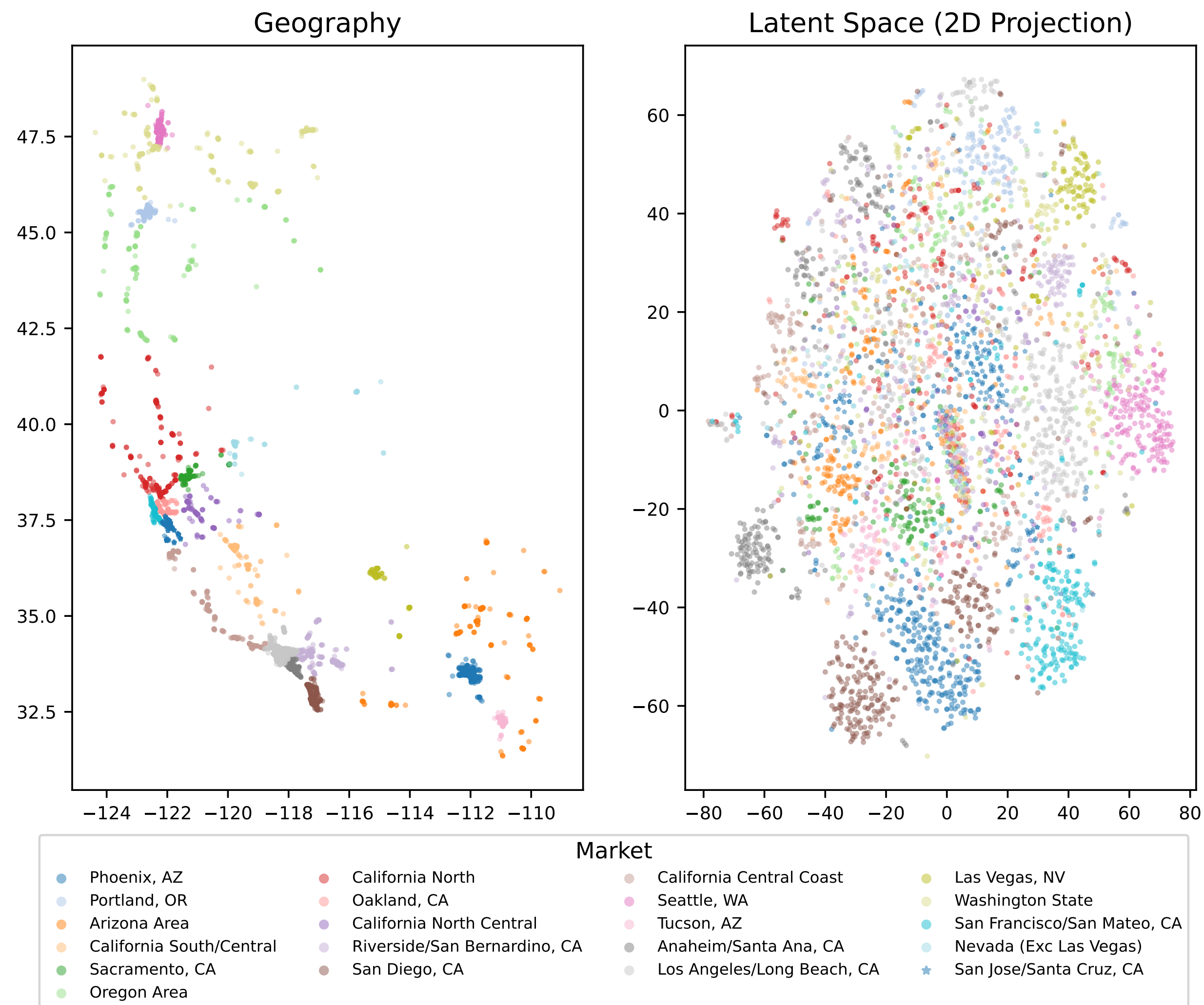



Figure 1: Two-Dimensional Representation of Learned Latent Space by Market

spondence of geographic markets to clusters suggests that search data may be used in future work to better inform market structure in a data-driven way.

We now more formally characterize the information content of the embeddings learned from search data. To do so, we evaluate the efficacy of embeddings trained in the Latent Characteristics model to predict observable characteristics of hotels not included in estimation. Specifically, we evaluate whether search data alone can recover product differentiation, by validating that the search model trained with *no* observable characteristics accurately captures observable characteristics we know influence demand in the market for hotels. The thought exercise is that, if we were unable to observe any characteristics of products, but

| | Latent Characteristics | | Observable Characteristics | |
|---|---|---|---|---|
| | In-Sample | Out-of-Sample | In-Sample | Out-of-Sample |
| Latitude-Longitude | 0.555 | 0.470 | 0.045 | 0.060 |
| Independent Hotel | 0.052 | 0.010 | 0.426 | 0.345 |
| Near Airport | 0.254 | 0.203 | 0.059 | 0.038 |
| Hotel Size Category | 0.325 | 0.297 | 0.686 | 0.671 |

Table 4: Predictability (R-Squared) of Observable Characteristics from Latent and Other Observable Characteristics

knew products were not homogeneous, could we learn characteristics from search data that captured the product heterogeneity consumers care about?

We measure the information content of embeddings in predicting observable characteristics by estimating a flexible function $f(\cdot)$ that takes as inputs the latent characteristics $\gamma_j$ and attempts to predict an observable characteristic $X^o_{j,c}$. We split the sample of STR hotels randomly into an 80%-20% training and test sample, and then use 20% of the training sample as a validation sample to optimize hyperparameters of the neural network. We use a 10-layer deep neural network to estimate each characteristic $X^o_{j,c}$, using RELU activation functions for intermediate layers, and optimize over (a) the number of nodes per layer, (b) the regularization applied during training (c) the number of training iterations.[17] Then, after tuning the hyperparameters, to provide the best in-sample fit to the training data, we predict the held-out 20% test sample of hotels, and evaluate the predictability of characteristic $X^o_{j,c}$ using the $R^2$ metric:

$$R^2_c = \frac{\sum_{j=1}^{N_{test}}(X^o_{j,c} - f(\gamma_j))^2}{\sum_{j=1}^{N_{test}}(X^o_{j,c} - \bar{X}^o_{test,c})^2} \tag{15}$$

Which measures how much of the variation of characteristic $X^o_{j,c}$ in the test-sample is explained by the neural net $f$ optimized over the training sample.

We pick as our target variables the latitude-longitude location of each hotel, whether a hotel is independent, whether a hotel is located near an airport, and the size category of each hotel. Table 4 reports the R-squared metric for each characteristic from the neural

[17]Hyperparameters chosen for each characteristic are reported in Appendix Table A1

net. To benchmark these results, we also estimate a neural network trained only on observable characteristics (except the target observable), which exploits the correlation across observable characteristics within each product. We find that latent characteristics are able to predict geographical location, whether the hotel is near an airport, and hotel size. The predictability of geographic variables from latent characteristics suggest that our learned latent preferences do not recover only idiosyncratic search patterns, but also correlate with characteristics we know consumers have preferences over when making a decision for purchasing products in our empirical setting, such as spatial differentiation. In Appendix A, we provide supplementary evidence to this exercise, by examining the correlation structure of latent consumer preferences, as well as implementing a classifier that shows hotels with similar latent characteristics share observables such as brand. The Appendix results are consistent with the evidence provided here: latent parameters recovered from search data are meaningful in explaining product differentiation in the hotel market.

# 5 Application 1: Entry Event Study

We evaluate the ability of unobserved characteristics learned from search data to capture substitution patterns in a reduced-form setting. Fundamental to the concept of substitution patterns is that consumers are more likely to substitute away from a particular product when there are competing products available that are "close" in the characteristic space over which consumers have preferences. Therefore, we expect when a new product enters the market, the incumbent suppliers that stand to lose the most are products whose characteristics are close to those of the entrant. We hypothesize that the learned characteristics we recover from our search model capture parts of the characteristic space consumers have preferences over, yet the econometrician does not observe. To test this formally, we estimate an event study that captures the heterogeneous effect of entry depending on whether an incumbent hotel is "close" in characteristic space to the entrant.

We perform this exercise as follows: first, we identify all hotel entries that occurred between January 2002 and March 2018 based on the listed open date in the STR characteristics dataset. Let $t_e$ denote the entry month of entrant $e$. We then take all hotels in the same geographical market for whom we have complete transaction data $\pm 12$ months around this entry to produced a balanced panel. Given a characteristic space $\mathcal{X}$, for each incumbent hotel $j$ included in the panel, we compute its distance in characteristic space to the entering hotel $e$ as follows:

$$d_{\mathcal{X}}(j,e) = ||X_j - X_e||_2, \quad X_j, X_e \in \mathcal{X} \tag{16}$$

where $||\cdot||_2$ is the L2 norm. let $\hat{H}_{\mathcal{X},e}$ denote the empirical distribution of distances for all hotels included in the panel for entrant $e$. We then classify a hotel as "close" in characteristic space if its distance $d$ is below the 10th percentile in distance among all hotels included in the panel for entrant $e$.

$$\mathbb{1}\{j \text{ close in } d_{\mathcal{X}}\} = \begin{cases} 1 & \text{if } H^{-1}_{\mathcal{X},e}(d_{\mathcal{X}}(j,e)) \leq 0.1 \\ 0 & \text{else} \end{cases} \tag{17}$$

We perform this calculation for all entries in the STR dataset, stack the entry panels for each entrant $e$, and estimate the following stacked event study specification:

$$\log(q_{j,t,e}) = \alpha_{j,e} + \delta_{t,e} + \sum_{s=-13}^{11} \beta_s \mathbb{1}\{j \text{ close in } d_{\mathcal{X}}\} \times \mathbb{1}\{t - t_e = s\} + \epsilon_{i,t} \tag{18}$$

where $\alpha_{j,e}$ denotes hotel-entry panel pair fixed effects, $\delta_{t,e}$ captures the effect of the entry in period $t$ on all hotels in the market, and $\beta_s$ measures the differential effect the hotel entry on hotels close in characteristic space in month $s$ relative to the entry date $t_e$. We expect $\beta_s < 0$ for $s \geq 0$, implying that hotels close in the characteristic space are more negatively effected in terms of sales after the new hotel enters.

We consider 3 characteristic spaces for this event study:

1. **Geographical Distance**: This measures the distance between hotels, determined by the euclidean distance in their latitude and longitude. Naturally, because preferences for hotels are in part spatially correlated, we expect incumbent hotels physically near an entrant to be more negatively effected.

2. **Distance in Other Observable Characteristics**: This measures the distance in all observable characteristics provided by STR, excluding longitude and latitude. Because the units of the observable characteristics vary, we construct the observable characteristics distance metric using the Mahalanobis distance. This is identical to the euclidean distance implied by the L2 norm using the following transformed variable:

$$\tilde{X}_j = LX_j^o, \quad \text{where } LL' = \Sigma \text{ and } \Sigma = Cov(X_j^o, X_j^o) \tag{19}$$

   Thus, we compute the covariance matrix of observable characteristics, perform a Cholesky decomposition of this matrix to recover $L$ then multiply the characteristics by $L$ to recover the standardized observable characteristics.

3. **Distance in Latent Characteristics**: Using the characteristics learned from the Latent Characteristics model described in Section 3.3, we compute the euclidean distance in latent characteristics, so that $X_j = \gamma_j$. No normalization is required to make units comparable, since these embedding characteristics are learned in units of utility.[18]

We estimate the event study for all 3 distance metrics and compare their effectiveness in capturing the substitution patterns of consumers when a new product enters their choice set.

In Figure 2, we plot the event study coefficients $\beta_s$ for all three distance metrics described above. Effects are relative to month $s = -1$. For all distance metrics, we see that prior to the entry month $s = 0$, there is no effect of being close to the entrant hotel, since it has not yet entered the market. After $s \geq 1$, demand for nearby hotels in the latent space

[18]Estimating the event study using the Mahalanobis distance in latent characteristics produces very similar results.

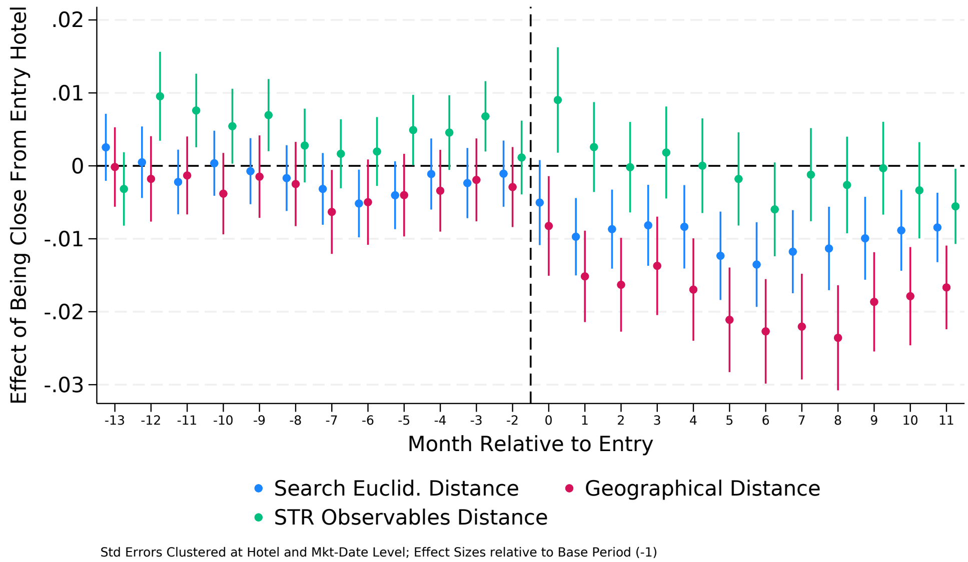



Figure 2: Effect of Closeness (< 10th percentile in distance) of Hotel Entry on Incumbent Hotel Demand, By Distance Metric

decreases by 1%, and this effect is consistent for the subsequent periods included in the event study sample. Distance in the observable characteristics space yields largely noisy estimates. By contrast, we see a statistically significant decrease in demand for incumbent hotels that are close in either geographic or latent characteristic space to the entrant. Because competition in the hotel market is geographically spatial, the statistically significant effect on geographic distance is not surprising. What is interesting is that the latent characteristics learned from search data are also significant and therefore have "learned" some notion of geography relevant for demand. The effect sizes are smaller though, suggesting that the latent characteristics are unable to fully capture spatial competition. This may due to differential factors contributing to how consumers decide to search for hotels versus actual purchases. Nonetheless, we conclude that data on search behavior alone is able to capture meaningful ways in which products are differentiated that may not be observable to researchers.

# 6 Application 2: Merger Analysis

We estimate the choice model described in Section 3 and compare various specifications of consumer demand for merger analysis. To quantify how effective the latent search parameters are at predicting demand, we benchmark the model against a traditional mixed logit using only observable characteristics and random consumer heterogeneity.

In November 2015, Marriott International announced it would be acquiring the competing Starwood Hotels company, creating the largest hotel chain in the world (Dogru, Erdogan, and Kizildag 2018). In Appendix B, we provide direct evidence that prices decreased by a large proportion (5%) in markets with a high concentration of Marriott-Starwood hotels post-merger. We use these large price changes to test whether our estimated demand model specifications can accurately predict demand in an out-of-sample period which experienced large price changes, causing demand substitution across products. We evaluate the fit of these demand models by predicting demand after a major merger in the hotel industry that induced large price changes. To perform this test, we estimate a demand system on the pre-merger hotel transaction data data (2012 to 2015), and evaluate the model's ability to predict demand changes for hotels post-merger announcement (2016 to 2018), which is held out when we optimize the GMM objective function of BLP. To allow for a fair comparison between the models based on latent and observable characteristics, we limit this exercise to the 4,021 hotels that have at least one user search the hotel in our Expedia dataset (98.3% of total purchases in our sample).[19]

**Specifications.** We consider two specifications of $X_j$, corresponding to each of our two search models: (1) the observable characteristics from STR ($X_j = X_j^o$, 13 dimensions), and (2) the latent characteristics estimated from search data ($X_j = \gamma_j$, 12 dimensions). We assume that $v_i$, the unobserved preferences, take 1 of 2 forms. Both sets of preferences are held constant across all markets in the data.

[19]Our full sample of 4,218 hotels used to estimate latent parameters includes hotels that were present in Expedia but not searched by any users during our sample.

1. $v_i \sim N(0, I)$. This follows the standard mixed logit assumption of (Berry, Levinsohn, and Pakes 1995). For normally distributed heterogeneous preferences $v_i$, we employ Gauss-Hermite quadrature over the unobserved heterogeneity.

2. $v_i \sim \hat{V}$ where $\hat{V}$ is the empirical distribution of (demeaned) search preferences for all consumers in our search data. Preferences are demeaned to have mean zero for each characteristic, and shifted in importance by the scaling matrix $\Sigma$.

This implies four demand specifications, in addition to the typical BLP mixed logit demand model. The specification with normally distributed preferences and latent characteristics benchmarks the informativeness of using latent product characteristics over observable product characteristics. Conversely, the specification with latent preferences and observable characteristics benchmarks the improvement in demand prediction from allowing more flexible consumer heterogeneity learned from search data. Finally, the specification with both search-learned characteristics and preferences benchmarks the performance of using only search microdata and market-level quantity data to inform consumer preferences.

**Estimation Details.** We define a market $t$ for this merger analysis as a STR-provided geographical market $\times$ year.[20] We use the industry-provided market definitions to mirror the typical market definition that would be used in demand estimation. We calibrate the mean price parameter, setting $\bar{\alpha} = -0.018$. This is the preferred estimate in Lewis and Zervas (2016) in their demand analysis on the same data, and implies an average demand elasticity of approximately -2.3, which seems reasonable. We choose to make this calibration because we are unable to find good supply-shifters for price. Lewis and Zervas (2016) are able to circumvent this problem because they estimate a full model of the supply side that accounts for capacity constraints; this is beyond the scope of this paper. Since our price coefficient

[20]We set market size according to a heuristic similar to that of Lewis and Zervas (2016): for each geographical market $g$, we take the month-year with the largest total number of rooms sold by all hotels in that market, and set market size $M_g$ to 1.5 times this quantity, multiplied by 12 to account for the fact that we estimate demand at the annual level. Thus the size of each geographical market is constant over time.

is calibrated, the only parameters we estimate are the the non-linear parameters, $\Sigma$, and the hotel fixed effects $\delta_j$. $\Sigma$ captures the relative importance of preference heterogeneity for each characteristic in $X_j$. This heterogeneity, along with the choice of characteristic space itself (i.e. including only observable or latent characteristics), will determine the substitution patterns relevant to merger analysis. Hotel fixed effects are included so that each model predicts the same "levels" of demand for each hotel, and thus differences across specifications load solely on the substitution patterns (as measured by $\Sigma$) that each model is able to estimate.

To estimate these demand models, we perform two-step GMM use the `pyblp` package (Conlon and Gortmaker 2019). We use the quadratic differentiation instruments proposed by (Gandhi and Houde 2016) to construct our first-step instruments $\mathbf{Z}$.[21] For each characteristic $c$ in $\vec{X}_j$, we construct instruments $Z_{j,t,c} = \sum_{k \in t, k \neq f_{j,t}} (X_{j,c} - X_{k,c})^2$, where $f_{j,t}$ denotes the set of hotels with the same affiliation as $j$ in market $t$. We then use these to form empirical moments $\hat{g}_c(\Sigma) = \frac{1}{N} \sum_{j,t} \xi_{j,t}(\Sigma) \times Z_{j,t,c}$, where $\xi$ is implicitly a function of $\Sigma$ due to the contraction mapping. In the first step, we use the weighting matrix $(\mathbf{Z}'\mathbf{Z})^{-1}$. In the second step, we use the first step estimates to construct an approximate form of the optimal instruments, $Z^{opt}_{j,t} = \frac{1}{\hat{\sigma}^2_\xi} E[\nabla_\theta \xi_{j,t} | \xi_{j,t} = 0, Z_{j,t}]$, where $\hat{\sigma}^2_\xi$ is the first-stage estimate of the variance of $\xi$, as recommended by (Conlon and Gortmaker 2019). We then solve the objective once again to obtain our final estimates for each model.[22]

While we are able to estimate $\xi$ in the pre-merger data to match market shares/demand exactly, this is unavailable in the post-merger data. We evaluate the prediction of the model assuming $\xi_{j,t} = 0$; This is to be consistent with the moments $E[\xi_{j,t} | \mathbf{Z}_{j,t}] = 0$.

[21] We only include the rival differentiation instruments in our estimation. Due to our hotel fixed effects, variation in our differentiation instruments comes solely from entry/exit in markets from competing brands during our sample.

[22] For normally distributed preferences, we found that the objective function with optimal instruments had flat regions that prevented the optimizer from converging at multiple starting points. For these preference, we used the original instruments and the efficient weighting matrix to obtain reasonable standard errors. The out-of-sample MAE and MSE are qualitatively very similar using optimal instruments.

| Preferences | Characteristics | Pre-Merger Error MAE | Pre-Merger Error MSE | Post-Merger Error MAE | Post-Merger Error MSE | Post-Merger % Decrease from Logit MAE | Post-Merger % Decrease from Logit MSE |
|---|---|---|---|---|---|---|---|
| None | Observables | 0.138 | 0.058 | 0.303 | 0.201 | - | - |
| Normal | Observables | 0.137 | 0.056 | 0.293 | 0.171 | -3.32% | -14.99% |
| Normal | Price & Latent | 0.134 | 0.054 | 0.287 | 0.164 | -5.21% | -18.19% |
| Search | Observables | 0.102 | 0.042 | 0.186 | 0.103 | -38.40% | -48.92% |
| Search | Price & Latent | 0.124 | 0.054 | 0.238 | 0.184 | -21.44% | -8.34% |

Table 5: Prediction Errors of Structural Demand Model

Predicted demand in the post-merger period(s) takes the following form:

$$\hat{q}_{j,t}(\Sigma, G(i)) = M_g \cdot \Big(\frac{1}{B} \sum_{i \sim G(i)} \frac{\exp(\delta_j + \alpha_i p_{j,t} + \beta_i X_j)}{1 + \sum_{k \in J_t} \exp(\delta_k + \alpha_i p_{k,t} + \beta_i X_k)}\Big) \tag{20}$$

Our loss functions for evaluating demand prediction is the mean-squared error and mean absolute error of log demand in the post-merger dataset:

$$MSE(\Sigma, G) = \frac{1}{N_{\text{post-merger}}} \sum_{j,t} (\log(\hat{q}_{j,t}) - \log(q_{j,t}))^2 \tag{21}$$

$$MAE(\Sigma, G) = \frac{1}{N_{\text{post-merger}}} \sum_{j,t} |\log(\hat{q}_{j,t}) - \log(q_{j,t})| \tag{22}$$

**Merger Results** Table 5 displays the MSE and MAE of each demand model's predictions, both pre and post-merger, when we set the demand shock $\xi_{j,t}$ to zero.[23] The first row shows the predictive performance of a simple logit model of demand with no unobserved heterogeneity in preferences. We benchmark the performance of each model relative to this baseline logit model in the out-of-sample post-merger data by computing the relative decrease in an error metric for model $m$ from the logit model. The second row shows the improvement in performance from a standard BLP model with normally distributed heterogeneity over observable characteristics. BLP is able to improve upon the logit in

[23] In Appendix Table A4, we report the full parameter estimatess from each of the 4 models considered.

predicting both pre and post merger demand, improving the prediction MSE by 15% out-of-sample. Using only latent instead of observable characteristics outperforms standard BLP along both metrics, suggesting the unobserved characteristics learned from search data are informative in capturing preferences at the purchase stage. Including both latent preferences and characteristics improves upon both according to mean absolute error, but does poorer according to mean squared error, due to some outlier predictions. When we add search-based preferences to the observable characteristics only BLP model, in place of normally distributed heterogeneity, our out-of-sample fit improves over the logit by 48% according to MSE, and 38% according to MAE. Across our specifications, the in and out of sample performance are comparable, suggesting we do not overfit in our estimation. Our usage of transfer learning, by estimating a high dimensional latent characteristic and preference space in a first stage with our large search dataset, allows us to estimate a small number of parameters during demand estimation, which prevents overfitting, while still capturing important substitution patterns. Taken together, the results suggest that preferences and characteristics learned from search data have wide scope to improve demand estimation. Moreover, these may work in tandem with observable characteristics when available, as seen by the specification with observable characteristics and search-based preferences, which provides the best out-of-sample fit.

To better understand the performance differences across the various models, Figure 3 plots the implied substitution patterns across demand specifications (with observable characteristics) for products that are close in the characteristic space. For each model, we calculate the mean diversion ratio (Conlon and Gortmaker 2019) for products similar in characteristics. The diversion ratio is a normalized version of the cross-price elasticity that can be interpreted as follows: given a marginal price increase in $j$, what fraction of consumers that no longer purchase product $j$ switch to $k$? By construction, the diversion ratios across all other choices sum to 1. In the bottom row, we report the fraction of consumers that switch to the outside option. This captures the relative substitutability of all other products: if consumers have strong preferences for the characteristics of the product they

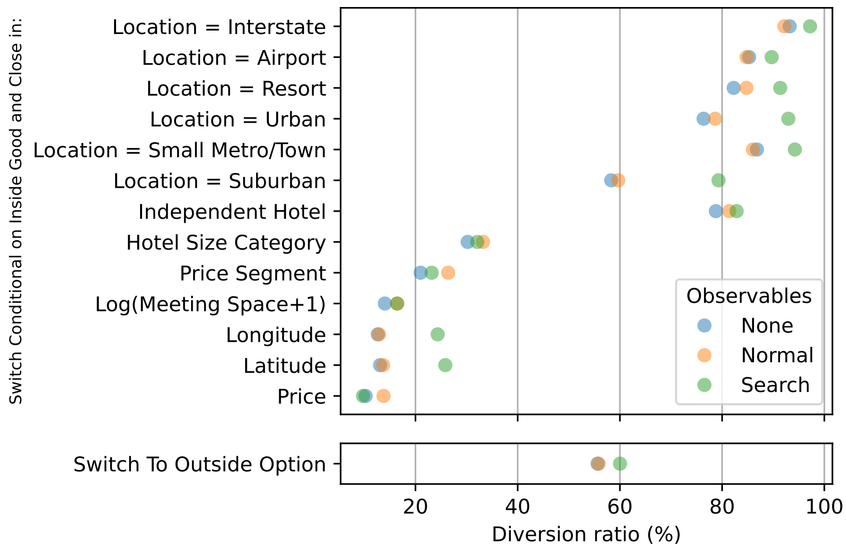



Figure 3: Diversion Ratios Across Demand Models with Observable Characteristics

purchase, switchers will be more likely to not purchase at all than switch to another product with differing characteristics. In the remaining rows, we report the probability of switching to a product that is "close" along a particular characteristic, conditional on not switching to the outside good. We define closeness as we did in Equation 17 for our event study application: for each hotel, being within the 10% percentile of absolute distance among all other products in the same market. For dummy indicators, such as being an independent hotel, this captures the fraction of switchers that choose a hotel with the same characteristic.

Relative to the logit model, with no unobserved consumer heterogeneity, we find small increases in substitution, primary along hotel location types and size category when we add normally distributed preference heterogeneity. In contrast, for the specification with consumer heterogeneity learned from search data and observable characteristics, our best-performing model for predicting post-merger demand, we find an increase in substitution in the outside option, implying a higher degree of product differentiation. Additionally, along nearly every dimension, the likelihood of switching to a product similar in characteristics increases substantially. This is particularly salient for geography-based characteristics, such as geographical coordinates and our set of location type indicators. This suggests that this model is able to better predict post-merger demand better because it is able to capturing the geographical substitution patterns determining demand. In Appendix Figure A2, we similarly estimate that, for latent characteristics, search preferences imply lower substitution along certain dimensions of the latent characteristic space. In Appendix Figure A3, we show that the both models based on learned search preferences exhibit a steeper gradient between substitutability and geographical distance, consistent with geographic preferences driving the improvements from using preferences learned from search data.

# 7 Conclusion

We have presented an approach for using search data to augment traditional demand estimation, in a setting in which search is observed in one dataset, and choice is observed in another. In particular, we use this auxiliary search data to recover latent parameters that are useful in predicting market-level purchasing decisions. The key identification strategy in our methodology is that because there are multiple choices made during the search process, latent preferences and product characteristics can be simultaneously estimated. Through an event study, we show these latent characteristics are able to predict the relative losers on the supply side from a new entrant to a market, suggesting they are meaningful and informative to market structure. In our analysis of the Starwood-Marriott merger, it appears that the main value added lies in the estimation of consumer preferences from search, since this allows us to model choice with a flexible correlation structure rather than making strong parametric restrictions of random heterogeneity.

Many questions are left open in this work. One is how best to use the data when both search and choice are observed on the same platform, and the goal is counterfactual prediction of events on that platform. A second is how the platform's own choices of product display and prominence (i.e. search rankings) should be incorporated into the analysis. Last, a natural application that is left unexplored is to a market where observable characteristics are poor predictors of choice, such as the market for books, so that latent characteristics recovered from search data could better illuminate which products are close substitutes.

# 8 Competing interests disclosure

All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript. The authors have no funding to report.

# References

Amano, T., A. Rhodes, and S. Seiler (2019): "Large-scale demand estimation with search data," .

Athey, S., D. Blei, R. Donnelly, F. Ruiz, and T. Schmidt (2018): "Estimating Heterogeneous Consumer Preferences for Restaurants and Travel Time Using Mobile Location Data," *arXiv preprint arXiv:1801.07826*.

Bajari, P., D. Nekipelov, S. P. Ryan, and M. Yang (2015): "Machine learning methods for demand estimation," *American Economic Review*, 105(5), 481–85.

Berry, S., J. Levinsohn, and A. Pakes (1995): "Automobile prices in market equilibrium," *Econometrica: Journal of the Econometric Society*, pp. 841–890.

——— (2004): "Differentiated products demand systems from a combination of micro and macro data: The new car market," *Journal of political Economy*, 112(1), 68–105.

Cardell, N. S. (1997): "Variance components structures for the extreme-value and logistic distributions with application to models of heterogeneity," *Econometric Theory*, 13(2), 185–213.

CONLON, C., AND J. GORTMAKER (2019): "Best practices for differentiated products demand estimation with pyblp," *Unpublished Manuscript*.

CORTES, D. (2018): "Cold-start recommendations in collective matrix factorization," *arXiv preprint arXiv:1809.00366*.

DE LOS SANTOS, B., AND M. WILDENBEEST (2017): "E-book pricing and vertical restraints.," *Quantitative Marketing and Economics*, 15, 85–122.

DEATON, A., AND J. MUELLBAUER (1980): "An Almost Ideal Demand System," *The American Economic Review*, 70(3), 312–326.

DOGRU, T., A. ERDOGAN, AND M. KIZILDAG (2018): "Marriott Starwood merger: what did we learn from a financial standpoint?," *Journal of Hospitality and Tourism Insights*.

DONNELLY, R., A. KANODIA, AND I. MOROZOV (2020): "A Unified Framework for Personalizing Product Rankings," *Available at SSRN 3649342*.

EINAV, L. (2007): "Seasonality in the U.S. motion picture industry," *The RAND Journal of Economics*, 38(1), 127–145.

ELROD, T. (1988): "Choice map: Inferring a product-market map from panel data," *Marketing Science*, 7(1), 21–40.

ELROD, T., AND M. P. KEANE (1995): "A factor-analytic probit model for representing the market structure in panel data," *Journal of Marketing Research*, 32(1), 1–16.

GANDHI, A., AND J.-F. HOUDE (2016): "Measuring substitution patterns in differentiated products industries," *University of Wisconsin-Madison and Wharton School*.

HONKA, E. (2014): "Quantifying search and switching costs in the US auto insurance industry," *The RAND Journal of Economics*, 45(4), 847–884.

Honka, E., S. Seiler, and R. Ursu (2023): "Consumer Search: What Can We Learn from Pre-Purchase Data?," .

Johnson, C. C. (2014): "Logistic matrix factorization for implicit feedback data," *Advances in Neural Information Processing Systems*, 27.

Keane, M. P., et al. (2013): *Panel data discrete choice models of consumer demand.* Nuffield College.

Kim, J. B., P. Albuquerque, and B. J. Bronnenberg (2011): "Mapping online consumer search," *Journal of Marketing Research*, 48(1), 13–27.

Kingma, D. P., and J. Ba (2014): "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980.*

Koren, Y., R. Bell, and C. Volinsky (2009): "Matrix factorization techniques for recommender systems," *Computer*, (8), 30–37.

Koulayev, S. (2014): "Search for differentiated products: identification and estimation," *The RAND Journal of Economics*, 45(3), 553–575.

Lancaster, K. J. (1966): "A New Approach to Consumer Theory," *Journal of Political Economy*, 74(2), 132–157.

Lewis, G., and G. Zervas (2016): "The welfare impact of consumer reviews: A case study of the hotel industry," *Unpublished manuscript.*

Maaten, L. v. d., and G. Hinton (2008): "Visualizing data using t-SNE," *Journal of machine learning research*, 9(Nov), 2579–2605.

Magnolfi, L., J. McClure, and A. T. Sorensen (2022): "Embeddings and Distance-based Demand for Differentiated Products," *Available at SSRN 4113399.*

McFadden, D. (1973): "Conditional logit analysis of qualitative choice behavior," *Frontiers in Econometrics*, pp. 105–142.

Moraga-González, J. L., Z. Sándor, and M. R. Wildenbeest (2023): "Consumer search and prices in the automobile market," *The Review of Economic Studies*, 90(3), 1394–1440.

Nevo, A. (2001): "Measuring Market Power in the Ready-to-Eat Cereal Industry," *Econometrica*, 69(2), 307–342.

Petrin, A. (2002): "Quantifying the benefits of new products: The case of the minivan," *Journal of political Economy*, 110(4), 705–729.

Rendle, S., C. Freudenthaler, Z. Gantner, and L. Schmidt-Thieme (2009): "BPR: Bayesian personalized ranking from implicit feedback," in *Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence*, pp. 452–461. AUAI Press.

Rudolph, M., F. Ruiz, S. Mandt, and D. Blei (2016): "Exponential family embeddings," in *Advances in Neural Information Processing Systems*, pp. 478–486.

Ruiz, F. J., S. Athey, and D. M. Blei (2017): "SHOPPER: A Probabilistic Model of Consumer Choice with Substitutes and Complements," *arXiv preprint arXiv:1711.03560.*

Salakhutdinov, R., and A. Mnih (2008): "Bayesian probabilistic matrix factorization using Markov chain Monte Carlo," in *Proceedings of the 25th international conference on Machine learning*, pp. 880–887. ACM.

Sams, J. A. (2019): "Learning or Herding? Understanding Social Interactions and the Distribution of Success on a Social Music Sharing Platform," Ph.D. thesis, Stanford University.

Schein, A. I., A. Popescul, L. H. Ungar, and D. M. Pennock (2002): "Methods and metrics for cold-start recommendations," in *Proceedings of the 25th annual international*

*ACM SIGIR conference on Research and development in information retrieval*, pp. 253–260.

URSU, R. M. (2018): "The power of rankings: Quantifying the effect of rankings on online consumer search and purchase decisions," *Marketing Science*, 37(4), 530–552.

WEITZMAN, M. (1978): *Optimal search for the best alternative*, vol. 78. Department of Energy.

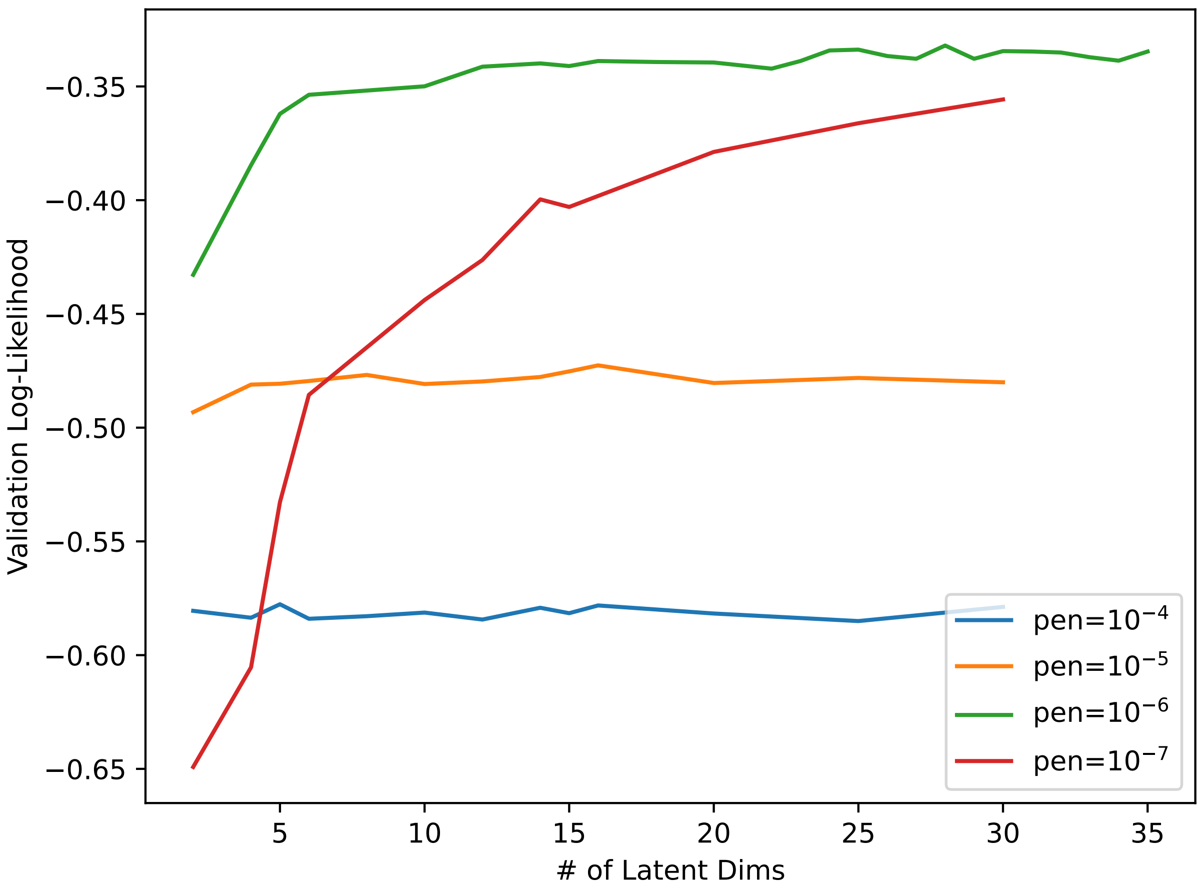



Figure A1: Validation Log-likelihood for Tuning # of Latent Dimensions

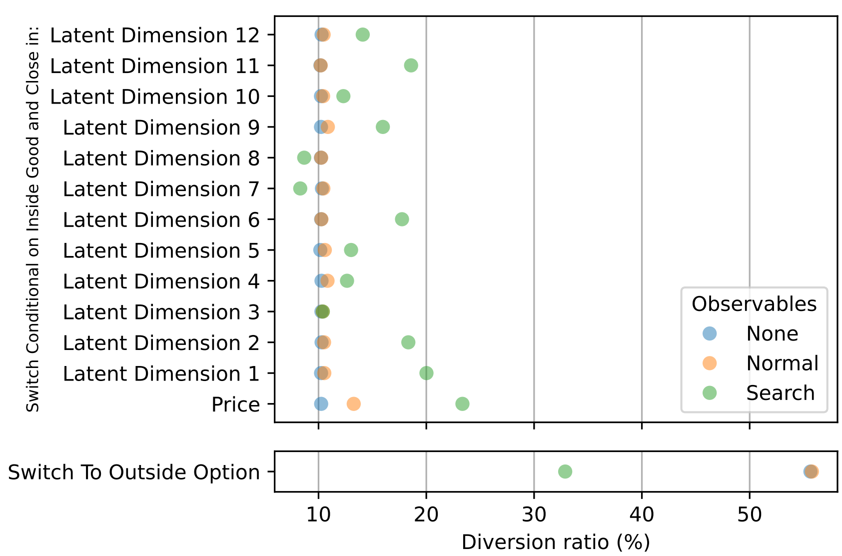
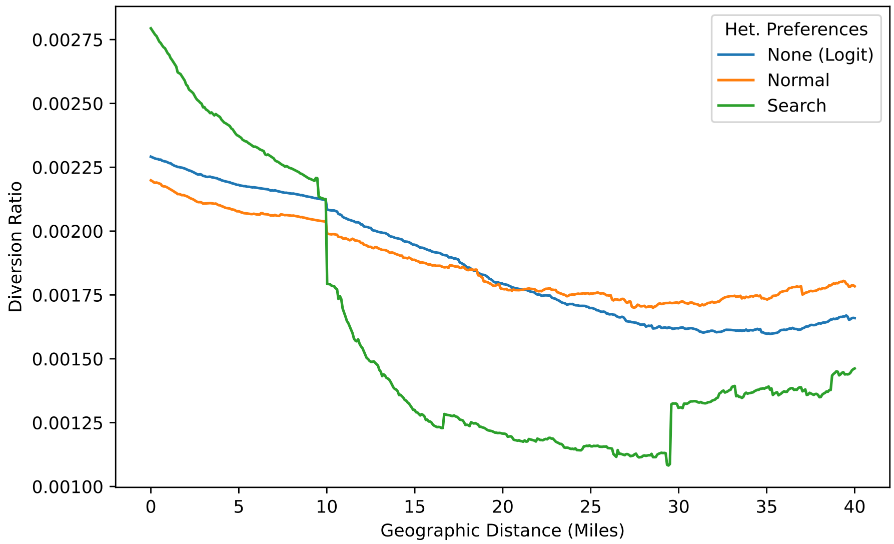
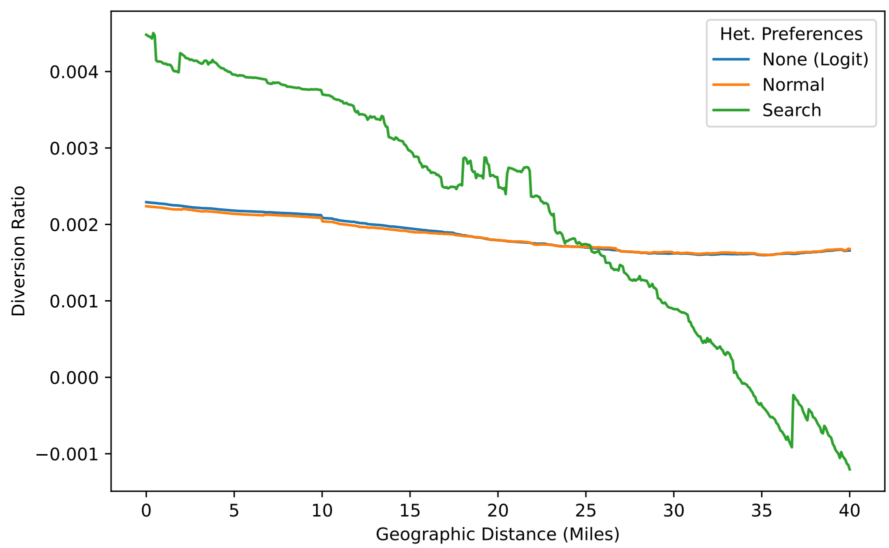



Figure A2: Diversion Ratios Across Demand Models with Latent Characteristics




(a) Observable Characteristics




(b) Latent Search Characteristics

Figure A3: Diversion Ratio As a Function of Distance, by Preference Specification

Figure shows the estimated mean diversion ratio from a kernel regression (uniform kernel, bandwidth = 10 miles) between hotels from our estimated demand model, as a function of the geographical distance between hotels. Het. Preferences denote whether there is no consumer heterogeneity (None), normally distributed heterogeneity (Normal), or heterogeneity based in preferences learned from search data (Search). Panel (a) plots comparisons across models using observables characteristics, while Panel (b) plots comparisons across models using only the price and the learned latent characteristics learned from search data.

| | Latent Characteristics | | | Observable Characteristics | | |
|---|---|---|---|---|---|---|
| | # Iterations | Nodes/Layer | Regularization | # Iterations | Nodes/Layer | Regularization |
| Latitude-Longitude | 8325 | 10 | 0.0001 | 4856 | 5 | 0.0046 |
| Independent Hotel | 430 | 5 | 0.0001 | 735 | 40 | 0.0000 |
| Near Airport | 180 | 40 | 0.0001 | 456 | 20 | 0.0000 |
| Hotel Size Category | 160 | 30 | 0.0046 | 873 | 40 | 0.2154 |

Table A1: Hyperparameters Chosen for Predicting Observable Characteristics

| | | Pre-Merger (Training Data) | | Post-Merger (Test Data) | |
|---|---|---|---|---|---|
| Preferences | Characteristics | MAE | % Decrease from Logit | MAE | % Decrease from Logit |
| None | Observables | 0.138 | - | 0.303 | - |
| Normal | Observables | 0.137 | -0.96% | 0.293 | -3.32% |
| Normal | Price \& Latent | 0.134 | -2.82% | 0.287 | -5.21% |
| Search | Observables | 0.102 | -26.04% | 0.186 | -38.40% |
| Search | Price \& Latent | 0.124 | -10.38% | 0.238 | -21.44% |

Table A2: Prediction Errors of Structural Demand Model (Mean Absolute Error)

| | | Pre-Merger (Training Data) | | Post-Merger (Test Data) | |
|---|---|---|---|---|---|
| Preferences | Characteristics | R-Squared | Increase From Logit | R-Squared | Increase From Logit |
| None | Observables | 0.916 | - | 0.687 | - |
| Normal | Observables | 0.923 | 0.007 | 0.732 | 0.046 |
| Normal | Price & Latent | 0.927 | 0.011 | 0.761 | 0.075 |
| Search | Observables | 0.940 | 0.023 | 0.840 | 0.153 |
| Search | Price & Latent | 0.922 | 0.006 | 0.713 | 0.026 |

Table A3: Fit of Structural Demand Model (R-Squared)

| Characteristics Preferences | Observables | Normal | Search | Price & Latent | Normal | Search |
|---|---|---|---|---|---|---|
| | Price | 0.0129 | 0.564 | Price | 0.00991 | 0.978 |
| | | (0.00116) | (0.798) | | (0.00415) | (0.264) |
| | Latitude | 3.49e-11 | 0.00434 | Latent Dimension 1 | -3.02e-09 | -1.12 |
| | | (4.35e-09) | (0.926) | | (1.41e-09) | (1.3) |
| | Longitude | -2.5e-11 | -0.000721 | Latent Dimension 2 | 1.29e-08 | -0.808 |
| | | (6.14e-09) | (0.539) | | (3.25e-08) | (2.34) |
| | Log(Meeting Space+1) | 7.31e-11 | 0.0284 | Latent Dimension 3 | -1.05e-09 | -0.359 |
| | | (9.27e-09) | (21.8) | | (2.06e-09) | (2.45) |
| | Price Segment | -3.13e-11 | 0.0438 | Latent Dimension 4 | 0.108 | -0.0928 |
| | | (3.73e-10) | (81.1) | | (1.37) | (3.18) |
| | Hotel Size Category | -1.19e-10 | 0.0852 | Latent Dimension 5 | 8.1e-10 | 0.00602 |
| | | (2.72e-08) | (73.8) | | (1.28e-08) | (4.13) |
| | Independent Hotel | 3.91e-10 | 0.372 | Latent Dimension 6 | -3.73e-09 | 0.355 |
| | | (1.15e-08) | (49.3) | | (1.54e-08) | (2.08) |
| | Location = Suburban | -1.22e-10 | -1.58 | Latent Dimension 7 | -4.06e-10 | -0.00261 |
| | | (2.96e-09) | (221) | | (1.16e-09) | (5.76) |
| | Location = Small Metro/Town | 2.49e-10 | -0.868 | Latent Dimension 8 | 1.11e-09 | -0.386 |
| | | (9.36e-09) | (304) | | (7.46e-09) | (3.43) |
| | Location = Urban | -0.23 | -33.3 | Latent Dimension 9 | 1.34e-10 | 0.323 |
| | | (25.4) | (48.3) | | (1.12e-09) | (0.755) |
| | Location = Resort | -3.65e-10 | -2.38 | Latent Dimension 10 | -1.26e-10 | -0.0616 |
| | | (2.53e-08) | (173) | | (1.91e-10) | (2.35) |
| | Location = Airport | -5.62e-10 | -17.1 | Latent Dimension 11 | 1.33e-09 | -1.11 |
| | | (1.63e-08) | (92.8) | | (1.41e-09) | (1.19) |
| | Location = Interstate | -2.18e-10 | -31 | Latent Dimension 12 | -5.56e-10 | -0.0416 |
| | | (7.28e-09) | (153) | | (7.57e-10) | (3.05) |

Table A4: Estimated Consumer Heterogeneity Demand Parameters

# A Understanding the Characteristics and Preference Space: Supplementary Evidence

In this section, we provide supplementary evidence to Section 4 that our recovered latent preferences and product characteristics from search data yield a sensible characterization of the market for hotels.

We can assess whether our estimated preferences are sensible by examining the correlation within-consumer of their preferences over various characteristics of hotels. In Figures A4 and A5, we plot the correlation matrix of preferences over characteristics for each of the two estimated models. In Figure A4, where we estimate preference heterogeneity only over observable characteristics, we find that preferences for location types are strongly negatively correlated, while tastes for meeting space and size category (number of rooms) are positively correlated. The former correlation represents a preference for the overall spatial location of hotels, while the latter represents an overall taste for large hotels both in guest capacity and ability to host large business conferences.

Second, we use an off-the-shelf method for classifying products, a top-$k$ classifier, to see if hotels that are close in the latent attribute space also share observable characteristics. Specifically, given a candidate hotel $j$, we examine the characteristics of hotels that are the top-10 closest in euclidean distance to $j$ in the latent characteristic space:

$$d(j,k) = ||\gamma_j - \gamma_k||_2 = \sqrt{\sum_{l=1}^{L}(\gamma_{j,l} - \gamma_{k,l})^2} \tag{A1}$$

where $\vec{\gamma}_j$ is the 12-d vector of latent characteristics learned in the latent characteristics only model. We do this for all hotels $k$ within the same STR-defined geographical market as $j$. We then see what share of hotels classified as close according to the top-10 classifier share the same brand, management company, ownership company, Parent company, and how close they are in miles to the candidate hotel $j$. We perform this exercise for all hotels in our dataset, then average across the classifiers for all $J$ hotels in our dataset to see if, on average, the hotels close in unobserved characteristics share observable attributes more often than those that are classified as far away in the latent characteristic space.

Table A5 plots the result of the top-10 classifier for our targeted characteristics. Observations differ across target variables because not all hotels have a management, owner, parent company, or brand (e.g. independent hotels have no parent company). We also use as a comparison the results from a top-10 classifier based on the observable characteristics of hotels, to benchmark our results.[24] We find that in general, hotels closer in latent characteristic are more likely to share supply-side characteristics such as brand/chain affiliation, compared to a random hotel in the same market. At the same time, the classifier based on observable characteristics performs better for all characteristics, except distance. To an extent, this is not surprising since some characteristics (such as class/price segment) are defined at the brand level, and many hotel chains implement uniform characteristics across their locations.

[24]We use the Mahalanobis distance on observable characteristics to standardize the scale of each characteristic. Because latitude and longitude is included in our input observable characteristics, we exclude it when classifying for the distance metric.

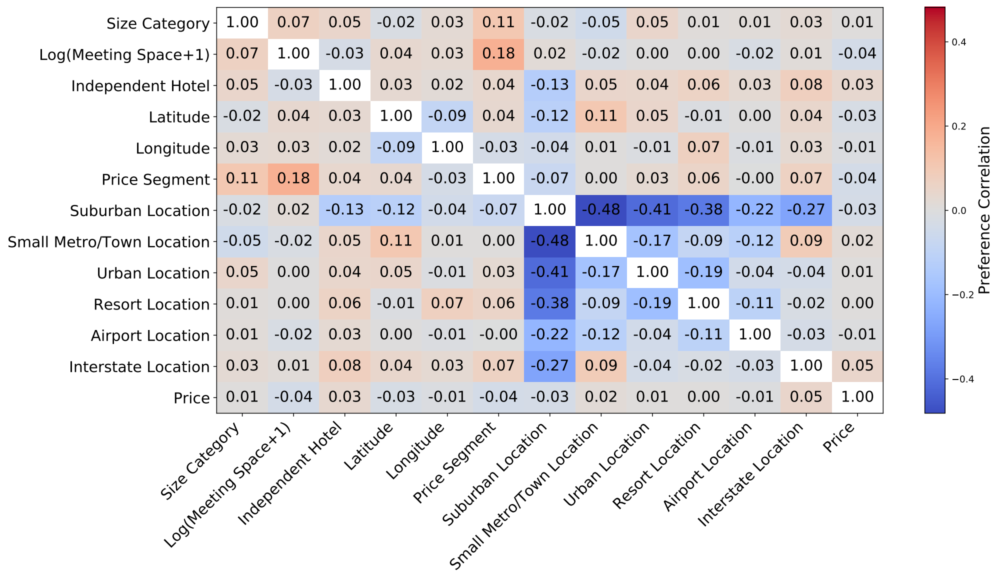



Figure A4: Correlation Matrix of User Search Preferences: Observable Characteristics Model

| | Baseline All Hotels In Market | Top-10 Classifier Based on Characteristics Latent | Observations | # Observations |
|---|---|---|---|---|
| Pr(Same Brand) | 0.029 | 0.039 | 0.093 | 3438 |
| Pr(Same Mgmt Company) | 0.027 | 0.051 | 0.098 | 1690 |
| Pr(Same Owner) | 0.037 | 0.065 | 0.117 | 1202 |
| Pr(Same Parent Company) | 0.133 | 0.164 | 0.251 | 3438 |
| Distance (Miles) | 51.413 | 30.265 | 37.261 | 4218 |

Table A5: Predictive Performance of Top-10 Classifier based on Distance in Latent and Observable Characteristic Space

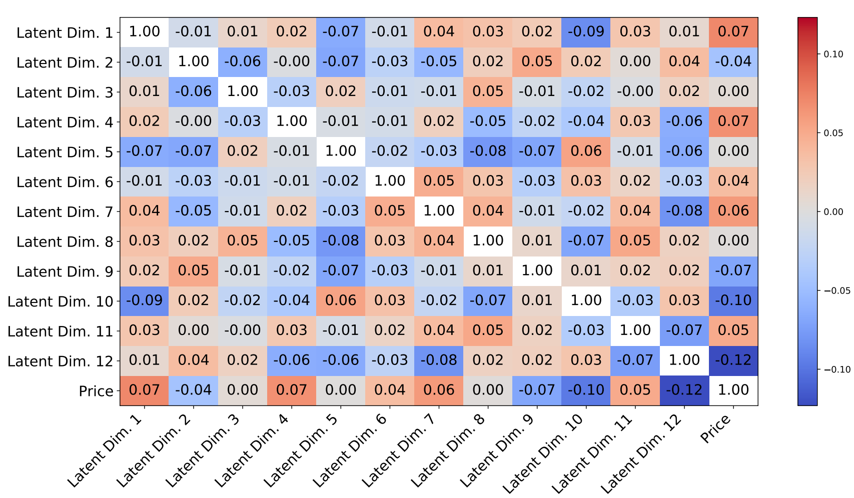



Figure A5: Correlation Matrix of User Search Preferences: Latent Characteristics Model

# B Merger Impact on Hotel Conduct

In order to evaluate how effective our proposed demand model is in capturing actual substitution patterns, we exploit a large merger that occurred during our sample period in the hotel industry that induces plausibly exogenous price changes. On November 16, 2015, Marriott International Inc announced that it would be acquiring the Starwood Hotels company. The merger was completed on September 23, 2016 (Dogru, Erdogan, and Kizildag 2018). After the merger completed, Marriott International became the largest hotel chain in the world. After this transaction occurred, prices in markets with high concentration of Starwood / Marriott hotels changed noticeably. Recall that the brand affiliations of each hotel in our STR dataset is an anonymized ID, so we cannot observe which hotels in our transaction data were Marriott or Starwood affiliates before the merger. Through further coordination with STR, we were able to obtain counts of each hotel brand within each geographical market and class segment, as of December 2015. We use this data to construct an "exposure index" to the merger, $Pr_j(\text{Starwood or Marriott}|\text{Market}_j, \text{Class}_j)$, and measure the effect of exposure to the merger on post-merger prices. The rationale behind this exposure index is that if Marriott/Starwood hotels changed conduct in price-setting after a merger, then hotels belonging to the same geographical market/ class segment as Marriott/Starwood hotels will also respond and change their price-setting behavior. The exposure index then captures both "direct effects" of Marriott-Starwood hotels changing their price behavior due to backend changes in costs and increased market power, as well as "indirect effects" of competing hotels responding to the new price-setting behavior of Marriott/Starwood hotels.

We estimate the effect of exposure to the merger via the following event study specification:

$$\log(p_{j,t}) = \alpha_{j,\text{month}(t)} + \delta_t + \sum_{q=2013Q1}^{2019Q1} \beta_q Pr_j(\text{Starwood or Marriott}|\text{Market}_j, \text{Class}_j) + \epsilon_{j,t} \quad \text{(A2)}$$

where $\alpha_{j,\text{month}(t)}$ denotes hotel×month-of-year fixed effects , to capture time-invariant differences in hotel prices as well as seasonalities in pricing structure, $\delta_t$ is a geographical market $\times$ month $\times$ year fixed effect, to capture common demand shocks occurring in each market-month-year, and $\beta_s$ is the effect of the exposure index in quarter $q$ on hotel $j$. We aggregate the event-study specifications to the quarterly-level due to power concerns given the large number of fixed effects. The control group in this event study are hotels in the same market as Marriott-Starwood hotels but a different class segment. Because consumers have differential demand, hotels belonging to a different class are not as exposed to the changed pricing behavior of Marriott-Starwood hotels after the merger.

Figure A6 plots the estimated event study. The dashed red line represents the time of the merger announcement, while the solid red line represents the date the merger was completed. The blue dots represent the estimated coefficients of the above specification, while the orange dots replace the market-month-year fixed effects with market-month-year-location type (e.g. urban vs suburban hotels in Los Angeles in May 2018) fixed effects.

In general, there do not appear to be strong pre-trends before the merger announcement. We see large price *decreases* following the completion of the merger, which is consistent with discussions surrounding the merger of cost reductions via centralization of sales and customer service operations between the acquiring and target firms (Dogru, Erdogan, and Kizildag 2018).

We use this event study as plausible evidence that the Marriott-Starwood merger led to exogenous price changes by hotels in markets with high exposure to Marriott and Starwood hotels, independent of demand fluctuations. Therefore, measuring the ability of our demand models to accurately predict demand post-merger may serve as a test to whether a demand model augmented with search data may perform "better" in predicting substitution patterns after exogenous price changes.

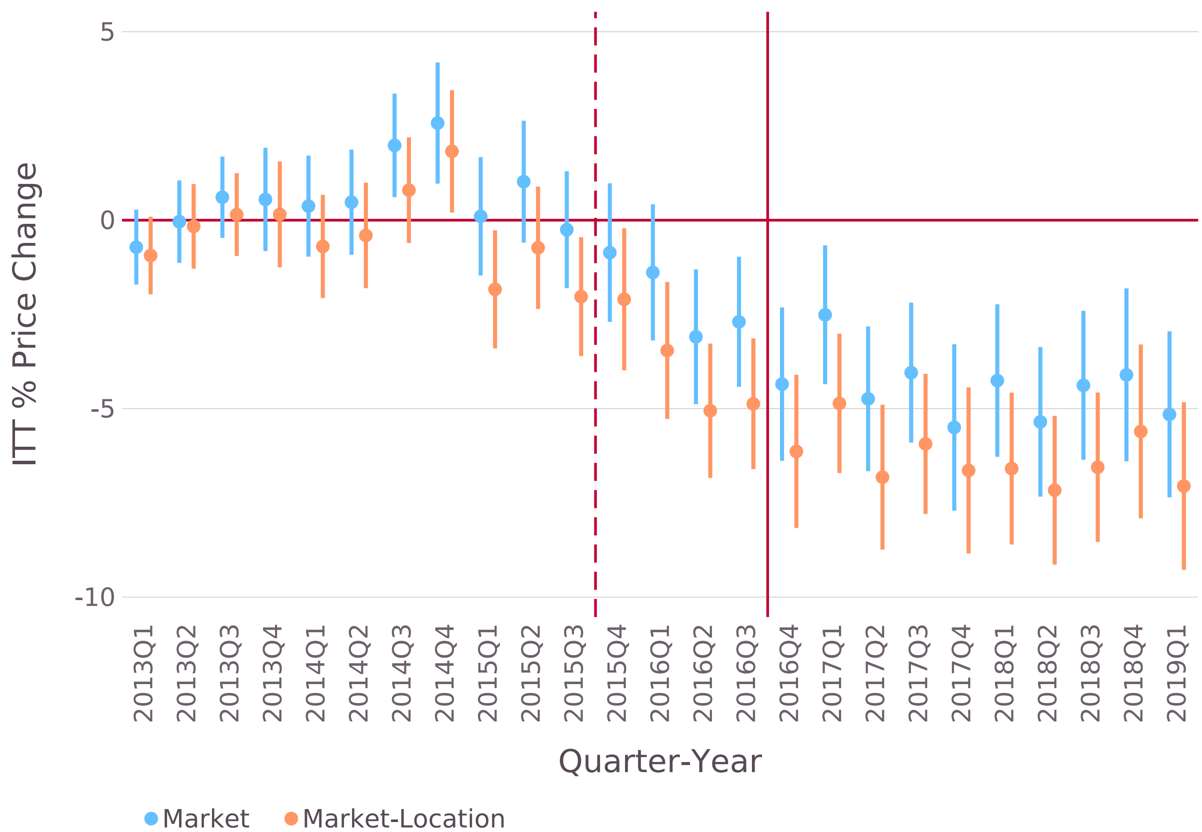



Figure A6: Effect of Marriott-Starwood Merger on Prices in Regions with High Marriott-Starwood Presence

# C Derivation of Choice Market Shares

In this section, we derive the choice probabilities implied by the search and choice model presented in Section 3.2. Recall that expected utility is defined as

$$E[u_{i,j,t}] = \bar{\delta}_{i,j,t} = \delta_{j,t} - \alpha_i p_{j,t} + \vec{\beta}_i \vec{X}_j + \zeta_{i,j,t} \tag{A3}$$

Since realized utility is $u_{i,j,t} = E[u_{i,j,t}] + (1 - \sigma_\zeta)\tilde{\epsilon}_{i,j,t}$, we can divide by $(1 - \sigma_\zeta)$ to obtain a cardinally equivalent expression of utility as $\tilde{u}_{i,j,t} = \bar{\delta}_{i,j,t}/(1 - \sigma_\zeta) + \tilde{\epsilon}_{i,j,t}$, so that utility is expressed as a location shifter plus an i.i.d. extreme value type 1 (T1EV) random variable with scale 1 and location 0. From here, we follow the notation of (Moraga-González, Sándor, and Wildenbeest 2023). Under this re-scaled utility measure, The reservation value under simultaneous search is defined as:

$$\tilde{r}_{i,j,t} = \bar{\delta}_{i,j,t}/(1 - \sigma_\zeta) + H_0^{-1}(\tilde{c}_{i,j,t}) \tag{A4}$$

Where $c_{i,j,t}$ is the user-specific search cost. Let $w_{i,j} = \min(\tilde{r}_{i,j}, \tilde{u}_{i,j})$. This is the minimum of the reservation value associated with searching product $j$, and actual utility derived from the product. Moraga-González, Sándor, and Wildenbeest (2023) show that under the sequential search environment, the optimal search strategy is to choose the product with the highest $w_{i,j}$.[25]

We assume in the main section of the paper that the distribution of search costs same distribution as Moraga-González, Sándor, and Wildenbeest (2023), therefore by Proposition 1 of in that paper, the distribution of $\tilde{w}_{i,j}$ follows a Gumbel distribution with location parameter $(\bar{\delta}_{i,j,t} - \mu)/(1 - \sigma_\zeta)$ and scale 1.[26]

[25] Formally, in (Moraga-González, Sándor, and Wildenbeest 2023), there are multiple products within each "search group", so they find it optimal to search the group with the highest $w$, then choose the best product in that group. In our context, we assume all consumers pay an equal search cost to search one product, so this second stage choice is degenerate, as there is one product in each group.

[26] We express the search location parameter $\mu$ in terms of the original utility space.

Since a T1EV variable with location $\delta$ and scale $\sigma$, multiplied by a constant $c$, is T1EV with location $c \cdot \delta$ and scale $c \cdot \sigma$, it follows that the random variable $(1-\sigma_\zeta)w_{i,j}$ is distributed according to a T1EV distribution with location $(\bar{\delta}_{i,j,t} - \mu)$ and scale $(1 - \sigma_\zeta)$. Since the mapping $f(w) = (1 - \sigma_\zeta)w$ is monotonic, it must also be the case that the optimal search strategy is to choose the product with the highest $(1 - \sigma_\zeta)w_{i,j}$.

Finally, since constant terms are additively separable from a T1EV distribution, we can express $(1 - \sigma_\zeta)w_{i,j}$ as

$$(1 - \sigma_\zeta)w_{i,j} \sim \left(\delta_{j,t} - \alpha_i p_{j,t} + \vec{\beta}_i \vec{X}_j - \mu\right) + \zeta_{i,j,t} + (1 - \sigma_\zeta)\eta_{i,j,t} \tag{A5}$$

Where $\eta_{i,j,t}$ is a T1EV random variable with location 0 and scale 1. Because $\zeta$ is a conjugate to the T1EV by Theorem 2.1 of (Cardell 1997), The distribution of $\zeta_{i,j,t} + (1 - \sigma_\zeta)\eta_{i,j,t}$ is also Gumbel with location 0 and scale 1, which implies that $(1 - \sigma_\zeta)w_{i,j}$ is distributed as T1EV with location $\delta_{j,t} - \alpha_i p_{j,t} + \vec{\beta}_i \vec{X}_j - \mu$ and scale 1. The probability that any particular $(1 - \sigma_\zeta)w_{i,j}$ is largest is given by the max-stability property of the T1EV distribution.

$$Pr(i \text{ choose } j) = Pr((1-\sigma_\zeta)w_{i,j} > \max_{k \neq j}(1-\sigma_\zeta)w_{i,k}) = \frac{\exp(\delta_{j,t} - \alpha_i p_{j,t} + \vec{\beta}_i \vec{X}_j - \mu)}{1 + \sum_{k \in J_t} \exp(\delta_{k,t} - \alpha_i p_{k,t} + \vec{\beta}_i \vec{X}_k - \mu)} \tag{A6}$$

We can then integrate over the distribution of consuemr preferences $\alpha_i, \beta_i$ to get market-level shares, which concludes the derivation.

# D Derivation of Search Likelihood

Consider the normalized utility presented in Section C, where we divide utility by $(1 - \sigma_\zeta)$. Let $\delta_{i,j,t} = \delta_{j,t} - \alpha_i p_{j,t} + \beta_i X_j$, so that $E[\tilde{u}_{i,j,t}] = (\delta_{i,j,t} + \zeta_{i,j,t})/(1 - \sigma_\zeta)$ and reservation values are given by Equation A4. Under simultaneous search, the optimal search strategy is to search products in decreasing order of their reservation value. Our inequalities derived for search behavior require us to compare reservation values across searched and unsearch products; therefore, the relevant probability to consider is:

$$Pr(r_{i,j} > r_{i,k}) = Pr(r_{i,j} - r_{i,k} > 0) = Pr(-\frac{\Delta\zeta_{i,j,k}}{(1-\sigma_\zeta)} - \Delta H^{-1}_{i,j,k} < \frac{\Delta\delta_{i,j,k}}{(1-\sigma_\zeta)}) \tag{A7}$$

Where $\Delta\delta_{i,j,k} = \delta_{i,j,t} - \delta_{i,k,t}$, $\Delta\zeta_{i,j,k} = \zeta_{i,j,t} - \zeta_{i,k,t}$, and $\Delta H^{-1}_{i,j,k} = H_0^{-1}(\tilde{c}_{i,j,t}) - H_0^{-1}(\tilde{c}_{i,k,t})$. Given a guess of the latent parameters, The first component, $\Delta\delta_{i,j,k}$ is constant and serves to simply shift the CDF threshold. So to understand the global properties of the distribution, we ignore it. According to Cardell (1997), the difference of two gumbel conjugate with scale $\sigma_\zeta$ is logistic with scale $\sigma_\zeta$. Therefore, the distribution of $-\Delta\zeta_{i,j,k}/(1-\sigma_\zeta)$ is simply a logistic distirbution with location 0 and scale $\sigma_\zeta/(1-\sigma_\zeta)$. The distribution of $-\Delta H^{-1}_{i,j,k}$ is unknown. In this appendix, we will simulate from it, and evalulate whether the logistic approximation for reservation values (which would be exact if $\Delta H^{-1}_{i,j,k}$ is zero) is a valid approximation.

The search cost distribution used in Moraga-González, Sándor, and Wildenbeest (2023) sets search costs equal to zero with probability $\exp(-\mu)$. This presents a problem, as $H_0^{-1}(0)$ is undefined, and implies a reservation value of $r_{i,j} = \infty$ (e.g. the product is always searched). This also implies that reservation utilities are equal for two products with zero search costs (e.g. they are both always searched). This implies the following probability that one product is searched and the other is not:

$$\begin{aligned} &Pr(\text{search } j, \text{do not search } k) = Pr(r_{i,j,t} > r_{i,k,t}) \qquad \text{(A8)} \\ &= Pr(c_{i,j,t} = 0, c_{i,k,t} > 0) + Pr(c_{i,j}, c_{i,k} > 0) \cdot Pr(r_{i,j,t} > r_{i,k,t} | c_{i,j,t}, c_{i,k,t} > 0) \\ &= Pr(c_{i,j,t} = 0) \cdot Pr(c_{i,k,t} > 0) + Pr(c_{i,j,t} > 0) \cdot Pr(c_{i,k,t} > 0) \cdot Pr(r_{i,j,t} > r_{i,k,t} | c_{i,j,t} > 0, c_{i,k,t} > 0) \\ &= \exp(-\mu)(1-\exp(-\mu)) + (1-\exp(-\mu))^2 \cdot Pr(-\frac{\Delta\zeta_{i,j,k}}{1-\sigma_\zeta} - \Delta H^{-1}_{i,j,k} < \frac{\Delta\delta_{i,j,k}}{1-\sigma_\zeta} | c_{i,j,t} > 0, c_{i,k,t} > 0) \end{aligned}$$

Where the second equality follows from the fact that search costs are independent. Notice that $\mu$ is not a parameter of our search model, so it is fixed during optimization and does not impact the likelihood. Therefore, maximizing the likelihood of our search model is equivalent to maximizing:

$$Pr(-\frac{\Delta\zeta_{i,j,k}}{1-\sigma_\zeta} - \Delta H^{-1}_{i,j,k} < \frac{\Delta\delta_{i,j,k}}{1-\sigma_\zeta} | c_{i,j,t} > 0, c_{i,k,t} > 0) \tag{A9}$$

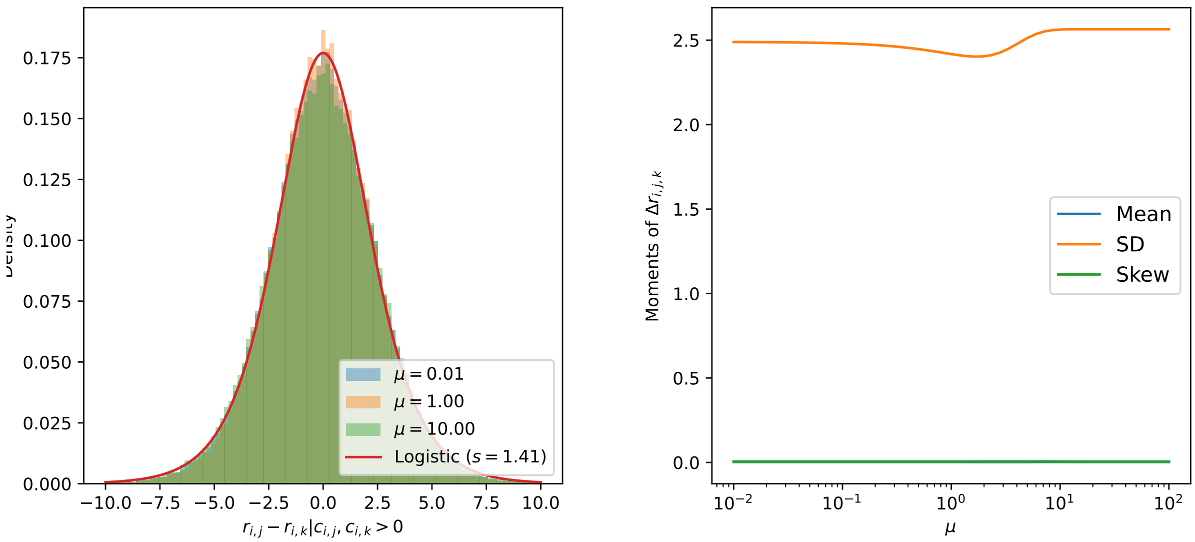
Since $\zeta$ is independent of search costs, its distribution is unchanged, conditional on $c_{i,j,t} > 0$. The CDF of $H_0^{-1}(c_{i,j})$ conditional on $c_{i,j} > 0$ is:

$$\begin{aligned} F_H(z|\mu) = Pr(H_0^{-1}(c) \leq z|\mu, c > 0) &= Pr(c \geq H_0(z)|\mu, c > 0) \\ &= 1 - Pr(c \leq H_0(z)|\mu, c > 0) \\ &= 1 - \frac{S(H_0(z)|\mu) - S(0|\mu)}{1 - S(0|\mu)} \\ &= \frac{F_\epsilon(z+\mu) - F_\epsilon(z)}{(1-F_\epsilon(z))(1-\exp(-\mu))} \end{aligned} \tag{A10}$$

Where $F_\epsilon(x) = \exp(-\exp(-x))$ is the CDF of a standard T1EV distribution. This suggests the search cost component of the reservation values, $H_0^{-1}(c_{i,j,t})$, is similar in shape to a (scaled) truncated T1EV distribution. We sample from this distribution by taking a uniform sample $U \sim [0,1]$, then finding the nonlinear root $z$ satisfies $F_H(z|\mu) = U$. We then use these simulated draws to sample from the distribution of reservation values (conditional on positive search costs).

Figure A7 displays simulations of reservation values across values of $\mu$, the location of search costs. For this exercise, we normalize the scale of $\Delta\zeta$ to be 1 (e.g. $\sigma_\zeta = .5$), so that values of $\mu$, which index the search cost component of reservation values, are relative to the dispersion of $\zeta$. In Panel (a), we plot the distribution of $\Delta r_{i,j,k}$, the difference in reservation values across two products, conditional on positive search costs, for three values of $\mu$ (100,000 draws each). For comparison, we also show the PDF of a logistic distribution with identical variance with $\mu = 10$. The logistic approximation is quite accurate, irrespective of the choice of $\mu$.

In Panel (b), we show the moments of the simulated $\Delta r_{i,j,k}$, across values of $\mu$. The mean and skewness are close to zero across the parameter space. The variance increases




(a) Simulated $\Delta r_{i,j,k}$

(b) Moments of Distribution across $\mu$

Figure A7: Simulated search cost Components and reservation values

siightly $\mu = 5$, but is mostly constant, even up to implausibly large values of $\mu$ such as 100. It appears that a logistic distribution with scale $\approx s = 1.4$ is a good approximation to the distribution of $\Delta r_{i,j,k}$, irrespective of the value of $\mu$. This suggests that our logistic approximation will be accurate even for extreme values of search costs, and there will be little loss of information in our usage of a quasi likelihood.