

Using Search Data to Crowd-source Unobserved Substitution Patterns for Demand Prediction

Mehtab Hanzroh

September 28, 2024

Abstract

Many demand models rely on the characteristics-space approach to represent products and estimate consumer preferences. A practical limitation with the approach in some markets is that if demand-relevant characteristics are not observed, the substitution patterns the model predicts are unreliable. To address this limitation, this paper proposes a method of learning substitution patterns directly from search data. The approach is to treat the sets of products that consumers search for as their revealed consideration set, and measure product substitution between a pair of product by their frequency of co-searches across all consumers' search sets. This substitution measure can then be mapped to vectors of latent characteristics representing each product. I validate the latent characteristics by using them as an input to a simple predictive demand model applied to data on online shopping at a large UK eCommerce platform. The aim is to predict which product a consumer will purchase given the set of previously searched products, as in a recommender system. I find that representing products with latent characteristics leads to improvement in prediction performance. These findings are supported by replicating the empirical analysis within a Monte Carlo simulation of consumer search.

I would like to thank Robert Clark for his excellent guidance. I also thank Andrew Steck, participants at CEA 2023, seminar participants at Queen's University and members of the IO Working Group at Queen's University for their invaluable comments and suggestions. I am a graduate student at Queen's University. This work is part of my thesis. All mistakes are my own. Please contact m.hanzroh@queensu.ca for correspondence.

Introduction

Consumers navigating online marketplaces typically search a variety of substitute goods before making purchases. The sequence of these searches, captured through clickstream data, provides a window into the decision-making process of consumers, revealing a “search set” of products considered before a purchase.¹ Therefore, one way to infer consumers’ substitution patterns is to analyze the patterns of products that consumers search for before making a purchase.

Understanding substitution patterns is imperative in order to, for example, evaluate the effects of price changes or changes in market structure on consumer choices. The status quo approach, following the ideas of Lancaster and McFadden (Lancaster (1966), McFadden (1973)), is to assume consumers have preferences over the characteristics of products. With this approach substitution between products is dependent on proximity in characteristic-space; products with similar characteristics are closer competitors since consumers value their characteristics similarly. However, a key practical concern in many markets is whether easily observable product characteristics indeed determine demand for a product. If demand-relevant characteristics are not observed, substitution patterns derived using a characteristic-space approach will be unreliable.

This paper proposes a novel approach to estimating latent characteristics of products learned from search data that can be used to augment models following the characteristic-space approach. The key rationale of the approach is to treat each consumer i ’s search set, S_i , as their revealed consideration set. Since consumers typically search for products that compete with the product they purchase, the latent characteristics learned from these search data reflect *all* consumers’ underlying substitution patterns.

The main contribution of the paper is the approach taken in mapping search data to latent characteristics. The search data come from a multicategory eCommerce platform, with a large number ($\approx 60,000$) of products. Thus, S_i may include products across different product categories, for instance because consumers browse complementary products. In addition, the approach needs to be scalable to allow for estimation of latent characteristics for many products. I take a simple reduced-form approach. I compute a simple pairwise index of substitution based on the co-search frequency of a pair of products. Empirically, I find that consumers are more likely to search for substitute products, since most search sets result in the purchase of only one product.² Thus, in this paper I do not try to explicitly measure complementarity, but the approach could be easily extended by considering complementarity measures in addition to substitution. Using this index, I use the t-distributed Stochastic Neighbour Embedding (t-SNE) method (van der Maaten and Hinton (2008)) to compute a vector representation of each product that rationalizes the substitution index, and take these representations as each product’s latent characteristics. The index used is a pairwise measure of distance between products that measures pairwise substitution, and the t-SNE method computes corresponding latent characteristics such that distances based on the latent characteristics match the substitution index.

¹These products constitute a subset of the consumer’s broader consideration set.

²In addition, most consumers (>80%) search products within one specific category. Search sets that result in multiple purchases (<1%) are dropped from the analysis. Nevertheless, other search datasets, such as those for market baskets may display higher rates of cross-category search and purchase.

I demonstrate the efficacy of the approach in a few ways. First, in a data-driven manner, I augment a sequential probabilistic demand model following the characteristic-space approach which aims to predict which product j consumers will purchase given their previously searched products up to time t , S_{it} . In this model, the probability of purchasing j depends on the interaction between j and each $j' \in S_{it}$. That is, each previously searched product affects the probability of purchase. Empirically, since consumers tend to purchase a close substitute of their previously searched options, the interactions should reflect pairwise substitution patterns. I test whether representing products using the learned latent characteristics improves predictions, so that interactions between j, j' are better determined by latent characteristics than observables which better reflect substitution patterns. I show that using latent characteristic representations of products in addition to observables leads to better predictions of purchase choices.

I also verify that the latent characteristics encode demand-relevant product characteristics. The products that consumers search for are jointly determined by their own preferences and the platform’s search algorithm. I show that the co-searches are not entirely determined by products that are searched close together in time, so that the co-search rate picks up more than just the search algorithm’s recommendations. This fact is also demonstrated visually, by showing that distances based on the latent characteristics do not create tight clusters of products separated by the observed categories alone.

This approach is useful for analyses on markets where demand-relevant characteristics are not easily obtainable, such as the market for books (Hong and Shum (2006)), movies, or even with online search data such as the popular comScore web-browsing data where researchers need to collect product characteristics (Bronnenberg et al. (2016), Shiller (2020)). Additionally, the approach easily scales to large search data.

The predictive demand models are estimated using a clickstream dataset from a UK e-commerce platform, that contains users’ sequences of clicks - and purchases, if any - onto different product pages. Using these data I estimate demand-prediction models based on multinomial-logit and LightGBM (Ke et al. (2017)), a cutting edge machine learning method. Each model uses different measures of substitution as outlined above and their prediction performance are compared. I find that model rankings are consistent across both classes of models.

To assess each model’s performance, I compare mean log-likelihoods across all observations in and out-of-sample. The latent model achieves a log-likelihood of -0.1628 in-sample and -0.1627 out-of-sample. In comparison, the baseline characteristics model achieves a log-likelihood of -0.1670 in-sample and -0.1679 out-of-sample, illustrating that the latent model’s performance is comparable. Additionally, a combined model using both observed and latent characteristics to represent products achieves a log-likelihood of -0.1612 in-sample and -0.1603 out-of-sample. These results suggest that the approach based on co-search patterns picks up demand-relevant characteristics of products that are not easily observed, especially since the combined model shows a further performance improvement. Thus when demand-relevant product characteristics are not available, using the JC measure to inform substitution is effective in building predictive demand models. Since the latent model performs comparably to the characteristics model, this suggests that the JC measure

can be used to effectively crowdsource substitution patterns even in the absence of demand-relevant characteristics. Nonetheless, these approaches can be combined as in the combined model, so that when some demand-relevant characteristics are available incorporating the JC measure into prediction models can still lead to improvements in performance.

To support the findings from the empirical analysis, I conduct a Monte Carlo simulation. In each period in the simulation setting, consumers are faced with the decision to either purchase a product in the product space or an outside option. I simulate consumer search sequences using a sequential search model in the style of [Weitzman \(1979\)](#). The DGP is designed to generate substitution patterns that depend on both observable and unobservable characteristics. Thus, the simulation setting is one where the researcher observes demand-relevant characteristics. Once these data are simulated, I estimate demand-prediction models identically to the empirical analysis. In this simulation, the latent model performs significantly better in prediction relative to the characteristics model, even when observing demand relevant characteristics. The product space is smaller in the simulation than in the main empirical analysis, so that co-searches are more common across all products. This suggests that the latent characteristics can perform even better in other empirical settings where co-searches are common. This may be settings where researchers or platforms have access to larger data, or cases with smaller product spaces.

Related Literature

This paper contributes to the literature on learning demand-relevant latent characteristics, which are often computed through embedding algorithms ([Rudolph et al. \(2016\)](#), [Liang et al. \(2016\)](#), [Barkan and Koenigstein \(2016\)](#)). The effectiveness of these latent characteristics hinges on their capacity to accurately represent underlying consumer preferences. However, these characteristics may not adequately capture the impacts of items already in a consumer’s basket, potentially failing to reflect true patterns of substitution and complementarity. This issue is particularly prevalent in online market settings where essential drivers of search like marketing influences are often unobserved. The Joint Consideration (JC) measure introduced in this study addresses this gap by providing a robust method for constructing latent characteristics that represent all consumers’ substitution patterns effectively.

The revealed preference approach used to compute the JC measure for measuring substitution is closely related to the work by [Armona et al. \(2021\)](#), who also compute latent characteristics using search data by computing an embedding. Their embedding relies on a revealed preference approach; products that consumers search have higher utilities than products that were not searched resulting in a revealed preference inequality for each pair of searched and unsearched products. This approach could lead to dimensionality issues due in a setting with a many products, and in comparison the JC measure is very simple to compute. [Magnolfi et al. \(2022\)](#) take a different empirical approach to compute embeddings by collecting survey data in which respondents reveal their beliefs about cereals they perceive as similar in the form of triplet

comparisons, which they apply to a triplet embedding algorithm (van der Maaten and Weinberger (2012)). They use their computed embeddings as latent characteristics in a demand model, and show that using these latent characteristics outperforms using observed characteristics in predicting aggregate shares.

The use of repeated decisions of consumers in identifying substitution patterns can also be seen in the broader demand estimation literature. In the marketing literature, panel-data models of demand often include estimation of latent characteristics jointly with the parameters of the demand model (Elrod (1988), Elrod and Keane (1995), Keane (1997)). Identification of the latent characteristics, which in these models represent market structure, comes from consumers who switch from one product to another over time implying that they are substitutes. Since Berry et al. (1995), the characteristic-space approach has been more widely used in the empirical industrial organization literature, and relies on consumer preferences over product characteristics to determine substitution patterns. Within this framework, Berry et al. (2004) show how to use repeated purchase data to better identify substitution patterns. More recently, the marketing literature has devoted much effort in modeling consumer choices using search data, where a consumer’s search process is explicitly modelled (De Los Santos et al. (2017), Ke and Villas-Boas (2019), Ursu et al. (2020), Moraga-González et al. (2023)).³ In addition to the papers referenced above, Kim et al. (2011) develop a method to estimate latent characteristics using search data and in addition estimate preferences over these characteristics in a search model. As their results show, since most search models also follow the characteristic-space approach in modelling utility, the use of latent characteristics leads to improved model fit. My paper combines these ideas by estimating an easy-to-compute measure using search data that represent consumers’ underlying substitution patterns and are explicitly included as explanatory variables (latent characteristics) in a predictive demand model.

A few papers have used measures similar to the JC measure in other contexts. Kumar et al. (2020) apply a heuristic similar to the JC measure in order to study product bundling. Using a clickstream dataset in which they observe users’ shopping baskets they identify complements and substitutes using products purchased together and products considered but not purchased together respectively. Note that the latter idea is used in the definition of the JC measure. They use this method of categorizing complements and substitutes to augment an embedding algorithm and study how to optimally bundle products using the embedding representations of each product. Similarly, Ringel and Skiera (2016) use an asymmetric measure similar to the JC measure to create a market map - a visualization of market structure in the LED TV market. The measure is also similar to measures of ‘Mutual Information’ in the information retrieval literature (e.g. Bouma (2009), Mikolov et al. (2013)). I view the findings in this paper as complimenting the literature’s results by empirically investigating the relevance of search patterns in identifying substitution patterns for models following the characteristic-space approach.

The literature also shows uses of measures similar to the JC measure for the case of repeated purchases. Atalay et al. (2023) use the likelihood of a household ever purchasing both of a pair of products in the

³This is a rapidly expanding literature. The interested reader should see Honka et al. (2019) for a survey of the search and consideration set literature and Ursu et al. (2023) for a recent survey of the sequential search literature.

Nielsen market basket panel as an input to an agglomerative clustering method. They then treat each computed cluster as a nest within a nested-logit demand model. To the best of my knowledge, my paper is the first to use this kind of repeated co-search measure for the purpose of a predictive demand model that takes past purchase behavior into account. [Magnolfi et al. \(2022\)](#) use a measure similar to the JC measure based on repeated purchases as an input to the t-SNE method. They show that using these embeddings as latent characteristics results in a better fitting demand model than using only observed characteristics, when predicting aggregate shares in the ready-to-eat breakfast cereal market.

In addition to [Atalay et al. \(2023\)](#), other models in the emerging flexible demand estimation literature employ search data to better inform substitution patterns. [Donnelly et al. \(2024\)](#) and [Amano et al. \(2022\)](#) estimate a two-step consideration-then-choice demand model. The former paper estimates latent factors of utility that drive substitution while the latter constrains the correlations in a component of utility to rationalize product co-search patterns. [Dotson et al. \(2018\)](#) and [Dotson et al. \(2024\)](#) instead estimate a multinomial probit choice model and allow the utility covariance structure to depend on the similarity between products.

In order to evaluate the efficacy of the JC measure, I use it for explanatory variables in a demand prediction model and observe the resulting demand prediction performance. A number of papers have investigated models that fit and perform well in predicting demand. [Bajari et al. \(2015\)](#) apply several methods to estimate demand for salty snacks from panel data for a grocery store chain. They compare the fit of these methods as measured by RMSE and find that machine learning methods such as Random Forest show substantially superior predictive accuracy compared to traditional statistical methods such as linear/logistic regression. For brevity, I use just two methods for each choice of explanatory variables; logistic regression and a very closely related algorithm to the [Breiman \(2001\)](#) Random Forest algorithm - LightGBM developed by [Ke et al. \(2017\)](#). LightGBM and Random Forest are both aggregated tree-based methods but differ in the way trees are constructed. LightGBM uses a more computationally feasible method in constructing trees, which works well given that I use a very large dataset. I describe the details of this method in the Methodology section. I show that the effective performance from using the JC measure holds for both estimation methods. I also evaluate the models using two common measures of model fit in the machine learning literature; the Area Under the Curve (AUC) of each model’s Receiver-Operating-Characteristic (ROC) plots and an aggregated Cross-Entropy Loss⁴. Both measures of fit suggest the same ranking of models as the misclassification rate.

There is a vast literature in predicting consumer choices for eCommerce. As clickstream data has become more accessible to researchers, a number of recent papers have developed predictive demand models that can use consumers’ past search choices to predict purchases. Recently in the marketing literature, such models have been developed by improving upon existing machine learning methods ([Gabel and Timoshenko \(2022\)](#), [Jacobs et al. \(2016\)](#)). These recent models allow for consumer preferences over goods in different categories

⁴Both measures of fit are explained in the Methodology section.

by computing latent characteristics, often using embeddings, to represent all products in one characteristic-space. The latent characteristics estimated using the JC measure in this paper accommodates for products in different categories to affect purchase probabilities in this way. Earlier models often included additional parameters to capture cross-category relationships (Manchanda et al. (1999), Russell and Petersen (2000)). Ruiz et al. (2020) develop a model to predict the next item a consumer will choose for their shopping basket given the set of items already chosen. The effect of products already in a basket on future choices is modelled through a term capturing the effect of characteristics of products already in the basket. In contrast, the model of Shiller (2020) measures the impact of past behavior on the probability of purchase using product fixed effects.

While not directly focusing on price discrimination, this paper highlights the potential of using clickstream data to develop predictive demand models, which is vital for strategies like personalized pricing. Dubé and Misra (2023) develop a model of personalized pricing by estimating a demand equation using machine learning methods to deal with dimensionality issues. Shiller (2020), Smith et al. (2023), Waldfogel (2015), and Zhang et al. (2014) estimate individual level demand as a function of consumers’ observables and past purchase behavior, and find large improvements in predicted revenues particularly when taking past purchase behavior into account. By demonstrating the capability to extract demand-relevant latent characteristics from clickstream data alone, this paper provides valuable insights for online retailers looking to optimize demand prediction without extensive product characteristic data.

Data

I use a substantial clickstream dataset detailing consumer product browsing behaviors on a UK eCommerce store during October and November 2019, available via the “Kaggle.com” data repository (Kechinov (2019)). The specific identity of the eCommerce platform is undisclosed; however, a histogram of product categories, shown in Figure 4 in the Appendix, provides insight into the platform’s product offerings. The dataset encompasses detailed records of user interactions with product pages, classifying each interaction as one of four event types: purchase, view, add-to-cart, remove-from-cart. Additionally, product characteristics data such as price, main-category and sub-category labels, and brand are included. The latter three are used as discrete characteristics, and log-price is used as a continuous characteristic.

The data are then transformed to construct search sets. I make use of a “session_id” variable; a user’s events with the same session-id occurred during one continuous period of activity on the platform. I combine multiple sessions from the same user into the same search spell under the following conditions: (a) the end of the first session and the start of the next session are at-most 1 day apart, (b) the first session did not result in a purchase, and (c) a majority of products in both sessions share a main-category. Note that this potentially generates multiple search spells for each consumer. Next, I split search spells by purchase events. Events up to and including a purchase event are treated as one search sequence, and events in the same

session after a purchase event may be combined with events from subsequent sessions if the conditions above are satisfied. Lastly, if this process leads to a search sequence that lasts longer than one week, it is dropped entirely from the analysis.

A natural challenge with clickstream data for predictive modeling is that some observations capture users at a late stage in the buying process. Typically, a user’s search history will show a view followed by an add-to-cart event shortly before a purchase. Directly using this sequential information as explanatory variables can increase the model’s prediction accuracy; however, it does not translate effectively into real-world applications. For instance, if this model were used in real-time to trigger product advertisements, these would likely reach users who have already decided to buy the product and are merely completing the necessary purchasing steps. This scenario would render the advertisements redundant, potentially wasting marketing resources. To circumvent this limitation, I ensure that the last event in the sequence of events used to construct the search sequences is associated with a different product than the outcome event.

Methodology

Predictive-demand models play a crucial role in marketing by leveraging historical consumer behavior to forecast future purchases. For instance, marketers often aim to predict which product j a consumer i is likely to purchase, based on a set of products S_i they have previously searched for the purpose of targeted advertising or product recommendations.

The models considered in this paper follow this line of thinking. As mentioned previously, search sequences in the dataset conclude with the purchase of one product, suggesting that the products within each sequence are likely substitutes. Therefore, to effectively measure the interaction of each previously searched product $j' \in S_i$ with the purchase choice j , the substitutability of j, j' should be considered. One can do this with observed characteristics by considering proximity in observed characteristic-space. This kind of model serves as the benchmark characteristics model in this paper.

However, this paper explores an innovative approach by estimating latent characteristics and considering substitution patterns through proximity in latent characteristic-space. These latent characteristics are derived by mapping pairwise product co-search patterns, referred to as Joint Consideration (JC), to vector representations of products. The construction of this Joint Consideration (JC) measure, the mapping to latent characteristics, and their application in predictive models are detailed in the following sections. This approach not only leads to high predictive accuracy but also provides deeper insights into how substitution patterns can be learned from consumer search patterns.

Latent Characteristics

The primary contribution of this paper is the use of search data to learn demand-relevant latent characteristics. The strategy is to first compute a simple pairwise measure of substitution derived from *all* consumers’

search sets. Then, I estimate latent characteristics for each product that represent the pairwise substitution measures. I refer to the substitution measure as the "Joint Consideration" (JC) measure. This measure leverages the entire dataset to crowdsource a broad understanding of consumers' perception of substitute goods. The rationale is that if products j and j' are frequently in the same search sets among all users in the data, then many users perceive them as competing products. So when evaluating whether a consumer will purchase j , having observed the user browse j' on the platform already should be informative of the choice the user will make when considering j . This measure is similar to measures of "Mutual Information" in the information retrieval literature, and similar to what is used by [Ringel and Skiera \(2016\)](#) to visualize market structure.

A consumer i 's search set S_i is defined as the collection of products associated with a "view" event before a "purchase" event. The JC measure is calculated by first identifying all observed search sets. Note that the dataset's structure ensures that the purchased product is also tagged as viewed, thereby including it in the search set.

The JC measure for a pair of products j and j' is given by the following:

$$JC(j, j') = \frac{\sum_i G_{ij} \times G_{ij'}}{\sum_i G_{ij} + G_{ij'} - H_{ijj'}}, \quad (1)$$

where $G_{ij} = 1(j \in S_i)$ and $H_{ijj'} = 1(j, j' \in S_i)$.

This generates a symmetric matrix \mathbf{JC} . I next change this to its equivalent pairwise distance matrix, $1 - \mathbf{JC}$, and use it as an input to the t-SNE ([van der Maaten and Hinton \(2008\)](#)) embedding method. This algorithm takes a distance measure as an input, and computes a low-dimensional vector representation of each product such that the distances of products in the low-dimensional space reflect the inputted distance measure, thereby reducing the dimensionality of data while maintaining the relative distances between points.⁵ The algorithm is as follows:

Define the similarity between j, j' as

$$p_{jj'} = \frac{\exp(-d(j, j')^2/2\sigma^2)}{\sum_{k \neq l} \exp(-d(k, l)^2/2\sigma^2)}.$$

Let x_j be the l -dimensional vector representation of product j . Define the similarity between j, j' in the low dimensional space as

$$q_{jj'} = \frac{\exp(-\|x_j - x_{j'}\|^2)}{\sum_{k \neq l} \exp(-\|x_k - x_l\|^2)}.$$

Choose x_j 's that minimize the Kullback-Leibler divergence:

$$KL(P||Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}}.$$

⁵For complete details of the method, the interested reader should see [van der Maaten and Hinton \(2008\)](#).

I use $d(j, j') = 1 - JC(j, j')$ as the input distance measure in this algorithm. To gain intuition behind the method, note that the KL-divergence determines how close the similarities q_{ij} that are based on the estimated embeddings are to the similarities p_{ij} that are based on the JC measure. When these two similarities are close, the contribution of pair i, j to the overall divergence is close to zero. The similarity p_{ij} is based on the inputted distance measure, $1 - JC(i, j)$, while the similarity q_{ij} is based on the euclidean distance between the embeddings x_i, x_j . Putting these pieces together, the algorithm finds embeddings x_j for all j such that measuring similarity by comparing x_j 's matches the JC measure. The optimization problem is easily solved by stochastic-gradient descent, even in my case with nearly 60,000 products.⁶

A practical concern is that the JC matrix may be sparse. There are 57,048 products in the data and so if too few search sets are observed, there may be many pairs of products that are not observed to be co-searched that might be with more data. Thus the resulting latent characteristics may be unreliable. This necessitates using a large portion of the data to construct **JC**. Specifically, **JC** is constructed using the set of observed search sets from roughly half of the data; the first month of data from October. More precisely, only search sequences that end in October are used, which allows for a sufficiently dense **JC** matrix in my testing.⁷ The models which use the JC measures are then restricted to be estimated on only data from November. This also ensures that data from the latter month are not used to construct the measure which in turn is used to predict purchases from the same month, which cannot occur in practice. The characteristics model, which does not use the measure, is allowed to be estimated on the data from both months. This allows a fair comparison in the sense that the models which do not use the JC measure do not need to set aside a portion of the data to construct an explanatory variable.

Table 1: JC Measure Summary Statistics

Statistic	Mean	Std	25%	50%	75%	90%	99%	Min	Max
Unconditional JC	0.007	0.021	0.001	0.002	0.006	0.015	0.071	0.000	1.000
Within Search Set	0.010	0.023	0.001	0.002	0.007	0.027	0.115	0.000	1.000
Within Search Set (Weighted)	0.023	0.041	0.002	0.009	0.029	0.063	0.160	0.000	1.000

Table 1 provides summary statistics for the JC measure. The first row gives summary statistics for the JC measure for all possible pairs of products. The second row instead restricts the pairs to those that appear at-least once in a search set. Finally, the third row weights the statistics by the frequency with which each pair of products appears in a search set. The JC measure for products within a search set typically are higher than for any two products. Thus, after observing a user's browsing history, it is sensible that a product with a high JC measure is more likely to be purchased than those with a lower JC measure.

⁶I use the t-SNE implementation in the "scipy" Python package.

⁷I find that the visualization patterns in the next section tend to be uninformative when constructing **JC** with fewer and fewer data.

Table 3: Search Sequence Categories

	Mean	SD	Max	Share One
Num. Main-Categories	1.17	0.48	9.00	0.87
Num. Sub-Categories	1.34	0.89	34.00	0.79
Num. Brands	1.89	2.06	72.00	0.65

Table 4: Distribution of JC within categories

Main Category	Mean	SD
Electronics	0.001	0.017
Computers	0.002	0.022
Appliances	0.001	0.017
Construction	0.002	0.032
Auto	0.004	0.042
Furniture	0.001	0.027
Kids	0.002	0.032
Apparel	0.001	0.019
Sport	0.003	0.048
Stationary	0.014	0.109
Accessories	0.003	0.035
Medicine	0.072	0.208

Table 2: Search Sequences Summary Statistics

Statistic	Mean	Std	25p	50p	75p	90p	Min	Max
Number of Events	7.67	12.62	3.00	4.00	8.00	16.00	1.00	933.00
Number of Products	3.39	5.37	1.00	2.00	3.00	7.00	1.00	301.00
Duration (Hours)	18.32	37.33	0.02	0.10	16.52	73.81	0.00*	168.00

* 139,741 sequences only consist of one search or one purchase and have a duration of 0.

Table 2 provides summary statistics on the 1,211,593 observed search sequences from both months of data. Search sequences are typically short; the 75th percentile of the number of events in a sequence is 13, and is 9 for the number of products in a sequence. Most search sequences end in a short time period, with the 50th percentile for search duration at 2.44 hours. Table 3 shows the degree to which search occurs across product categories and brands.

Around 79% of search sequences correspond to search within only one sub-category. This is not surprising since the eCommerce platform’s own search ranking algorithm and product recommendations help shape consumers’ search sets. Thus a natural question to ask is whether the JC measure is similar across all pairs of products within a category, which would suggest that JC does not provide useful information beyond product categories. Table 4 shows the mean and spread of the distribution of JC measures for pairs of products that share a main-category.

JC measures remain small in magnitude even for pairs of products within categories, but there is con-

siderable dispersion relative to the magnitudes of JC measures. As the visualization below further suggests, measuring substitution using JC measures does not tightly cluster *all* products within a category together.

Visualization of Substitution Patterns

To visualize differences in the substitution measures used in the paper, I apply the t-distributed Stochastic Neighbour Embedding (t-SNE) algorithm (van der Maaten and Hinton (2008)) to a measure of distance based on the observed characteristics, and compare them to the embeddings based on the JC measure, in both cases computing 2-dimensional embeddings. When the embeddings are plotted, they often reveal clusters in high-dimensional data, making it ideal for visualizing complex relationships like product substitution analyzed in this paper (van der Maaten and Hinton (2008)).

The observed characteristics are a mix of qualitative and quantitative variables. Therefore, I use Gower’s measure (Gower (1971)) to find the distance between two products as it can accommodate both qualitative and quantitative data as characteristics. If each product has p characteristics indexed by a , the distance between products $x_j = (x_{j,1}, \dots, x_{j,p})$ and $x_k = (x_{k,1}, \dots, x_{k,p})$ is given by the following:

$$D_{\text{Observed}}(j, k) = \frac{1}{p} \sum_{i=1}^p d_a(x_j, x_k), \quad (2)$$

$$d_a(x_j, x_k) = 1(x_{j,a} \neq x_{k,a}) \text{ when } a \text{ is qualitative}, \quad (3)$$

$$d_a(x_j, x_k) = \frac{|x_{j,a} - x_{k,a}|}{R_a} \text{ when } a \text{ is quantitative}, \quad (4)$$

where R_a is the range of characteristic a in the data. The two-dimensional embeddings when using the JC measure are computed as described in the previous section.

Visual inspection of the positions of points representing each product in Figures 1 and 2 illustrates distinct differences between the two distance measures used. As discussed previously, each of the embeddings aim to identify substitute products as those with a small distance, which in these Figures will be represented by points close in the two-dimensional Euclidean space. Thus, any differences in identifying substitutes for each distance measure will be revealed by visual differences in each two-dimensional embedding.

Some interesting differences between using D_{Observed} and D_{JC} are apparent. First, D_{Observed} results in clusters where products within a cluster are generally within a product category. D_{JC} results in regions of points with products sharing a product category, but these regions are both not as tight, and are not separated as drastically as with D_{Observed} . This suggests that the latent characteristics are picking up on substitution patterns that are not captured by simply comparing product categories and prices. As is shown later, demand prediction benefits from using latent characteristic representations of products with suggests that these substitution patterns that D_{JC} captures come about due to demand-relevant characteristics not captured by the observed characteristics.

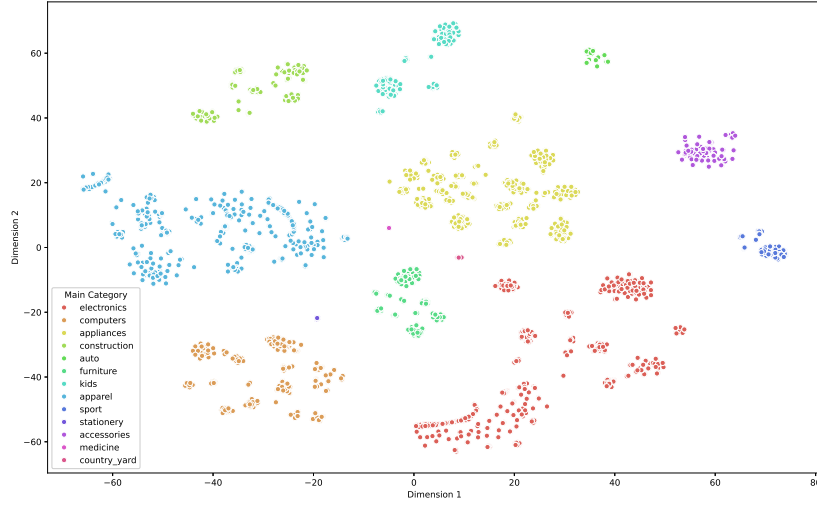


Figure 1: Observed characteristics

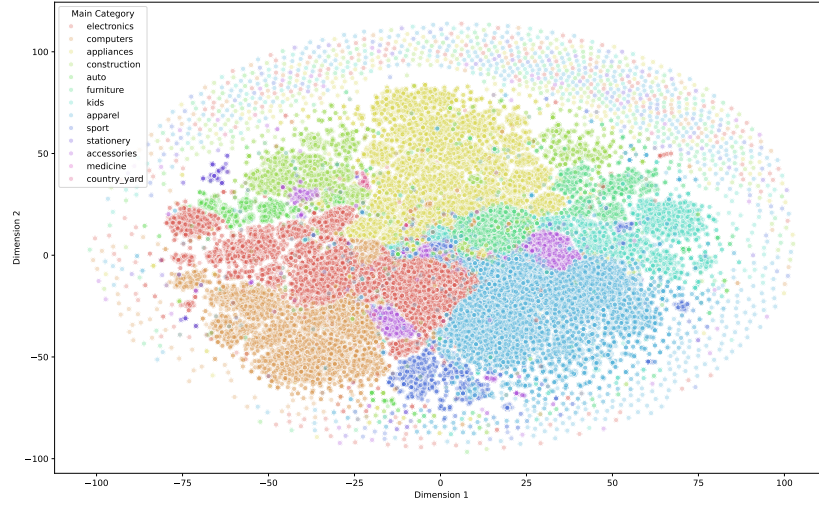


Figure 2: JC Measure

Evolution of Searched Characteristics

Table 5 presents mean values for the observed product characteristics and JC measure, conditioned on the outcome of the previous event. For example, if the most recent event is a purchase, the outcome product shares the main category with the most recent event's product in 60.4% of observations, compared to 68.1% following non-purchase events. Thus, as consumers are browsing products, they substitute away

Table 5: Conditional Means of Explanatory Variables

Variable	Outcome of Previous Event	
	Purchase	Non-purchase
Main Category	0.604	0.681
Sub Category	0.535	0.622
Brand	0.765	0.470
JC	0.358	0.573
The proportion of purchases is 0.065 and the mean price is £374.82.		

from products in the same main category more aggressively after a purchase. The same pattern is true for the subcategory of the product and the JC measure. However, the brand of the product exhibits the opposite trend, suggestive of a persistent brand value effect. This suggests a pattern in search behavior: once a satisfactory product within a category is purchased, search often shifts to a different product type. Equivalently, consumers tend to browse products close in characteristic-space before making a purchase decision.

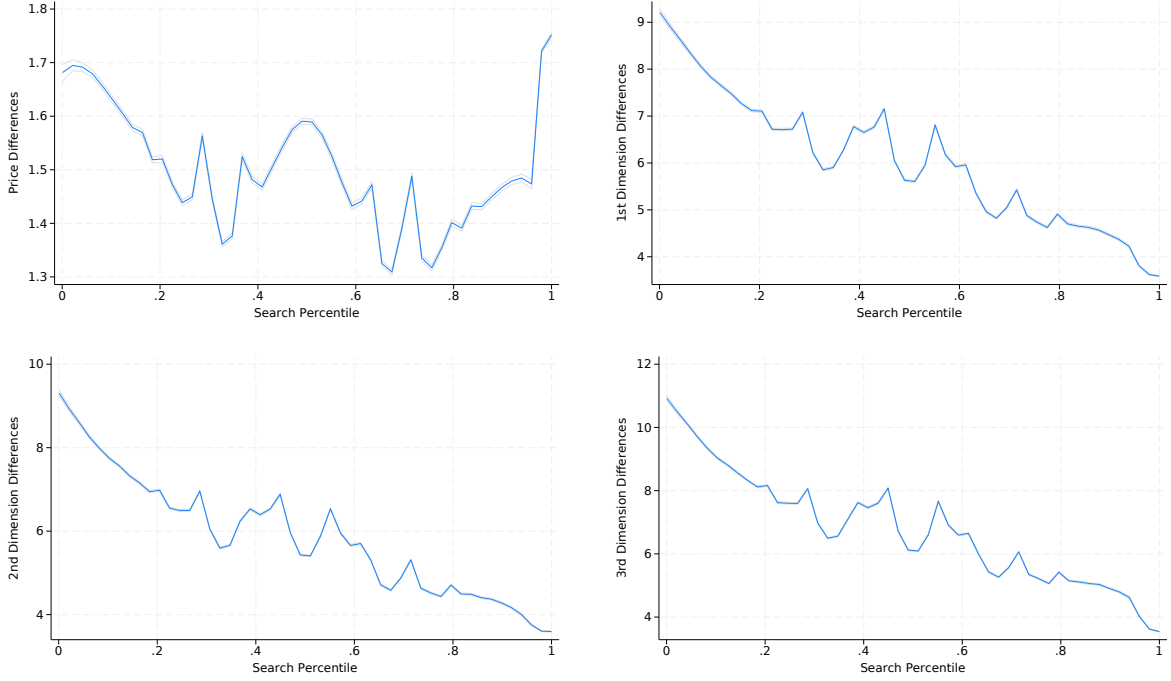
To further investigate, I now show patterns in the characteristics of searched products up to and including the purchased product - products within a search set S_i . Specifically, I test whether a finding from [Bronnenberg et al. \(2016\)](#), or BKM for short, holds when using latent characteristics. BKM find that consumers converge over their search spells in characteristic space to their purchase choice. If the latent characteristics indeed are demand-relevant, the same pattern should hold as well. Otherwise, substitution patterns obtained by comparing latent characteristics would not suggest that consumers browse close substitutes prior to purchase. For this test I use the estimated latent characteristics with $l = 3$, and show convergence in the price and three dimensions of latent characteristics. Figure 3 shows that BKM’s finding holds even when representing products with latent characteristics.

In summary, the most recent events tend to correspond to products that share characteristics with that of the outcome variable, and changes in search behavior depend on whether the consumer has decided to make a purchase. I take this as evidence of the patterns that the latent characteristics pick up - consumers tend to search products that compete with the product they purchase and so latent characteristics of co-searched products should be similar.

Demand Prediction

This section estimates a simple demand prediction model that showcases the ability of the learned latent characteristics to pick up on substitution patterns. The model follows the characteristic-space approach to representing products, denoted x_j , and allows for previously searched products to influence consumers’ purchase probabilities. The three models estimated are differentiated by the characteristics included in x_j :

Figure 3: Coverage in Prices and Embedding Space



This plots a search-level local polynomial regression of differences in characteristics of searched and eventually-purchased products on search percentile. They are estimated using the default tuning parameters in Stata with an epanechnikov kernel. Search percentile is the relative position of each search within its sequence, defined as the ratio of search number divided by total searches in the sequence.

1. **Characteristics Model:** x_j is based on the observed characteristics, and includes dummy variables for each of the three qualitative characteristics (main-category, sub-category, brand) and price.
2. **Latent Model:** x_j is the vector of price and the latent characteristics learned from the search data.
3. **Characteristics and Latent Model:** x_j includes both observed characteristics and latent characteristics.

The exercise is to compare model fit and prediction performance between each of the models so as to judge the ability of latent characteristics to encode demand-relevant information.

The general model to be estimated is a probabilistic model of purchase, or in the language of computer science, a recommender system, and is framed as a multinomial logit model of demand. Specifically, the objective is to determine the product j that the consumer would be most likely to purchase, given the set of M products $S_{t,M} = \{x_{t1}, \dots, x_{tM}\}$ that the consumer has previously searched (but not yet purchased). Note that this is only a subset of the consumer's search set S_i . Let \mathcal{J} denote the set of all products. Formally, we wish to assign a probability of purchase $P(j|S_{t,M})$ for each possible $j \in \mathcal{J}$. It is reasonable to assume that the consumer's choice of j depends on the set of previously searched products since learning about their preferences from previously searched products may inform the consumer about other similar products. For instance, see recent work in the empirical search literature (Hodgson and Lewis (2023)). Motivated by the

scale of the data, I take a more reduced-form approach than writing down a complete structural model of search and purchase. I posit that the consumer's utility from choice j is given by

$$U(j, S_{t,M}) = f(j|S_{t,M}) + \xi_{jt} \quad (5)$$

Here, $f(j|S_{t,M})$ is the deterministic part of utility dependent on the chosen product j and the set of previously searched items $S_{t,M}$, and ξ_{jt} is a logit error term. Thus, purchase probabilities are given by

$$P(j|S_{t,M}) = \frac{\exp(f(j|S_{t,M}))}{\sum_{k \in \mathcal{J}} \exp(f(k|S_{t,M}))} \quad (6)$$

Each product is represented by a vector of characteristics x_j . I assume the functional form of $f()$ is given by

$$f(j|S_{t,m}) = \beta^T x_j + \gamma d(x_j, \tilde{x}_t), \quad (7)$$

where $d(x_j, x_k) = (\sum_a \omega_a (x_{ja} - x_{ka})^2)^{1/2}$ is a weighted measure of distance in characteristic space of products j, k . The weights allow the model to flexibly determine the most important characteristics, a , that determine substitution patterns. It also controls for the fact that different characteristics may have different scales. The products in each search set $S_{t,m}$ are aggregated and represented by the vector \tilde{x}_t , which has the typical element

$$\tilde{x}_{tk} = \begin{cases} Mo(x_{tmk}) & \text{if } k \text{ is discrete} \\ \bar{x}_{tmk} & \text{if } k \text{ is continuous} \end{cases}, \quad (8)$$

where $Mo(x_{tmk})$ denotes the mode of characteristic k across all $m \in S_{t,m}$, and \bar{x}_{tmk} similarly denotes the mean.

The parameter β captures preferences over both prices and non-price product characteristics. I assume the relation between each pair of x_j and $x_{tm} \in S_{t,M}$ is captured by the term $\gamma d(x_j, \tilde{x}_t)$. That is, purchase choices depend on the proximity within characteristic space of the purchase choice and all previously searched products. The parameter vector γ captures the degree to which distance affects utility. Note that the utility from a “No-Purchase” outside option denoted $j = 0$ is normalized to zero.

The inclusion of this distance measure in the utility specification is motivated by theoretical and empirical findings in the literature. Search theory would suggest that purchase options close in characteristic space to past searched products have a higher likelihood of being purchased; consumers are unlikely to purchase a wildly different product to what they have been considering. Empirically, [Bronnenberg et al. \(2016\)](#) show that consumers converge in characteristic space over their search spells to their purchased products. I replicate this finding in Figure 3. Thus, the purchased product is likely to be a close substitute to past searched products. Accordingly, we would expect that γ is negative. This parameter is identified by variation in the distance between purchase choices and previously searched products.

With this framework in mind, the model is a multinomial logit model of purchase that aims to estimate the parameters $\varphi = (\beta^T, \gamma, \omega)$. The model is estimated by maximum likelihood. The main computational challenge in estimation is that the denominator of 6 requires summing over many exponentials. Instead of evaluating the exact log-likelihood, I estimate the model using a lower bound of the exact log-likelihood as in [Titsias \(2016\)](#). Specifically, they show empirically that the parameters that maximize the lower bound also maximize the exact log-likelihood. The advantage of the approach is that the lower-bound is easier to compute.

Note that endogeneity of prices may be present since prices for j and each $j' \in S_{t,m}$ are included in the specification. Thus, the model cannot answer counterfactual questions related to prices.⁸ This is not a problem for the exercise here, since the goal is to judge the improvement in prediction performance from including the latent characteristics in x_j .

Estimation Details

This section describes the implementation of the approximate log-likelihood procedure used to estimate the model. [Titsias \(2016\)](#) show that a lower bound to the purchase probabilities in Eq. 6 is given by

$$P(j|S_{t,M}) \geq \prod_{k \in \mathcal{J} \setminus j} \sigma(f(j|S_{t,M}) - f(k|S_{t,M})), \quad (9)$$

where $\sigma(x) = \frac{1}{1+e^{-x}}$ is the sigmoid function. The exact likelihood is given by

$$L(\varphi) = \prod_{t=1}^T P(j_t|S_{t,M}). \quad (10)$$

Plugging in the lower bound gives

$$L(\varphi) \geq \prod_{t=1}^T \prod_{k \in \mathcal{J} \setminus j_t} \sigma(f(j_t|S_{t,M}) - f(k|S_{t,M})). \quad (11)$$

Finally, taking logs we obtain the lower bound on the exact log-likelihood.

$$\log L(\varphi) \geq \sum_{t=1}^T \sum_{k \in \mathcal{J} \setminus j_t} \log(\sigma(f(j_t|S_{t,M}) - f(k|S_{t,M}))) = \log l(\varphi) \quad (12)$$

I obtain an unbiased estimate of the lower-bound by subsampling observations t and products k over which the inner summation is calculated. Specifically, each evaluation of the log-likelihood samples a set \mathcal{B}_T of observations. Each evaluation of the log-likelihood of an individual observation - the inner summation -

⁸This is partially due to data limitations. I cannot obtain reasonable price instruments with the data at hand. However, with other clickstream data researchers have access to such as the comScore web-browsing panel, or the sophisticated data eCommerce platforms have, the same exercise could be done while estimating a BLP-type demand model.

samples a set $\mathcal{B}_{\mathcal{J}}^t$ of products. The unbiased estimate is given by

$$\log l(\varphi) = \frac{T}{|\mathcal{B}_T|} \sum_{t \in \mathcal{B}_T} \frac{J}{|\mathcal{B}_{\mathcal{J}}^t|} \sum_{k \in \mathcal{B}_{\mathcal{J}}^t} \log(\sigma(f(j_t|S_{t,M}) - f(k|S_{t,M}))). \quad (13)$$

I set $|\mathcal{B}_T| = 10000$ and $|\mathcal{B}_{\mathcal{J}}^t| = 1000$.

Regularization

Since the objective is prediction we must also deal with the issue of overfitting. Logit models in particular are prone to overfitting when the number of parameters is relatively high (Hastie et al. (2001)), which may be an issue for the characteristics and latent model in particular. I employ L2 regularization, also known as ridge regularization, to deal with this issue. L2 regularization applies a penalty term given by the size of the estimated parameters which has the effect of shrinking the parameters closer to zero, thus reducing the number of explanatory variables that have a large impact on the prediction. Thus, it becomes less likely that irrelevant explanatory variables can be used to greatly influence predictions and thus overfit the model. The L2 regularization technique chooses parameters by solving the optimization problem given by the following:

$$\hat{\varphi} = \arg \min_{\varphi} -\log l(\varphi) + \lambda \sum_i |\varphi_i|^2, \quad (14)$$

where λ is a tuning parameter that controls the importance of the L2 penalty term. A higher value of λ corresponds to a more aggressive penalty. I choose the value of λ that minimizes the misclassification rate in the validation sample by grid search over the values $\{0.1, 0.2, \dots, 1\}$ independently for each multinomial logit model.

Gradient Boosted Random Forests

I also validate whether a more sophisticated algorithm from the machine learning literature designed for prediction also benefits from representing products using learned latent characteristics. For instance, one may worry that a model better suited to predictions may be able to flexibly pick up substitution patterns from observed characteristics alone.

LightGBM, developed by Ke et al. (2017), is a very similar estimation procedure to the more popular random forest model of Breiman (2001). The LightGBM model differs in the estimation of a classification tree f_b in step 2 (b) below and the use of gradient boosting. In the Breiman (2001) random forest model, estimation of classification trees is based on the CART algorithm. In the LightGBM model, estimation is based on a “leaf-first” method. The leaf-first method is much cheaper computationally, and generally leads to better results. Each tree is also estimated through a gradient boosting procedure, which the literature has shown to work better than the simpler Random Forest model (Hastie et al. (2001)). The interested reader should see Breiman (2001) and Ke et al. (2017) for details on these tree estimation methods.

I briefly describe the LightGBM algorithm below.

1. Sample, with replacement, B training samples (y_b, X_b) .
2. For each b in B
 - (a) Randomly sample from the set of predictors X_b .
 - (b) Estimate a classification tree f_b on sample (y_b, X_b) .
3. The final prediction is given by $\hat{f} = 1(\frac{1}{B} \sum_{b=1}^B f_b(y_b, X_b) \geq 0.5)$, or in other words majority vote.

In addition to the evidence in the literature (Bajari et al. (2015)) that random forest models work well in predicting demand, they are typically robust to over-fitting issues (Hastie et al. (2001)). Since the set of predictors is randomly sampled in each tree, and the predictions of the trees are aggregated, there is a low chance that any tree which overfits the data to a set of irrelevant variables significantly influences the aggregated prediction. The same applies to the LightGBM model as well.

The LightGBM model has a very large number of tuning parameters. I use the “LightGBM Classifier” from Python’s “LightGBM” package. Following the recommendations of Ke et al. (2017), I set the majority of tuning parameters to the default values, but increase the number of trees estimated, B , to 500 and decrease the learning rate to 0.01. Both these changes have the effect of increasing the effectiveness of the predictions. For a full discussion of all tuning parameters, see Ke et al. (2017).

For all methods, I randomly sample 50% of observations as the training sample, 25% of the observations as the validation sample, and the remaining 25% of observations are used as the test sample. The validation sample is used to assess the sensitivity of the results to the choice of the dimension of the latent characteristic vectors k and to choose λ for the logistic regression model.

Results

As mentioned in the methodology section, the estimated latent characteristics are evaluated on their ability to capture substitution patterns in a reduced-form demand prediction model. The model is set up so that past-searched products affect the probability of purchase, following the rationale that consumers are more likely to purchase a product that competes with what they have previously searched for. In the model, competing (substitute) products are determined by their distance in characteristic space. Thus, the exercise is to judge which characteristics best represent products when proximity in characteristic space directly affects the model’s fit; observed characteristics, latent characteristics, or both.

Table 6 presents the in and out-of-sample fit across all three models based on the average log-likelihood and average Cross-Entropy Loss. The Cross-Entropy Loss of an individual observation i , CEL_i , is equivalent to the log-likelihood contribution of a single observation. This measure indicates how well the predicted

Table 6: In and Out-of-Sample Fit

Model	Multinomial Logit		LightGBM	
	In-Sample	Out-of-Sample	In-Sample	Out-of-Sample
Characteristics	-0.1670	-0.1679	-0.1847	-0.1895
Latent	-0.1628	-0.1627	-0.1885	-0.1848
Characteristics and Latent	-0.1612	-0.1603	-0.1799	-0.1825

Multinomial Logit results are average log-likelihoods across observations in and out-of-sample. LightGBM results are the average cross-entropy loss across observations in and out-of-sample.

probabilities approximate the underlying purchase decisions. Let d_{ij} be an indicator variable that is one when observation i corresponds to a purchase of j . Then the Cross-Entropy Loss is given by Eq. 15:

$$CEL_i = \sum_j d_{ij} \log(\hat{p}_{ij}) \quad (15)$$

The first comparison of interest is between the characteristics and latent models. The latent model performs significantly better than the Characteristics model, and using the latent characteristics along with observed characteristics leads to the best fitting model. I view these results as validating the approach. The latent characteristics encode demand relevant information, since measuring the impact of past search behavior on purchase choices through distance in latent characteristic space meaningfully improves predictions. Further, that predictions are further improved when used in conjunction with observed characteristics suggests that the latent characteristics capture substitution patterns that cannot be captured by observed characteristics alone.

Recall the pattern of the summary statistics in Table 5 that products in the most recent events have relatively high JC measures with the outcome product. In combination with these results, they suggest that in predicting the purchase of product j , observing the user having browsed products which are close substitutes as per the JC measure is informative of consumers' purchase decisions. Thus, two use cases seem apparent for this approach. Since the latent model performs comparably, if demand relevant characteristics data were not available to researchers, then these results suggest that learning latent characteristics that rationalize substitution patterns revealed through search behavior can be used to build an effective predictive model. In addition, the results suggest that JC captures substitution patterns that observed characteristics may not capture. Thus, even in a case with some demand-relevant characteristics, the approach may pick up substitution patterns that cannot easily be captured by an observed characteristic. For instance, in the market for books or movies, not all demand-relevant characteristics are salient and easy to measure, in which case this approach can be useful.

Monte Carlo Simulation

As a second check on the effectiveness of the JC measure in the predictive models, I conduct a similar analysis as before using search data simulated via a [Weitzman \(1979\)](#) style sequential search model. In the simulations users search across available products with uncertain utilities until they find a sufficient product to purchase, which includes the outside option with a known utility normalized to zero.

In each repetition, I simulate one search sequence for each of 100,000 consumers indexed by i . There are 64 products in the product space, each assigned one of four main-categories (mc), one of four sub-categories (sc), and one of four brands (b). Consumer i 's utility for product j is given by

$$u_{ij} = \alpha_i^{mc} + \beta_i^{sc} + \gamma_i^b - p^j + \varepsilon_{ij} + \mu_j.$$

Here α_i^{mc} is the utility to consumer i in period t of purchasing a product within main-category mc . Similarly, β_i^{sc} is the utility of sub-category sc , and γ_i^b the utility of brand b . The consumer knows the value of these three components, so that they know how much they generally value all products in a category from a specific brand. The term ε_{ij} is an idiosyncratic shock that is known to consumers prior to search. The utility from the unobservable characteristic μ_j is unknown. Thus, it can be thought of as the unobserved quality of the product. This structure generates substitution patterns between products that are explained by both observed and unobserved characteristics. Both shocks ε, μ are i.i.d. drawn from $N(0,1)$. Consumers can search each period, at a cost c , for a specific product j to learn the value of μ_j . The consumer can purchase any previously searched options and can always purchase the outside option without paying any cost. Denote the set of products the consumer has not yet searched at time t by \bar{S}_{it} and the highest realized utility by u_{it}^* . The consumer's problem is a dynamic programming problem described by the following Bellman equation:

$$V(\bar{S}_{it}, u_{it}^*) = \max\{u_{it}^*, \max_{j \in \bar{S}_{it}} \{-c + V(\bar{S}_{it+1}, u_{it+1}^*)\}\}.$$

Every period, the consumer must decide between purchasing the product with the highest realized reward, or searching one of the unsearched products by paying cost c . The process stops when the consumer purchases a product. This is a classic characterization of search based on [Weitzman \(1979\)](#), which outlines the consumer's optimal search and purchase choices. Optimal behavior depends on reservation utilities z_{ij} defined by:

$$\int_{z_{ij}}^{\infty} (u_{ij} - z_{ij}) dF_{ij}(u_{ij}) = c,$$

which in the above utility specification results in the following equation:

$$z_{ij} = \alpha_i^{mc} + \beta_i^{sc} + \gamma_i^b - p^j + \varepsilon_{ij} + m(c),$$

where the value of $m(c)$ is obtained by solving the following equation:

$$c = \phi(m) + m \times \Phi(m) - m.$$

The reservation utilities are those values that equate the cost of searching j with the expected benefit of searching j when the consumer can receive z_{ij} with certainty. That is, they describe the opportunity cost of searching j . Optimal behavior is then described by the following three rules:

- **Selection Rule:** The consumer ranks all unsearched options by their reservation utilities z_{ij} , and searches in decreasing order of reservation utilities.
- **Stopping Rule:** The consumer stops search when the highest realized utility is greater than the highest reservation utility of unsearched options.
- **Choice Rule:** When the consumer stops search, they choose the product with the highest realized utility.

After simulating 100,000 search sequences once, I compute the corresponding latent characteristics for each product. I then simulate another 100,000 search sequences which are used for estimating each of the three prediction models as in the main analysis.

This simulation has a few advantages over the main empirical analysis. First, it allows me to verify that the JC measure is capturing substitutability between products that comes from information unobservable to the researcher. The characteristics in the utility DGP are all observed apart from the unobserved shocks. Thus, the characteristics model includes most demand-relevant characteristics so that the latent model will perform well relative to the characteristics model only if it effectively captures the degree of substitution between products with similar values for unobserved quality. This would also imply that the characteristics plus latent model should lead to an improvement relative to the characteristics model. In the main analysis, one might also be concerned that the observed characteristics are not so demand relevant, so that the improvement in predictions found in the main analysis is larger than one might expect in other settings. The simulation results can speak to whether the latent characteristics are really comparable to demand-relevant characteristics. In addition, the simulation allows for the calculation of confidence intervals. The results in Table 7 present the mean log-likelihood, as well as confidence intervals using the 2.5th and 97.5th percentile of log-likelihood across all repetitions.

The simulation results exhibit the same patterns as in the main analysis. The latent model performs comparably to the characteristics model across all fit metrics, and the characteristics and latent model leads to an improvement in performance. Confidence intervals for the models using latent characteristics are relatively wide, but nonetheless support the claim that performance is better than the characteristics model. We cannot however, differentiate the performance of the latent model with the characteristics and latent model. Together, this suggests that the latent characteristics are driving the majority of the improvement

Table 7: In and Out-of-Sample Fit

Model	Multinomial Logit	
	In-Sample	Out-of-Sample
Characteristics	-0.4302 [-0.4201,-0.4434]	-0.4261 [-0.4149,-0.4383]
Latent	-0.0477 [-0.0421,-0.1937]	-0.0492 [-0.0483,-0.2091]
Characteristics and Latent	-0.0391 [-0.0219,-0.2562]	-0.0406 [-0.0255,-0.2574]

These results are average log-likelihoods across observations in and out-of-sample. Confidence intervals report the 5th and 95th percentile of average log-likelihoods across all 10,000 repetitions.

in performance.

One interesting difference to the main analysis is that the latent models performance is an order of magnitude better than the characteristics model. This is likely due to the smaller product space in the empirical analysis, so that co-searches are more commonly observed across all products. This allows the latent characteristics to represent substitution patterns very well. This suggests that performance would scale well in other applications with more data or smaller product spaces, where we would find a less sparse JC matrix. To conclude, these results support the findings of the empirical analysis.

Conclusion and Managerial Implications

In this paper I present a method to extract substitution patterns of consumers using product search data. I construct a measure, called JC, of substitution between products by aggregating consumers' search sets, effectively crowdsourcing information on substitution patterns. I then map the JC measure to a set of latent characteristics that represent each product, so that comparing products' latent characteristics reflects the JC measure. To demonstrate the latent characteristics' utility, I use them in a predictive demand model in which users' previous search choices affect purchase decisions. I show that by effectively capturing the degree of substitution between two products relative to observed characteristics, the JC measure alone can be used to effectively predict purchases even when demand relevant product characteristics are not available. I also find that prediction performance is highest when both measures are used together.

To support these findings, I carry out a Monte Carlo simulation procedure in which users sequentially browse through products using a [Weitzman \(1979\)](#) style sequential search model. Then, applying the same methods as in the empirical analysis, I show that the results in the simulation are consistent with those of the empirical analysis. The simulation also suggests that the latent characteristics pick up on substitution

patterns that cannot be explained by observed characteristics, even when they may be demand-relevant. Lastly, the simulation results suggest that the approach would scale well to larger search data.

Together these results suggest that the JC-based latent characteristics effectively capture the degree of substitution between products. The results of the paper highlight a specific use-case in improving the predictive accuracy of a probabilistic demand model. However, together with the descriptive evidence comparing observed and latent characteristics the paper shows that they can be used more generally as an effective measure of substitution when detailed product characteristics data are not available. They can be used in a variety of applications, such as in the construction of recommender systems like the model developed in this paper, and in the design of marketing strategies. By identifying products that are close substitutes, firms can target marketing campaigns to consumers who have previously browsed these products. In addition, the JC measure can be used to construct recommender systems. By identifying products that are close substitutes, firms can recommend, for example through advertisements or the order of the user’s search results, products that are likely to be purchased given the consumer’s history of product browsing. This can increase the likelihood of a purchase and improve the user experience by enabling users to more quickly reach products they are interested in.

References

- AdBadger. Amazon advertising stats (2024 update), 2024. URL <https://www.adbadger.com/blog/amazon-advertising-stats/>.
- Tomomichi Amano, Andrew Rhodes, and Stephan Seiler. Flexible demand estimation with search data. *Available at SSRN*, May 2022. doi: <http://dx.doi.org/10.2139/ssrn.3214812>. URL <https://ssrn.com/abstract=3214812>.
- Luis Armona, Greg Lewis, and Georgios Zervas. Learning product characteristics and consumer preferences from search data. In *Proceedings of the 22nd ACM Conference on Economics and Computation*, EC '21, page 98–99, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450385541. doi: 10.1145/3465456.3467548. URL <https://doi.org/10.1145/3465456.3467548>.
- Enghin Atalay, Erika Frost, Alan T Sorensen, Christopher J Sullivan, and Wanjia Zhu. Scalable demand and markups. Working Paper 31230, National Bureau of Economic Research, May 2023. URL <http://www.nber.org/papers/w31230>.
- Patrick Bajari, Denis Nekipelov, Stephen P. Ryan, and Miaoyu Yang. Machine learning methods for demand estimation. *American Economic Review*, 105(5):481–85, May 2015. doi: 10.1257/aer.p20151021. URL <https://www.aeaweb.org/articles?id=10.1257/aer.p20151021>.
- Oren Barkan and Noam Koenigstein. Item2vec: Neural item embedding for collaborative filtering. In *2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6, 2016. doi: 10.1109/MLSP.2016.7738886.
- Steven Berry, James Levinsohn, and Ariel Pakes. Automobile prices in market equilibrium. *Econometrica*, 63(4):841–890, 1995. ISSN 00129682, 14680262. URL <http://www.jstor.org/stable/2171802>.
- Steven Berry, James Levinsohn, and Ariel Pakes. Differentiated products demand systems from a combination of micro and macro data: The new car market. *Journal of Political Economy*, 112(1):68–105, 2004. doi: 10.1086/379939. URL <https://doi.org/10.1086/379939>.
- Gerlof Bouma. Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL*, 30:31–40, 2009.
- Leo Breiman. Random Forests. *Machine Learning*, 45(1):5–32, October 2001. ISSN 1573-0565. doi: 10.1023/A:1010933404324. URL <https://doi.org/10.1023/A:1010933404324>.
- Bart J. Bronnenberg, Jun B. Kim, and Carl F. Mela. Zooming in on choice: How do consumers search for cameras online? *Marketing Science*, 35(5):693–712, 2016. doi: 10.1287/mksc.2016.0977. URL <https://doi.org/10.1287/mksc.2016.0977>.

- Babur De Los Santos, Ali Hortaçsu, and Matthijs R. Wildenbeest. Search with learning for differentiated products: Evidence from e-commerce. *Journal of Business & Economic Statistics*, 35(4):626–641, 2017. doi: 10.1080/07350015.2015.1123633. URL <https://doi.org/10.1080/07350015.2015.1123633>.
- Robert Donnelly, Ayush Kanodia, and Ilya Morozov. Welfare effects of personalized rankings. *Marketing Science*, 43(1):92–113, 2024. doi: 10.1287/mksc.2023.1441. URL <https://doi.org/10.1287/mksc.2023.1441>.
- Jeffrey P. Dotson, John R. Howell, Jeff D. Brazell, Thomas Otter, Peter J. Lenk, Steve MacEachern, and Greg M. Allenby. A probit model with structured covariance for similarity effects and source of volume calculations. *Journal of Marketing Research*, 55(1):35–47, 2018. doi: 10.1509/jmr.13.0240. URL <https://doi.org/10.1509/jmr.13.0240>.
- Jeffrey P. Dotson, Elea McDonnell Feit, and Mark A. Beltramo. Ratings-informed probit for predicting substitution. *Available at SSRN*, January 2024. doi: <http://dx.doi.org/10.2139/ssrn.2282570>. URL <https://ssrn.com/abstract=2282570>.
- Jean-Pierre Dubé and Sanjog Misra. Personalized pricing and consumer welfare. *Journal of Political Economy*, 131(1):131–189, 2023. doi: 10.1086/720793. URL <https://doi.org/10.1086/720793>.
- Terry Elrod. Choice map: Inferring a product-market map from panel data. *Marketing Science*, 7(1):21–40, 1988. ISSN 07322399, 1526548X. URL <http://www.jstor.org/stable/183912>.
- Terry Elrod and Michael P. Keane. A factor-analytic probit model for representing the market structure in panel data. *Journal of Marketing Research*, 32(1):1–16, 1995. ISSN 00222437. URL <http://www.jstor.org/stable/3152106>.
- Sebastian Gabel and Artem Timoshenko. Product choice with large assortments: A scalable deep-learning model. *Management Science*, 68(3):1808–1827, 2022. doi: 10.1287/mnsc.2021.3969. URL <https://doi.org/10.1287/mnsc.2021.3969>.
- J. C. Gower. A General Coefficient of Similarity and Some of Its Properties. *Biometrics*, 27(4):857–871, 1971. ISSN 0006-341X. doi: 10.2307/2528823. URL <https://www.jstor.org/stable/2528823>.
- Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.
- Charles Hodgson and Gregory Lewis. You can lead a horse to water: Spatial learning and path dependence in consumer search. Working Paper 31697, September 2023. URL <http://www.nber.org/papers/w31697>.
- Vladimír Holý, Ondřej Sokol, and Michal Černý. Clustering retail products based on customer behaviour. *Applied Soft Computing*, 60:752–762, 2017. ISSN 1568-4946. doi: <https://doi.org/10.1016/j.asoc.2017.02.004>. URL <https://www.sciencedirect.com/science/article/pii/S1568494617300728>.

- Han Hong and Matthew Shum. Using price distributions to estimate search costs. *The RAND Journal of Economics*, 37(2):257–275, 2006. doi: <https://doi.org/10.1111/j.1756-2171.2006.tb00015.x>. URL <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1756-2171.2006.tb00015.x>.
- Elisabeth Honka, Ali Hortaçsu, and Matthijs Wildenbeest. Chapter 4 - empirical search and consideration sets. In Jean-Pierre Dubé and Peter E. Rossi, editors, *Handbook of the Economics of Marketing, Volume 1*, volume 1 of *Handbook of the Economics of Marketing*, pages 193–257. North-Holland, 2019. doi: <https://doi.org/10.1016/bs.hem.2019.05.002>. URL <https://www.sciencedirect.com/science/article/pii/S2452261919300097>.
- Bruno J.D. Jacobs, Bas Donkers, and Dennis Fok. Model-based purchase predictions for large assortments. *Marketing Science*, 35(3):389–404, 2016. doi: 10.1287/mksc.2016.0985. URL <https://doi.org/10.1287/mksc.2016.0985>.
- Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning: with Applications in R*. Springer, 2013. URL <https://faculty.marshall.usc.edu/gareth-james/ISL/>.
- Guolin Ke, Qi Meng, Thomas Finely, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems 30 (NIP 2017)*, December 2017. URL <https://www.microsoft.com/en-us/research/publication/lightgbm-a-highly-efficient-gradient-boosting-decision-tree/>.
- T. Tony Ke and J. Miguel Villas-Boas. Optimal learning before choice. *Journal of Economic Theory*, 180:383–437, 2019. ISSN 0022-0531. doi: <https://doi.org/10.1016/j.jet.2019.01.005>. URL <https://www.sciencedirect.com/science/article/pii/S0022053119300092>.
- Michael P. Keane. Modeling heterogeneity and state dependence in consumer choice behavior. *Journal of Business & Economic Statistics*, 15(3):310–327, 1997. ISSN 07350015. URL <http://www.jstor.org/stable/1392335>.
- Michael Kechinov. ecommerce behavior data from multi category store, Dec 2019. URL <https://www.kaggle.com/mkechinov/ecommerce-behavior-data-from-multi-category-store>.
- Jun B. Kim, Paulo Albuquerque, and Bart J. Bronnenberg. Mapping online consumer search. *Journal of Marketing Research*, 48(1):13–27, 2011. ISSN 00222437. URL <http://www.jstor.org/stable/25764561>.
- Madhav Kumar, Dean Eckles, and Sinan Aral. Scalable bundling via dense product embeddings. Technical report, January 2020. URL <http://arxiv.org/abs/2002.00100>. arXiv:2002.00100 [cs, stat] type: article.
- Kelvin J. Lancaster. A new approach to consumer theory. *Journal of Political Economy*, 74(2):132–157, 1966. doi: 10.1086/259131. URL <https://doi.org/10.1086/259131>.

- Dawen Liang, Jaan Altosaar, Laurent Charlin, and David M. Blei. Factorization meets the item embedding: Regularizing matrix factorization with item co-occurrence. In *Proceedings of the 10th ACM Conference on Recommender Systems*, RecSys '16, page 59–66, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450340359. doi: 10.1145/2959100.2959182. URL <https://doi.org/10.1145/2959100.2959182>.
- Lorenzo Magnolfi, Jonathon McClure, and Alan T. Sorensen. Triplet Embeddings for Demand Estimation. *Working Paper*, August 2022. doi: 10.2139/ssrn.4113399. URL <https://papers.ssrn.com/abstract=4113399>.
- Puneet Manchanda, Asim Ansari, and Sunil Gupta. The “shopping basket”: A model for multicategory purchase incidence decisions. *Marketing Science*, 18(2):95–114, 1999. doi: 10.1287/mksc.18.2.95. URL <https://doi.org/10.1287/mksc.18.2.95>.
- Daniel McFadden. Conditional logit analysis of qualitative choice behavior. *Frontiers in Econometrics*, pages 105–142, 1973.
- Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper_files/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf.
- José Luis Moraga-González, Zsolt Sándor, and Matthijs R Wildenbeest. Consumer search and prices in the automobile market. *The Review of Economic Studies*, Volume 90(Issue 3):Pages 1394–1440, May 2023.
- Naoki Nishimura, Noriyoshi Sukegawa, Yuichi Takano, and Jiro Iwanaga. A latent-class model for estimating product-choice probabilities from clickstream data. *Information Sciences*, 429:406–420, 2018. ISSN 0020-0255. doi: <https://doi.org/10.1016/j.ins.2017.11.014>. URL <https://www.sciencedirect.com/science/article/pii/S002002551632134X>.
- Daniel M. Ringel and Bernd Skiera. Visualizing asymmetric competition among more than 1,000 products using big search data. *Marketing Science*, 35(3):511–534, 2016. doi: 10.1287/mksc.2015.0950. URL <https://doi.org/10.1287/mksc.2015.0950>.
- Maja Rudolph, Francisco Ruiz, Stephan Mandt, and David Blei. Exponential family embeddings. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/06138bc5af6023646ede0e1f7c1eac75-Paper.pdf.

- Francisco J. R. Ruiz, Susan Athey, and David M. Blei. SHOPPER: A probabilistic model of consumer choice with substitutes and complements. *The Annals of Applied Statistics*, 14(1):1 – 27, 2020. doi: 10.1214/19-AOAS1265. URL <https://doi.org/10.1214/19-AOAS1265>.
- Gary J Russell and Ann Petersen. Analysis of cross category dependence in market basket selection. *Journal of Retailing*, 76(3):367–392, 2000. ISSN 0022-4359. doi: [https://doi.org/10.1016/S0022-4359\(00\)00030-0](https://doi.org/10.1016/S0022-4359(00)00030-0). URL <https://www.sciencedirect.com/science/article/pii/S0022435900000300>.
- Erich Schubert and Peter J. Rousseeuw. Fast and eager k-medoids clustering: O(k) runtime improvement of the pam, clara, and clarans algorithms. *Information Systems*, 101:101804, 2021. ISSN 0306-4379. doi: <https://doi.org/10.1016/j.is.2021.101804>. URL <https://www.sciencedirect.com/science/article/pii/S0306437921000557>.
- Benjamin Reed Shiller. Approximating purchase propensities and reservation prices from broad consumer tracking. *International Economic Review*, 61(2):847–870, 2020. doi: <https://doi.org/10.1111/iere.12442>. URL <https://onlinelibrary.wiley.com/doi/abs/10.1111/iere.12442>.
- Adam N. Smith, Stephan Seiler, and Ishant Aggarwal. Optimal price targeting. *Marketing Science*, 42(3): 476–499, 2023. doi: 10.1287/mksc.2022.1387. URL <https://doi.org/10.1287/mksc.2022.1387>.
- Michalis K. Titsias. One-vs-each approximation to softmax for scalable estimation of probabilities. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS’16, page 4168–4176, Red Hook, NY, USA, 2016. Curran Associates Inc. ISBN 9781510838819.
- Raluca Ursu, Stephan Seiler, and Elisabeth Honka. The sequential search model: A framework for empirical research. *Available at SSRN*, 2023. URL <http://dx.doi.org/10.2139/ssrn.4236738>.
- Raluca M. Ursu, Qingliang Wang, and Pradeep K. Chintagunta. Search duration. *Marketing Science*, 39(5): 849–871, 2020. doi: 10.1287/mksc.2020.1225. URL <https://doi.org/10.1287/mksc.2020.1225>.
- Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. URL <http://jmlr.org/papers/v9/vandermaten08a.html>.
- Laurens van der Maaten and Kilian Weinberger. Stochastic triplet embedding. In *2012 IEEE International Workshop on Machine Learning for Signal Processing*, pages 1–6, 2012. doi: 10.1109/MLSP.2012.6349720.
- Donatella Vicari and Marco Alfó. Model based clustering of customer choice data. *Computational Statistics & Data Analysis*, 71:3–13, 2014. ISSN 0167-9473. doi: <https://doi.org/10.1016/j.csda.2013.09.014>. URL <https://www.sciencedirect.com/science/article/pii/S0167947313003381>.
- Joel Waldfogel. First degree price discrimination goes to school. *The Journal of Industrial Economics*, 63 (4):569–597, 2015. doi: <https://doi.org/10.1111/joie.12085>. URL <https://onlinelibrary.wiley.com/doi/abs/10.1111/joie.12085>.

Martin L. Weitzman. Optimal search for the best alternative. *Econometrica*, 47(3):641–654, 1979. ISSN 00129682, 14680262. URL <http://www.jstor.org/stable/1910412>.

W. J. Youden. Index for rating diagnostic tests. *Cancer*, 3(1):32–35, 1950.

Jonathan Z. Zhang, Oded Netzer, and Asim Ansari. Dynamic targeted pricing in b2b relationships. *Marketing Science*, 33(3):317–337, 2014. doi: 10.1287/mksc.2013.0842. URL <https://doi.org/10.1287/mksc.2013.0842>.

Appendix

Figure 4: Main Categories of Products

