

# Triplet Embeddings for Demand Estimation

Lorenzo Magnolfi  
University of Wisconsin  
[magnolfi@wisc.edu](mailto:magnolfi@wisc.edu)

Jonathon McClure  
University of Wisconsin  
[jmcclure2@wisc.edu](mailto:jmcclure2@wisc.edu)

Alan Sorensen  
University of Wisconsin & NBER  
[sorensen@ssc.wisc.edu](mailto:sorensen@ssc.wisc.edu)\*

July 2022

## Abstract

We propose a method for augmenting conventional demand estimation methods with crowd-sourced data on products’ locations in product space. In addition to the usual data on prices and quantities, our method incorporates triplets data (of the form “product A is closer to B than it is to C”) obtained from an online survey. The triplets are used to compute an embedding—i.e., a low-dimensional representation of the latent product space—which can then be used either (i) as a substitute for data on observed characteristics in a conventional mixed logit model, or (ii) to compute pairwise product distances that can discipline the cross-elasticity parameters of a simple log-linear demand model. We illustrate the performance of both approaches by estimating demand for ready-to-eat cereals, and find that the information contained in the embedding leads to more plausible substitution patterns and better model fit.

---

\*A previous version of this draft was circulated under the title “Embeddings and Distance-based Demand for Differentiated Products.” We thank Chris Sullivan, Giovanni Compiani, Jeff Thurk, and seminar participants at University of Maryland, IO<sup>2</sup>, and HOC for helpful comments. The results reported below represent our own analyses calculated (or derived) based in part on data from Nielsen Consumer LLC and marketing databases provided through the NielsenIQ Datasets at the Kilts Center for Marketing Data Center at The University of Chicago Booth School of Business. The conclusions drawn from the NielsenIQ data are our own and do not reflect the views of NielsenIQ. NielsenIQ is not responsible for, had no role in, and was not involved in analyzing and preparing the results reported herein.

# 1 Introduction

Estimating demand systems in differentiated product markets is fundamental in empirical Industrial Organization (IO), and the toolkit of methods can be roughly divided into two approaches.<sup>1</sup> The *product space* approach assumes that consumers have preferences over products, and product-level demand comes from the aggregation of those preferences. This is perhaps the most natural way to conceptualize demand, and has the advantage of yielding demand equations that are computationally simple to estimate (e.g., [Christensen, Jorgenson, and Lau, 1975](#); [Deaton and Muellbauer, 1980](#)). The *characteristics space* approach, pioneered by [Lancaster \(1966\)](#) and [McFadden \(1974\)](#), instead treats products as bundles of characteristics, and defines consumers’ preferences over these characteristics. Methods in this vein have their own advantages: they are based on theoretically grounded models of discrete choice, they have convenient analytical properties (e.g., closed-form solutions for firms’ predicted market shares), and with the inclusion of random coefficients on some characteristics (as suggested, for example, by [Berry, Levinsohn, and Pakes \(1995\)](#) (BLP) and [McFadden and Train \(2000\)](#)) they allow for rich patterns of substitution between products.

Of course, each of these approaches also has meaningful drawbacks. The principal challenge of estimating product space models is one of dimensionality: absent any restrictions, a market with  $J$  products will require estimation of separate parameters for each of the  $J^2$  demand elasticities. And while a key rationale for the characteristics space approach is that it collapses preferences over  $J$  products down to a set of  $K \ll J$  characteristics, in practice it can be challenging to define and/or collect data on the demand-relevant characteristics.

In this paper we propose a pragmatic approach for obtaining complementary data that can be used either to discipline the parameters in a product space model or to serve as (latent) characteristics in a characteristics space model. We first solicit product comparisons via an online survey to generate data of the form “product A is closer to B than it is to C”—commonly referred to as ‘triplets data’ in the machine learning literature—and then apply the t-Distributed Stochastic Triplet Embedding (tSTE) algorithm proposed by [Van Der Maaten and Weinberger \(2012\)](#) to compute an *embedding* of the products in low-dimensional space. Distances between products are then easily calculated from this embedding, and cross-price elasticities in a product space model can be estimated as a function of these distances as in [Pinkse, Slade, and Brett \(2002\)](#). Alternatively, the products’ coordinates in the embedding can be treated as the characteristics in a conventional mixed logit demand model like BLP. As we explain below, these embeddings are easy to generate, and the data required to compute them are straightforward to obtain.

---

<sup>1</sup>See [Berry and Haile \(2021\)](#) and [Gandhi and Nevo \(2021\)](#) for recent surveys.

We illustrate the method by estimating demand for ready-to-eat breakfast cereals—a common laboratory for evaluating demand estimation methodologies—augmenting the usual price-and-quantity data with survey data obtained from college students and Mechanical Turk workers. The triplets data from the survey lead to an embedding of products that appears quite sensible. We use the embedding to first show how it can be used in a product space model similar to that of [Pinkse et al. \(2002\)](#), where the (log) quantity demanded for any product is a linear function of its own (log) price and of all competing products’ (log) prices, allowing the cross-price elasticity parameters to be functions of pairwise product distances computed from the embedding. Estimates of this model are computationally trivial to obtain, and they yield reasonable own- and cross-price elasticities—similar to those reported in prior studies like [Nevo \(2001\)](#) and [Backus, Conlon, and Sinkinson \(2021\)](#). Importantly, we show that the distances computed from the embedding deliver meaningfully better estimates than distances computed from observed product characteristics.

We then show how the coordinates from the embedding can be used in more conventional discrete choice models like BLP. If we treat the products’ coordinates in the embedding as latent characteristics, essentially including them as the covariates in an otherwise standard BLP model, we obtain elasticity estimates that are very similar to those from a model that uses observable characteristics. This result is particularly encouraging because it suggests our method can deliver credible estimates even in markets where demand-relevant characteristics are more elusive, such as fashion apparel, movies, or music. Using survey data to obtain an embedding is essentially a way of crowd-sourcing data on product characteristics, an option that will be especially useful in cases where data on characteristics are otherwise difficult to collect.

While the specifics of our method are novel, we are not the first to propose the use of embeddings in demand estimation. [Bajari, Cen, Chernozhukov, Manukonda, Wang, Huerta, Li, Leng, Monokroussos, Vijaykumar et al. \(2021\)](#) use deep neural nets to generate numeric latent attributes (i.e., an embedding) from products’ images and text descriptions, and then use those attributes to estimate a hedonic price function for apparel items on Amazon.com. This is a nice example where the demand-relevant information about a product—say, a woman’s dress—cannot be easily summarized by a set of characteristics, even though humans can easily process and synthesize the relevant information from the product’s image and/or text description. With a similar motivation, [Han, Schulman, Grauman, and Ramakrishnan \(2021\)](#) use a deep neural net to compute an embedding describing the product space for fonts. [Armona, Lewis, and Zervas \(2021\)](#) show how to use Bayesian Personalized Ranking to learn products’ latent attributes from search data (consumers’ web-browsing histories), and use their method to estimate demand for hotels. Related articles that use data on consumers’ transactions and search to fit embeddings and estimate demand also include [Ruiz, Athey, and Blei \(2020\)](#), [Kumar, Eckles, and Aral \(2020\)](#) and [Gabel and Timoshenko \(2022\)](#).

An important distinction is that in these articles search or transactions data are informative about products’ latent attributes *and* consumers’ preferences for those attributes.<sup>2</sup> In our case, we use the triplets data from the survey to learn about products’ latent attributes, but our estimates of consumers’ preferences are driven by price and quantity data.<sup>3</sup>

We also view our approach as being similar in spirit to studies that employ auxiliary data to augment existing demand estimation methodologies. [Berry, Levinsohn, and Pakes \(2004\)](#) is a canonical study in which second-choice data from surveys are used to generate additional moments in the estimation of demand for automobiles. [Petrin \(2002\)](#) is an early example of combining demographic data with the usual price-and-quantity data to get richer estimates of substitution patterns. More recently, [Conlon, Mortimer, and Sarkis \(2021\)](#) show how demand estimates can be meaningfully improved by incorporating data on “second-choice diversion ratios,” in their case obtained from experimentally generated stockouts. They even show that the information contained in such data is powerful enough to enable estimation of a semi-parametric model that imposes much lighter assumptions than conventional mixed logits.

## 2 Demand Estimation and Linear Embeddings

Consider a market, indexed by  $t$ , where firms offer a set  $\mathcal{J}_t$  of differentiated products. Prices and quantities for each good  $j$  are denoted as  $p_{jt}$  and  $q_{jt}$ . The demand system that maps prices into quantities depends on two key sets of primitives: consumers’ preferences and demographics; and the product space. We assume that products can be represented by coordinates in the  $m$ -dimensional Euclidean space; thus, the product space in market  $t$  is a set of vectors  $\mathbf{x}_t \equiv \{x_{1t}, \dots, x_{J_t}\} \in \mathbb{R}^{m \times \mathcal{J}_t}$ . Hence, demand can be written as  $q_{jt} = \sigma_j(p_t; \mathbf{x}_t)$  for some function  $\sigma_j$ .

The product space  $\mathbf{x}_t$  is a key element of the empirical demand system under either estimation approach mentioned above. In the *characteristics space* approach, demand is assumed to arise from discrete choices of individual consumers, whose preferences are defined directly over the product space coordinates. Thus,  $x_{jt}$  enters consumers’ indirect utility for product  $j$ , interacted with preference parameters. In the *product space* approach, the functions  $\sigma_j$  are estimated directly, with functional form restrictions imposed (typically based on either convenience or a representative consumer micro-foundation). The importance of the product space  $\mathbf{x}_t$  is that it can play a role

---

<sup>2</sup>This is also the spirit of conjoint analysis in marketing.

<sup>3</sup>An earlier empirical study that aims to recover both attributes and preferences from the same data is [Goettler and Shachar \(2001\)](#). They use panel data on consumers’ television viewing choices in combination with a bliss-point model of demand to simultaneously estimate television shows’ latent attributes along with consumers’ preferences for those attributes. Though similar in spirit to our exercise, again the distinction is that we propose to compute an embedding from auxiliary data before using it as an input to demand estimation.

in disciplining the otherwise overabundant cross-elasticity parameters: as in [Pinkse et al. \(2002\)](#), cross-elasticities of demand between products  $j$  and  $k$  can be modeled as a function of the distance  $d_{jk}(\mathbf{x}_t)$  between the two products.

Within this framework, our method can be understood as a way of recovering  $\mathbf{x}_t$  from auxiliary data as an *embedding* when product characteristics are not observable or are otherwise difficult to codify. The next subsection provides an overview of embeddings, and the following sections summarize how they can be incorporated into either of the two main approaches to demand estimation.

## 2.1 Product embeddings

In machine learning, an *embedding* is a low-dimensional, learned continuous vector representation of discrete variables.<sup>4</sup> In our case the discrete variables are just product indicators (“this is product  $j$ ”), and the objective is to assign locations (real-valued vectors) to these products in a way that best satisfies the distance comparisons from a training dataset. As training data we use *triplets*—i.e., comparisons of the form “product A is closer to B than it is to C”—obtained from a survey that we describe in detail below in the context of our application.

Thus, given our set of products, we want to find a set of vectors  $\mathbf{x} \equiv \{x_1, \dots, x_J\} \in \mathbb{R}^{m \times J}$  that represent the products in  $m$ -dimensional space, and assume that this corresponds to the product space that enters the demand system. To learn the embedding from triplet data, we use the t-distributed Stochastic Triplet Embedding (tSTE) algorithm proposed by [Van Der Maaten and Weinberger \(2012\)](#). Letting  $\mathcal{T}$  be the set of triplet comparisons in our data, each one indicating that some product  $i$  is closer to  $j$  than it is to  $k$ , tSTE solves

$$\max_{\mathbf{x}} \sum_{(i,j,k) \in \mathcal{T}} \ln(\pi_{ijk}) \quad \text{where} \quad \pi_{ijk} = \frac{\left(1 + \frac{\|x_i - x_j\|^2}{\alpha}\right)^{-\frac{\alpha+1}{2}}}{\left(1 + \frac{\|x_i - x_j\|^2}{\alpha}\right)^{-\frac{\alpha+1}{2}} + \left(1 + \frac{\|x_i - x_k\|^2}{\alpha}\right)^{-\frac{\alpha+1}{2}}}$$

and  $\alpha$  is the degrees of freedom parameter for the underlying Student- $t$  kernel. To gain intuition about this program, note that to fit one single triplet the embedding assigns equal coordinates to products  $i$  and  $j$ , and infinitely far coordinates to products  $i$  and  $k$ , so that  $\ln(\pi_{ijk})$  diverges. As

---

<sup>4</sup>Common uses of embeddings in machine learning include image classification and natural language processing. For example, Google’s Word2Vec algorithm uses a neural network to assign vector representations to words so that the cosine similarity between any two words’ vectors can be used as a measure of their semantic similarity. Embeddings are also commonly used for visualizing high-dimensional data: collapsing to two or three dimensions allows for simple plots in which clusters and other patterns are easy to see.

we introduce further triplet comparisons, the solution is more intricate: the embedding has to fit more complex patterns because the same products are involved in multiple comparisons.

In our empirical application to ready-to-eat breakfast cereal we have  $J = 86$  products, so if we choose to fit a 6-dimensional embedding ( $m = 6$ ) then the above program is a numerical optimization problem with 516 free variables. It may thus seem remarkable that the problem is computationally tractable, but ordinary gradient descent algorithms converge in a matter of minutes.<sup>5</sup>

## 2.2 Using embeddings in product space demand models

Although product-space demand models have a long tradition in applied economics, they are often deemed unsuitable for IO applications. This is for one main reason: in even the simplest product-space demand systems (e.g., linear or log-linear) the number of parameters grows exponentially with the number of products, making them impractical in markets for differentiated goods.<sup>6</sup>

Various solutions to this problem have been devised;<sup>7</sup> in this paper we adopt the method proposed by Pinkse et al. (2002), who note that when competition among firms is spatial (i.e., it depends on some topology of the product space) the parameters that govern substitution between products can be projected on a flexible function of their distances. When products have an observable location in the physical space, as in the application of Pinkse et al. (2002) or in Houde (2012), distances are straightforward to measure. When spatial competition is only figurative, as in the case of Pinkse and Slade (2004)’s study of the UK beer market, distance can instead be modeled as a function of observable product characteristics.

Using an embedding computed from triplets data as described above, we can obtain a map of the product space even when the products’ characteristics are difficult to observe or quantify, and the distances between products in the embedding can be used in the framework of Pinkse et al. (2002). In the empirical exercise below, we estimate with product-level data the log-linear demand model

$$\ln(q_{jt}) = \alpha_j + \beta_j \ln(p_{jt}) + \sum_{k \neq j} f(d_{jk}; \gamma) \ln(p_{kt}) + \epsilon_{jt} \quad (1)$$

---

<sup>5</sup>We used a version of the MATLAB code provided by Laurens Van der Maaten: <https://lvdmaaten.github.io/stochastic-triplet-embedding.html>.

<sup>6</sup>Other reasons include the difficulties in incorporating (and estimating) heterogeneity across consumers, and in evaluating the demand for new products. See Gandhi and Nevo (2021) for more discussion.

<sup>7</sup>For example, the researcher can restrict substitution across categories of goods by modeling choice as a multi-stage budgeting problem (Gorman, 1959). Hausman, Leonard, and Zona (1994) use a model in this spirit based on Deaton and Muellbauer (1980)’s Almost Ideal Demand System.

where  $\alpha_j, \beta_j$  and  $\gamma$  are parameters, and  $\epsilon_{jt}$  is a consumer-product specific unobservable. The function  $f$  is a real-valued transformation of the pairwise distances among products we compute from the embedding; we discuss specific parameterizations of this function below.

The log-linear formulation we adopt is convenient because the coefficients on log prices can be interpreted directly as elasticities:  $\beta_j$  is the own-price elasticity for product  $j$ , and the cross-elasticity between products  $j$  and  $k$  is a function of their distance  $d_{jk}$ . As a consequence, this approach offers an obvious computational advantage: elasticities can be obtained from simple linear or nonlinear regressions once a functional form for  $f$  has been chosen, and suitable identifying assumptions have been made.<sup>8</sup> This is in contrast with state-of-the art implementations of discrete-choice demand models, which instead require computationally intensive nonlinear optimization routines.

While the log-linear specification is convenient for showcasing our method, other specifications of the model that incorporate distances are possible, and may be preferable depending on the application at hand.<sup>9</sup> First, from an econometric perspective, while the log-linear specification models demand as a regression – with one structural error per equation – this is a strong restriction that is violated in more flexible classes of models (Berry and Haile, 2021). Embeddings data could however be used to discipline flexible models of *inverse* demand. Second, as the log-linear model lacks economic structure, it may be preferable to use a specification corresponding to a micro-founded demand system—to enable welfare analysis for the representative consumer, and/or to enforce certain theoretical properties that might be important.<sup>10</sup> With this in mind, we discuss in Appendix A.2 an alternative specification based on the AIDS framework of Deaton and Muellbauer (1980).

## 2.3 Using embeddings in characteristics space demand models

The natural way to use an embedding in a conventional logit-style demand model (like BLP) is to treat the products’ coordinates in the embedding as characteristics (i.e.,  $x$  variables in the consumer’s indirect utility function). If an  $m$ -dimensional embedding is computed, then each of the  $m$  dimensions can be treated as a characteristic. Because each dimension of the embedding

<sup>8</sup>Identification and estimation of the model are discussed in Section 3.4 below.

<sup>9</sup>For instance, Anderson and Vycassim (2001) note that log-linear models have undesirable implications for retailer category pricing.

<sup>10</sup>For instance, Jaffe and Weyl (2010) establish that linear demand cannot be generated by discrete choice, and Jaffe and Kominers (2012) establish that in fact any demand system that is additively separable in own price cannot be generated by discrete choice. Of course, many markets that are modeled in a discrete-choice framework are not truly based on single discrete choices by consumers. Even in the market for breakfast cereal, which we study in our application, the true consumer decision problem is not simply which *one* of many cereals to buy, but which *set* of cereals to buy, and how much of each.

enters the model separately, this approach is more flexible than the one described above for the product space model; it allows the data to determine which dimensions of the embedding are most relevant to substitution.<sup>11</sup>

An obvious disadvantage of this approach is that these are latent characteristics without any automatic interpretation. However, we expect that in many cases the latent characteristics from a crowd-sourced embedding will give a better overall description of the products and their relationships to one another than could be obtained from observable characteristics. For example, the 2020 Toyota Camry and the 2020 MINI Clubman are very similar cars based on horsepower, fuel efficiency, passenger volume, and curb weight;<sup>12</sup> but we suspect consumers would not identify the two cars as being near each other in product space. In our cereal application, our survey appropriately indicates that Cocoa Pebbles are closer to Cocoa Krispies than Tootie Frooties, even though Tootie Frooties are closer based on sugar, fiber, and calories from fat.

### 3 Empirical application

We illustrate our method by estimating demand for breakfast cereals. This product category has been the subject of important studies on demand estimation (e.g., [Nevo, 2001](#)), and is well understood by IO economists. As fairly rich data on cereals’ nutritional and other characteristics are available, we will measure the usefulness of embeddings data in this category by comparing the performance of models using embeddings data with that of models using standard characteristics data.

We first describe the survey we used to collect the triplets data, and then summarize the embedding that we compute from those data. We then provide details of the demand estimation, both for the product space model and the characteristics space model. In each case we emphasize the comparison to demand estimates from the same model *without* the use of an embedding—i.e., either using pairwise product distances computed from observable characteristics in the product space model, or using observable characteristics as the “ $x$  variables” in the discrete choice model.

---

<sup>11</sup>In Section 4.2 below we discuss how to add similar flexibility to the log-linear product space model.

<sup>12</sup>The HP, MPG, volume, and weight specifications for the Camry (Clubman) are 203 (189), 34 (29), 100 (93), and 3241 (3235).

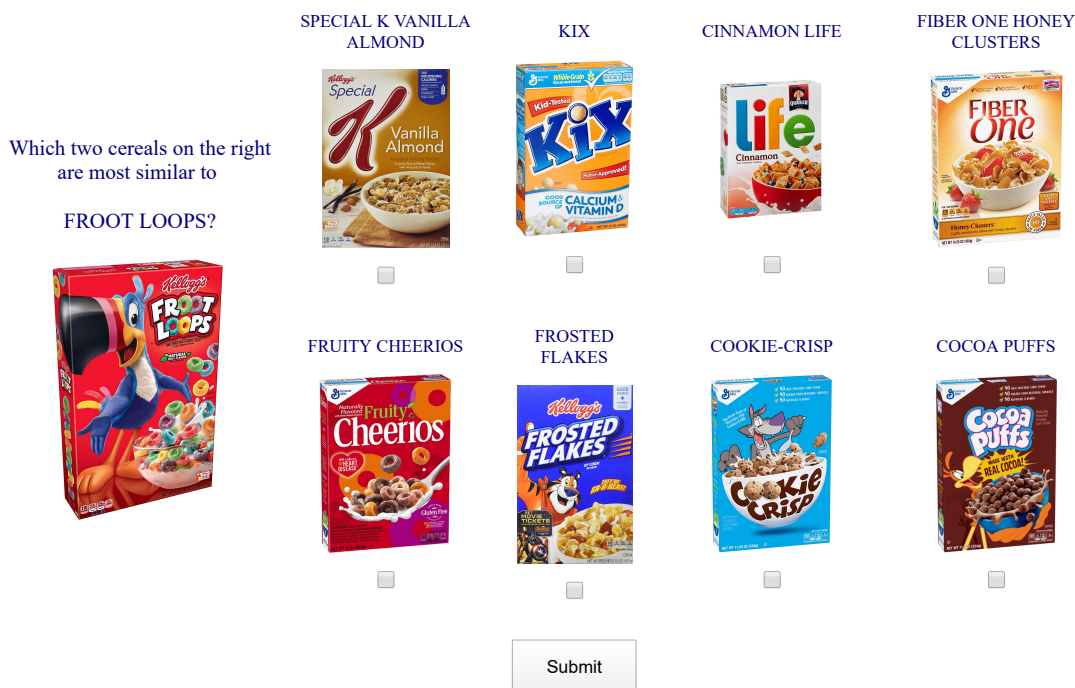
### 3.1 Survey

To obtain the triplets data needed to learn the embedding of cereal products, we conducted an online survey that asked respondents to make a series of product comparisons. Each page showed a reference product along with eight comparison products, and the respondent was asked to indicate which two were most similar to the reference product.<sup>13</sup> Figure 1 shows a sample page from the survey.

FIGURE 1: Sample survey page

Which two cereals on the right are most similar to

FROOT LOOPS?



SPECIAL K VANILLA ALMOND

KIX

CINNAMON LIFE

FIBER ONE HONEY CLUSTERS

FRUITY CHEERIOS

FROSTED FLAKES

COOKIE-CRISP

COCOA PUFFS

Submit

The figure shows a sample page from our online survey.

Each comparison page thus yields 12 triplets: each of the 2 checked products is considered closer to the reference product than the 6 unchecked products. Survey respondents were asked to complete up to 20 comparison pages, so each respondent generated as many as 240 triplet comparisons.

The survey respondents included 456 undergraduate students at the University of Wisconsin and 220 workers from Amazon’s Mechanical Turk platform. Respondents were first asked to indicate how often they eat cereal and how many different cereals they have tried (see Figure 7 in Appendix B), and were then shown the sequence of comparison pages. We found only negligible differences between the embeddings based on Turk workers’ responses vs. undergraduate students’ responses,

<sup>13</sup>This approach to obtaining triplet comparisons is discussed in Wilber, Kwak, and Belongie (2014).

so we pooled their responses when computing the embedding used in the analyses below. We discarded data from a very small percentage of respondents who indicated no prior experience with breakfast cereal, but this has little impact on the computed embedding. The final sample includes 107,820 triplet comparisons.

### 3.2 Computed embedding

For the demand estimation below we use 6-dimensional embeddings, but for purposes of visualization Figure 2 shows a 2-dimensional embedding computed from the same triplets data. Even with only two dimensions, the algorithm neatly organizes the products into reasonable clusters—for example, sugary fruity cereals (clustered in the northeast region of the figure) and sugary chocolatey cereals (clustered in the southeast).

FIGURE 2: Plot of two-dimensional embedding



The figure shows a two-dimensional embedding for ready-to-eat cereals estimated from the triplets data.

Based on distances from the 6-dimensional embedding, Table 1 lists the two nearest cereals to some of the highest-revenue brands in our sample. In general the embedding appears to be correctly identifying the most similar products. This should not be surprising, since identifying similar

cereals is not difficult for a human, and our procedure is essentially synthesizing thousands of comparisons made by humans.

TABLE 1: Examples of nearby brands based on 6-dimensional embedding

Brand	Nearest brand	Second-nearest brand
GM Honey Nut Cheerios	GM Honey Nut Cheerios Medley Crunch	Post Honey Graham Oh’s
Kellogg’s Frosted Flakes	Malt-o-Meal Frosted Flakes	Kellogg’s Corn Flakes
GM Cinammon Toast Crunch	GM French Toast Crunch	Malt-o-Meal Cinammon Toasters
Kellogg’s Froot Loops	Malt-o-Meal Tootie Fruities	Kellogg’s Apple Jacks
Kellogg’s Raisin Bran	Kellogg’s Raisin Bran Crunch	Post Raisin Bran
Kellogg’s Rice Krispies	GM Kix	Kellogg’s Corn Pops
GM Cocoa Puffs	Kellogg’s Cocoa Krispies	Post Cocoa Pebbles

The table reports, for the sample of ready-to-eat cereal brands in the first column, the nearest and second-nearest brand in the 6-dimensional embedding.

Nevertheless, a natural question is whether the product distances that come from the survey are the right distances for the purposes of demand estimation. It is important to note that the distances themselves are not intended to be measures of substitution. Like ordinary product characteristics in conventional discrete-choice methods, they are *inputs* into the demand estimation, which uses price and quantity data to measure substitution patterns. Ideally we want the demand estimation to use these inputs as flexibly as needed to deliver the true substitution patterns—much as allowing for random coefficients on product characteristics allows for flexible substitution in the discrete-choice framework—so it may not be enough to simply use Euclidean distances. For purposes of illustrating the method, we proceed with Euclidean distances when estimating the product space demand model below, but in Section 4.2 below we discuss how to incorporate the distances more flexibly.

Our use of surveys to elicit information about products’ positions is reminiscent of conjoint analysis, but our survey is intended to recover product attributes, not consumers’ preferences for those attributes. In conjoint studies, survey respondents are asked to rate the desirability of each product in a set of hypothetical offerings, and response data are then used to estimate preferences for the products’ observed attributes. A common use of this technique is to optimize the design of new products. Thus, conjoint analysis aims to learn consumers’ preferences for known product attributes. By contrast, the purpose of our survey is merely to learn products’ latent attributes. The embedding computed from the survey is then combined with revealed preference data (prices and quantities from actual markets) to estimate preferences.

### 3.3 Price and quantity data

Our data on prices and quantities come from Nielsen’s Retail Scanner data from the year 2017. The unit of observation in our analysis is a UPC-store-week. Our sample of UPCs consists of the highest-selling UPCs for the 86 brands that together account for 80% of total sales in the breakfast cereal category. We focus on large markets with many competing products, limiting the sample by (i) keeping product-market combinations which appear in all 52 weeks of the data, (ii) keeping markets with at least 50 UPCs, and (iii) keeping UPCs which appear in at least 50 markets. This results in a sample of 684,476 UPC-retailer-DMA-week observations, containing 43 retailer chains, 111 DMAs, and 189 unique retailer-DMA pairs across 52 weeks. Table 2 shows some basic summary statistics for the 86 products in the sample, as well as for the 189 retailer-DMA pairs.

TABLE 2: Summary statistics

<b>Cereal products (<math>N=86</math>)</b>					
	Mean	Std. Dev.	Percentiles		
			.10	.50	.90
Average price	3.58	0.83	2.50	3.51	4.77
Average weekly sales	216.16	649.63	9	55	480
Number of stores	153.06	33.27	105	165	189
<b>Retailer-DMA pairs (<math>N=189</math>)</b>					
	Mean	Std. Dev.	Percentiles		
			.10	.50	.90
# of cereal products carried	69.65	8.20	57	72	79
Avg. weekly cereal revenues (000)	45.30	77.48	3.71	17.15	117.46

The table reports summary statistics for the 86 cereal UPCs and 189 retailer-DMA pairs we use for demand estimation.

### 3.4 Log-linear demand estimates

To estimate demand in product space, we use the Nielsen price and quantity data to estimate the linear model shown above in equation 1, using the embedding to compute the pairwise product distances  $d_{jk}$ . Theory suggests that the function  $f(d_{jk}; \gamma)$  should be monotonically decreasing in  $d_{jk}$ , since more distant products should have lower substitution. While functional forms such as  $f(d_{jk}; \gamma) = \frac{\gamma}{1+d_{jk}}$  easily incorporate this, there are important reasons to estimate  $f(d_{jk}; \gamma)$  flexibly. First, it allows estimated substitution patterns to be driven more by the sales data than by the embedding. Second, if we are unsure that the embedding is returning reasonable product distances, a flexible distance function provides a method of validation. If the estimated distance function is non-monotonic or flat, it suggests the embedding is doing a poor job of capturing product attributes that are relevant to substitution.

We experimented with various flexible approaches, including sieves and b-splines, but found that a simple cubic polynomial in scaled distances worked well:

$$f(d_{jk}; \gamma) = \gamma_0 + \gamma_1 d_{jk} + \gamma_2 d_{jk}^2 + \gamma_3 d_{jk}^3 \quad \text{with} \quad d_{jk} = \frac{\|x_j - x_k\|}{\max_{\{h,i\}} \{\|x_h - x_i\|\}} \quad (2)$$

This is the functional form we used for the baseline results shown below; we return to discuss more flexible distance metrics in Section 4.2.

As noted in [Berry and Haile \(2021\)](#), identification of demand for differentiated products is complicated by two fundamental challenges: price endogeneity, and co-dependence of the demand of each product on the latent demand shocks for all other products in the market. For the purpose of showcasing the embeddings data in the simplest possible context, we impose strong assumptions to set aside these challenges. In particular, we make the strong assumption that prices and product distances are uncorrelated with the econometric unobservable  $\epsilon_{jt}$ . Coupled with the restrictions embedded in the specification of Equation (1), the assumption of exogenous prices allows us to estimate the model with OLS, which is robust and computationally easy in large datasets.

The assumption of exogenous prices deserves further discussion, as price endogeneity in demand estimation is typically a first-order concern, especially with cross-sectional data. To understand the assumption that prices are exogenous, it is useful to discuss the sources of variation in  $\epsilon_{jt}$  in our data. As we observe for each product weekly sales data at the DMA-retailer level, we can control via fixed effects for standard sources of variation in  $\epsilon_{jt}$  that are correlated with prices. For instance, unobserved product characteristics can be absorbed by product or brand fixed effects. Similarly, unobserved promotional activity (such as feature promotions, typically planned quarterly) can be captured by monthly or quarterly fixed effects. While exogenous prices and fixed effects are in general not sufficient for identification of demand ([Berry and Haile, 2021](#)), we assume that any residual variation in  $\epsilon_{jt}$  across time and DMA-retailers is due to variables such as wholesale prices, retailer strategy, or other retailer-specific costs that affect demand only through prices.

Alternatively, we can use instruments to identify the log-linear model. In our specification, we need instruments not only for own price, but also for the prices of all other products. In principle, Hausman or BLP instruments can be used in this context. But in a setting like ours with high-frequency data, these instruments may not generate estimators with good sampling properties, as argued by [Rossi \(2014\)](#). We proceed with the assumption that prices are exogenous and estimate the model with OLS, and leave the discussion of an IV specification to Appendix A.1.<sup>14</sup> We also

---

<sup>14</sup>[Hitsch, Hortacsu, and Lin \(2019\)](#) use a high-frequency dataset similar to ours and also argue it is best to estimate with OLS instead of instrumenting for price.

note that, in cases where the researcher wants to estimate a log-linear (or similar) specification using instruments, having distances from embeddings limits the number of parameters to be estimated and thus makes it easier for the exogenous variation provided by the instruments to pin down the parameters of the model.

In evaluating our results, the main comparison we want to make is to an alternative specification that relies only on observable characteristics to compute pairwise product distances. That is, we estimate the same log-linear model (1) but compute the product distances  $d_{jk}$  based on sugar (grams per serving), fiber (grams per serving), and calories from fat (per serving).<sup>15</sup> We use the same cubic polynomial distance function as in (2), but also add a term to reflect differences in the target demographic. Letting  $G_j \in \{\text{Kids, Adult, All Family}\}$  denote the category of cereal  $j$ , the modified distance function is

$$f(d_{jk}; \gamma) = \gamma_0 + \gamma_1 d_{jk} + \gamma_2 d_{jk}^2 + \gamma_3 d_{jk}^3 + \gamma_4 \mathbb{1}(G_j \neq G_k)$$

We expect  $\gamma_4$  to be negative, as two products in different categories should be less substitutable than two in the same category.

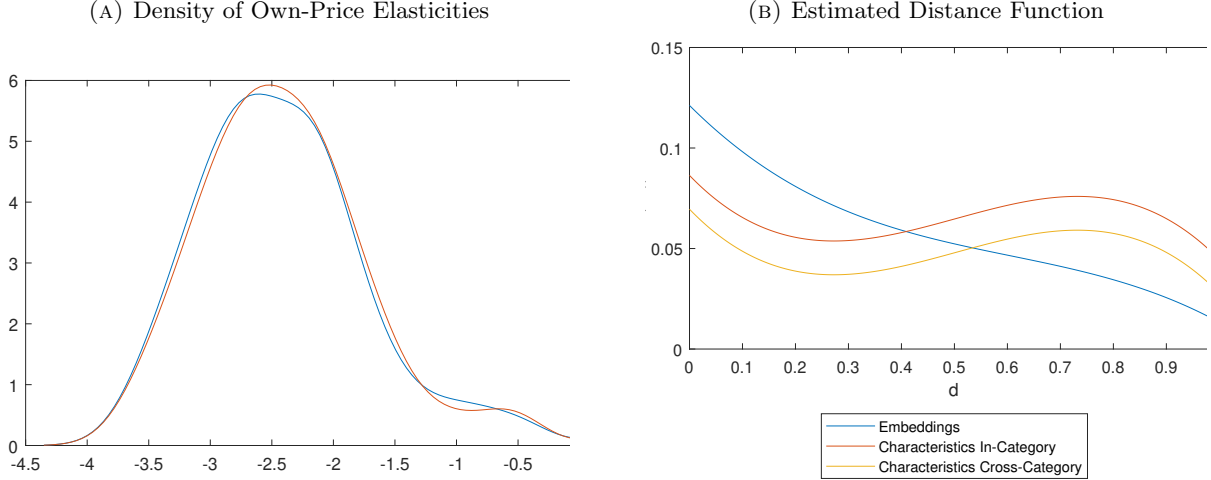
Figure 3 summarizes the distributions of estimated own- and cross-price elasticities from the two specifications. The left panel shows kernel density estimates of the own-price elasticities. These estimates fall in a reasonable range (all negative, mostly between -1 and -4) and are very similar between the two specifications. The similarity is not surprising, since own-price elasticity estimates are driven almost entirely by the price and quantity data. Where the two models differ is in the estimated cross-price elasticities, which depend on the estimated distance functions, shown in the right panel of Figure 3. When product distances are computed from the embedding, this function has the expected monotonically decreasing shape: nearby products are estimated to have larger cross-price elasticities. When distances are computed from observable characteristics, the estimated distance function has an implausible shape (non-monotonic) and is overall relatively flat, implying that cross-price elasticities for the “nearest” products are hardly different from those for the most distant products.

To illustrate the differences between the two specifications, Table 3 shows estimated cross-price elasticities for two pairs of very similar cereals. The model that uses the embedding delivers relatively high cross-elasticities between Honey Nut Cheerios and Honey Graham Oh!s (0.095) and between Cocoa Pebbles and Cocoa Krispies (0.106), and relatively low cross-elasticities between

---

<sup>15</sup>Data on cereals’ nutritional characteristics are collected from the Nutritionix database. We rescale the sugar, fiber, and calorie measures to have mean zero and unit variance.

FIGURE 3: Elasticity estimates for the log-linear model



Panel A shows the density of own-price elasticities  $\beta_j$  for the log-linear model (Equation (1)). Panel B shows  $f(d)$  of Equation (2) implied by the estimated  $\gamma$  parameters. The two parallel distance functions for the model based on observed characteristics represent estimated distances for products in the same vs. different categories (Kids, Adult, All Family).

dissimilar pairs (e.g., 0.045 between Honey Nut Cheerios and Cocoa Pebbles). By contrast, the model that uses observed characteristics produces cross-elasticities in a narrow range (0.040 to 0.069 for the example products in the table), and the cross-elasticities are actually highest for dissimilar products.

TABLE 3: Comparison of Elasticities between Similar Products – log-linear model

Cereal		1	2	3	4
Honey Nut Cheerios	1	-2.936	0.095	0.045	0.046
		-2.903	0.040	0.065	0.060
Honey Graham Oh!s	2	0.095	-1.734	0.043	0.044
		0.040	-1.653	0.041	0.043
Cocoa Pebbles	3	0.045	0.043	-3.301	0.106
		0.065	0.041	-3.277	0.069
Cocoa Krispies	4	0.046	0.044	0.106	-2.503
		0.060	0.043	0.069	-2.432

The table reports in each cell elasticities  $e_{jk}$  corresponding to the row model  $j$  and the column model  $k$ . Cells contain elasticities estimated with the log-linear model of Equation 1, with distances based either on the embedding (on top) or on observed characteristics (on bottom).

In addition to generating more plausible elasticity estimates, the specification based on the embedding also delivers a better fit of the data. When we use distances computed from the embedding, the  $R^2$  from regression (1) is 0.838, vs. 0.837 if we use distances based on observed characteristics. We can compare these to a baseline case where the distance function is simply a constant (meaning that all pairs of products are estimated to have the same cross-price elasticity), for which the  $R^2$  is 0.834. On the one hand, these numbers imply that making cross-elasticities a function of product

distances does little to improve the overall fit of the model. This is perhaps not surprising, since most of the variation in sales quantities is likely induced by changes in products’ own prices. On the other hand, the (small) gain from incorporating distances from the embedding is more than 1.5 times larger than the analogous gain from incorporating distances based on observed characteristics.

### 3.5 BLP demand estimates

To show how the embedding can be used in a characteristics space model, we estimate a standard BLP demand system similar to [Nevo \(2001\)](#) and [Backus et al. \(2021\)](#). We refer the reader to those articles for a full description of the micro-foundations of the model; in this section we briefly describe the details of our implementation. Our specification includes market and product fixed effects as variables that enter consumers’ indirect utility linearly. These fixed effects capture some unobserved determinants of utility, as in [Nevo \(2001\)](#). Variables that have a non-linear impact on demand are price, the constant, and product characteristics: either the product’s content of sugar, fiber, calories from fat, and indicators for whether the cereal is for kids or all-family (in the observable characteristics case), or the coordinates from a six-dimensional embedding calculated from the survey triplets. For convenience we will refer to the former model as Characteristics BLP and the latter as Embedding BLP. Aside from the different characteristics ( $x$  variables), everything in the two specifications is identical.

The effect of non-linear variables on demand is modeled via random coefficients in the indirect utility of a household  $i$ . These coefficients are  $\beta_i \sim N(\beta + \Pi D_i, \Sigma)$ , where  $\beta$  and  $\Pi$  are vectors of parameters, and  $D_i$  are demographic characteristics of household  $i$ . We estimate the diagonal elements in  $\Sigma$  corresponding to each non-linear variable.<sup>16</sup> The model includes demographic interactions  $\Pi D_i$  for prices and the nutritional variables (or embedding coordinates), with log household income and an indicator for the presence of children in the household as the included demographic characteristics.<sup>17</sup> We estimate a log-normal income distribution with/without kids and a binomial distribution for the presence of kids from the households in the Nielsen Consumer Panel data. Values of  $D_i$  correspond to 200 Halton draws per market from these distributions.

Instruments are needed to identify and estimate this model. To this aim, we create the quadratic differentiation IVs of [Gandhi and Houde \(2020\)](#). For  $\delta_{jk}(l) = x_{jl} - x_{kl}$ , given characteristic  $l$  and

<sup>16</sup>In the Characteristics BLP some values of  $\Sigma$  were consistently estimated to be near zero, so in the final specification we set them to zero to aid convergence.

<sup>17</sup>We exclude the interaction of demographics and the constant as this largely drives outside shares, and we have already calibrated market size at the market level.

products  $j, k$ , define:

$$z_{jt}^{quad} = \left\{ \sum_k \delta_{jk}^2(l), \sum_k \delta_{jk}(l) \times \delta_{jk}(\ell) \right\} \forall (l, \ell),$$

where  $l, \ell$  are the non-linear characteristics (price and observable characteristics or embedding coordinates). We then follow [Backus et al. \(2021\)](#) in interacting these variables with moments of the demographics in each market, taking the 10th, 50th, and 90th percentile incomes for households with and without children as well as the percentage of households with children. We construct thus a total of 168 instruments as follows:

$$z_{jt} = z_{jt}^{quad} \times \left\{ 1, \%kids_t, inc_t^{10\%,k}, inc_t^{50\%,k}, inc_t^{90\%,k}, inc_t^{10\%,nk}, inc_t^{50\%,nk}, inc_t^{90\%,nk} \right\}.$$

After estimating the model using 2-step GMM, we then utilize the approximation to the optimal instruments of [Reynaert and Verboven \(2014\)](#) to refine the results.

To estimate a discrete-choice demand model we also need to specify market size. We follow [Backus et al. \(2021\)](#) in estimating the market size as the number of individuals entering the store, using variation in purchases of staple products (milk and eggs) as predictors. For retailer-DMA  $c$  and week  $w$ , we estimate:

$$\ln(q_{cw}) = b_1 \ln(q_{cw}^{\text{milk}}) + b_2 \ln(q_{cw}^{\text{eggs}}) + \delta_c + \varepsilon_{cw},$$

where  $q_{cw}$  indicates the total quantity of cereals sold. We can thus estimate market size  $\hat{M}$  as<sup>18</sup>

$$\ln(\hat{M}_{cw}) = \lambda \cdot \left( \hat{b}_1 \ln(q_{cw}^{\text{milk}}) + \hat{b}_2 \ln(q_{cw}^{\text{eggs}}) + \hat{\delta}_c \right).$$

To keep computation manageable, we estimate this model on a subsample of our data. We limit the sample successively to (i) the top 15 DMAs by market sales, (ii) the top 15 retailers within that set of DMAs, and (iii) a random set of 20 weeks. Our final subsample for BLP estimation contains 32,385 observations, with 540 unique retailer-DMA-week markets.

Parameter estimates are reported in Table 4. The two specifications deliver similar results in most respects: price coefficients are negative and significant; the interactions of price and income are positive and significant; and the random coefficients on the constant and on price are statistically significant.<sup>19</sup> More importantly, the implied elasticities are similar in magnitude and positively correlated: the median own-price elasticity in the Characteristics BLP is -2.233, vs. -2.352 in the Embedding

<sup>18</sup>As in [Backus et al. \(2021\)](#), we scale  $\lambda$  so the the average outside good share  $s_{0t} = 0.723$ .

<sup>19</sup>The standard errors reported in the table for the Embedding BLP specification do not account for sampling error in the embedding coordinates; the embedding is simply treated as data. In future drafts we plan to bootstrap the embedding coordinates (by resampling the triplets data) and compute adjusted standard errors.

BLP;<sup>20</sup> and the correlation between the two specifications' own-price elasticities is 0.946. For cross-price elasticities, the medians are 0.016 and 0.009 (respectively), and the correlation is 0.704.

TABLE 4: Estimated Coefficients of BLP Model

Parameter	Variable	Characteristics		Embeddings	
$\beta$	Price	-2.667		-3.093	
		(0.363)		(0.301)	
$\Sigma$	Constant	3.766		4.193	
		(1.271)		(0.453)	
	Price	0.820		0.946	
		(0.036)		(0.037)	
	$x_{j1}$	-		0.015	
				(0.192)	
	$x_{j2}$	0.016		0.0003	
		(0.026)		(0.401)	
	$x_{j3}$	-		0.0002	
				(0.095)	
	$x_{j4}$	0.090		0.828	
		(0.099)		(0.194)	
	$x_{j5}$	-		0.0003	
				(0.193)	
	$x_{j6}$	-		1.572	
				(0.186)	
$\Pi$		Income	Kids	Income	Kids
	Price	0.121	-19.966	0.141	-0.096
		(0.035)	(0.00)	(0.027)	(0.064)
	$x_{j1}$	-0.169	-0.809	0.139	-0.081
		(0.018)	(0.00)	(0.015)	(0.021)
	$x_{j2}$	0.135	-	0.058	-0.087
		(0.019)		(0.027)	(0.034)
	$x_{j3}$	0.003	-	-0.139	0.072
		(0.020)		(0.023)	(0.033)
	$x_{j4}$	0.060	-	0.029	-0.151
		(0.140)		(0.021)	(0.039)
	$x_{j5}$	0.104	-	0.057	-0.107
		(0.137)		(0.018)	(0.025)
	$x_{j6}$	-	-	-0.153	0.226
				(0.033)	(0.044)
Observations		32,385		32,385	
Non-linear Variables		Observables		6D Embedding	
Median Own-price Elasticity		-2.352		-2.233	
Median Outside Diversion		0.150		0.311	

The table reports estimates (on top) and standard errors (below) for the parameters of the BLP model. Columns 1 refer to the specification that uses data on product characteristics. Columns 2 refer to the specification that uses embedding coordinates. Observable characteristics  $x_{j1}$  through  $x_{j5}$  refer to sugar, fiber, calories from fat, and indicators for whether the cereal is for kids or an all-family cereal.

Table 5 shows own- and cross-price elasticity estimates for the same examples as in Table 3 above.

<sup>20</sup>For comparison, Backus et al. (2021) get a median own-price elasticity of -2.665.

As with the log-linear product space model, the specification that uses the embedding delivers more plausible substitution patterns. For similar cereals, cross-elasticities from the Embedding BLP are higher than from the Characteristics BLP (e.g. 0.216 vs. 0.158 for the cross-elasticity between Honey Graham Oh!s and Honey Nut Cheerios); and for dissimilar cereals they are lower (e.g. 0.018 vs. 0.043 for the cross-elasticity between Honey Graham Oh!s and Cocoa Krispies). Both specifications show some evidence of logit-style substitution patterns, with generally higher diversion to products with high market shares (e.g. Honey Nut Cheerios), but much less so for the Embedding BLP.

TABLE 5: Comparison of Elasticities between Similar Products – BLP

Cereal		1	2	3	4
Honey Nut Cheerios	1	-2.378	0.034	0.014	0.017
		-2.483	0.024	0.021	0.035
Honey Graham Oh!s	2	0.216	-2.798	0.030	0.018
		0.158	-2.821	0.048	0.043
Cocoa Pebbles	3	0.072	0.016	-2.468	0.057
		0.120	0.040	-2.581	0.031
Cocoa Krispies	4	0.053	0.012	0.040	-1.906
		0.127	0.023	0.020	-2.338

The table reports in each cell elasticities  $e_{jk}$  corresponding to the row model  $j$  and the column model  $k$ . Cells contain elasticities from the Embedding BLP model on top, Characteristics BLP on bottom.

The Embedding BLP thus delivers elasticity estimates that are arguably more plausible—and at the very least similar—to what we obtain from observed characteristics. We take this as an encouraging result, because it means we can obtain credible estimates of demand *even when observable characteristics are unavailable*—a challenge we believe is inherent to many markets of interest.

## 4 Discussion and Extensions

### 4.1 Log-linear demand vs. BLP

For various reasons we noted above, mixed logit models like BLP have become the gold standard for estimating rich demand systems in differentiated product markets. However, in some contexts—most notably, in analyses of antitrust cases conducted by the DOJ or FTC—researchers need to obtain demand estimates more simply and more quickly than is feasible within the BLP framework. The results of our empirical exercise are encouraging in this regard. The log-linear (product space) model that uses distances from the embedding delivers elasticity estimates (both own- and cross-price) similar to those from BLP, and for some pairs of products the estimated cross-elasticities from the log-linear model are arguably even more plausible than those from BLP in our application. In other words, the simple log-linear model, augmented with crowd-sourced data on products’

locations, does a good job of approximating the state of the art, even though it is substantially easier to estimate.<sup>21</sup>

## 4.2 Incorporating more flexible distance metrics

In estimating the log-linear demand model described in Section 3.4, we calculated pairwise product distances  $d_{jk}$  as scaled Euclidean distances that equally weight each dimension of the embedding. A more flexible approach is to estimate demand regressions like Equation (1) in which different dimensions of the embedding are allowed to have different weights. For instance, the substitutability of two products may depend weakly on how close they are in the first dimension of the embedding, while depending strongly on how close they are in the second dimension. Since the dimensions of the embedding do not have natural interpretations, we may want to let the data determine which dimensions matter most for substitution.<sup>22</sup>

We explore this idea by estimating the log-linear model of Equation (1) with the same cubic polynomial distance function as in Equation (2), but defining pairwise product distances as

$$\tilde{d}_{jk} = \left[ \sum_m \omega_m (x_{jm} - x_{km})^2 \right]^{\frac{1}{2}}, \quad (3)$$

with  $\omega_1$  (the weight on the first dimension) normalized to one, and the remaining  $\omega_m$  coefficients left as parameters to be estimated for all other embedding dimensions  $m$ .

A notable disadvantage of this modification is that the regression is no longer linear in the parameters. Estimating with nonlinear least squares increases the computational burden, but can still be done with a single line of code (e.g. using Stata’s `nls` command).

We find that the modified regression using distances as defined in Equation (3) yields similar elasticity estimates and a meaningfully better fit. Parameter estimates for the baseline with Euclidean distance and for the flexible distance specifications are reported in Table 6. Interestingly, the distances in two dimensions of the embedding are estimated to be somewhat more important than the others ( $\omega_4$  and  $\omega_5$  are above 1.5), and one dimension is estimated to hardly matter at all ( $\omega_3$  is near zero). The median own- and cross-price elasticities from this model were  $-2.475$  and  $0.051$  (com-

---

<sup>21</sup>Estimates of the log-linear demand models took less than 20 seconds to compute on a Windows desktop. Estimates of the BLP models took over 100 times longer, even when using the limited sample and running on a powerful Linux server. But differences in computation time understate the overall difference in time and complexity between the two approaches, since arriving at reliable BLP estimates requires considerable back-and-forth on things like start values, scaling, etc., even with the aid of helpful software packages like pyBLP (Conlon and Gortmaker, 2020).

<sup>22</sup>We thank Rob Porter for suggesting this enrichment of the model.

pared to  $-2.486$  and  $0.051$  from the baseline specification), and the correlations of the own- and cross-price elasticities with those from our baseline specification were  $0.997$  and  $0.931$ , respectively. The more flexible model has an  $R^2$  of  $0.980$ , compared with  $0.838$  for the baseline specification.

TABLE 6: Comparison of Baseline and Flexible Distance Results

Parameter	Variable	Baseline (1)		Flexible (2)	
		Estimate	SE	Estimate	SE
$\gamma$	$d_{jk}^0$	0.121	0.002	0.123	0.002
	$d_{jk}^1$	$-0.263$	0.013	$-0.293$	0.011
	$d_{jk}^2$	0.346	0.026	0.440	0.023
	$d_{jk}^3$	$-0.191$	0.016	$-0.257$	0.015
$\omega_m$	$x_{j1}$	1	—	1	—
	$x_{j2}$	1	—	0.983	0.052
	$x_{j3}$	1	—	0.000	0.019
	$x_{j4}$	1	—	1.791	0.102
	$x_{j5}$	1	—	1.592	0.083
	$x_{j6}$	1	—	0.994	0.055
Observations		684,476		684,476	
R-squared		0.838		0.973	
Median Own-Elast		$-2.486$		$-2.475$	
Median Cross-Elast		0.051		0.051	

The table reports estimates of  $\gamma$  and  $\omega_m$  parameters from Equations (2) and (3). Column 1 refers to the baseline specification of the model, which uses Euclidean distances in the  $f$  function. Column 2 refers to the flexible model using the specification of Equation (3).

### 4.3 An embedding based on purchase correlations

The ideal scenario for a researcher aiming to estimate substitution patterns is to have price and quantity data paired with actual data on consumers’ second (and third and fourth...) choices (see e.g., [Berry et al., 2004](#)). Such “second-choice data” can be used to generate additional moments that, when combined with the BLP moment conditions, discipline the estimates to better predict actual patterns of substitution. While we do not have second-choice data for our empirical application to cereal, we can borrow an idea from [Atalay, Frost, Sorensen, Sullivan, and Zhu \(2022\)](#) that uses Nielsen’s Consumer Panel data to learn which products households consider to be substitutes.

[Atalay et al. \(2022\)](#) use the Consumer Panel data to determine sets of products that are ever purchased by the same household across a large number of shopping trips, and then gauge the substitutability of a given pair of products by how commonly the two products are purchased by the same household. The underlying premise is that if individuals within each household have preferences over products’ characteristics and these preferences are stable over time, then temporary changes in relative prices (e.g. due to periodic sales or stockouts) will induce consumers to

occasionally purchase substitutes for their preferred product. In our case, if a household sometimes purchases Frosted Flakes and sometimes purchases Froot Loops, but never purchases Raisin Bran, the implication is that Froot Loops is a closer substitute to Frosted Flakes than Raisin Bran for that household.

This idea is formalized by constructing a dissimilarity matrix  $\mathbf{D}$  with  $1 - \rho_{jk}$  as its  $(j, k)$ -th element, where  $\rho_{jk}$  is the pairwise purchase correlation between products  $j$  and  $k$ —i.e., a measure of how likely a household is to have ever purchased product  $k$  conditional on having ever purchased product  $j$ . An embedding can then be computed based on this dissimilarity matrix; we do this using the tSNE algorithm [Van der Maaten and Hinton \(2008\)](#).<sup>23</sup> A two-dimensional embedding is shown in Figure 4. As with the embedding based on the survey triplets, it clusters similar products together, such as sugary cereals in the northwest quadrant.

FIGURE 4: Two-dimensional embedding based on Consumer Panel



The figure shows a two-dimensional embedding for ready-to-eat cereals estimated from the Consumer Panel micro-data.

If we estimate our product-space demand model using distances from this alternative embedding, we get reasonably similar estimates of products' own- and cross-price elasticities. The magnitudes

<sup>23</sup>tSNE is analogous to tSTE, except that instead of triplets it uses feature data or (in our case) data on products' distances or dissimilarities to compute the embedding.

are similar,<sup>24</sup> and more importantly they are positively correlated with the elasticities we estimate using the embedding based on survey triplets. The correlation of the own-price elasticities is 0.987, and of cross-price elasticities is 0.507.

Thus, the distances from the survey-based embedding deliver results similar to those that would result from an embedding computed from micro-data on consumers’ actual choices. We interpret this as further validation of our approach. When data that directly reflect consumers’ substitution choices are available (e.g., second-choice survey data as in [Grieco, Murry, and Yurukoglu \(2021\)](#) or household panel data as in [Atalay et al. \(2022\)](#)), it certainly makes sense to use those data. But in the absence of such data our method is a viable alternative.

## 5 Conclusion

The demand estimation toolkit available to empirical researchers in industrial economics has seen many advances in the last few decades. In particular, we have learned how to specify, identify and estimate more and more flexible models. Complementary to this line of work, in this paper we do not propose new modeling approaches but instead introduce a new source of data: triplet comparisons obtained from an online survey. We use these data to compute an embedding that represents the latent product space. To showcase the usefulness of the data, we use the embedding in conjunction with data on prices and quantities to estimate two specifications: a simple log-linear model of demand, and a BLP model. In an application to the ready-to-eat cereals market, our method produces estimates that compare favorably with those obtained using standard data on product characteristics.

Beyond our illustrative application, embeddings are particularly valuable in empirical settings where characteristics are hard to observe or measure, thus making standard demand models hard to estimate. In future work, we plan to use the method to estimate demand in an important digital market: the market for mobile apps. Recovering credible substitution patterns in this market is essential to answer policy-relevant questions about market power and the effects of consolidation, but conventional discrete-choice methods are hard to apply because demand-relevant characteristics are difficult to define and measure. Our method promises to be a useful alternative in this setting.

---

<sup>24</sup>The mean own-price elasticity is -2.42 (identical to when we use the embedding based on survey triplets), and the mean cross-price elasticity is 0.052 (compared to 0.053).

## References

- ANDERSON, E. AND N. J. VILCASSIM (2001): “Structural demand models for retailer category pricing,” *London Business School Mimeo*.
- ARMONA, L., G. LEWIS, AND G. ZERVAS (2021): “Learning Product Characteristics and Consumer Preferences from Search Data,” in *Proceedings of the 22nd ACM Conference on Economics and Computation*, 98–99.
- ATALAY, E., E. FROST, A. SORENSEN, C. SULLIVAN, AND W. ZHU (2022): “Large Scale Estimation of Demand and Markups,” .
- BACKUS, M., C. CONLON, AND M. SINKINSON (2021): “Common ownership and competition in the ready-to-eat cereal industry,” Tech. rep., National Bureau of Economic Research.
- BAJARI, P. L., Z. CEN, V. CHERNOZHUKOV, M. MANUKONDA, J. WANG, R. HUERTA, J. LI, L. LENG, G. MONOKROUSSOS, S. VIJAYKUNAR, ET AL. (2021): “Hedonic prices and quality adjusted price indices powered by AI,” *Cemmap working paper*.
- BERRY, S., J. LEVINSOHN, AND A. PAKES (1995): “Automobile prices in market equilibrium,” *Econometrica*, 841–890.
- (2004): “Differentiated products demand systems from a combination of micro and macro data: The new car market,” *Journal of political Economy*, 112, 68–105.
- BERRY, S. T. AND P. A. HAILE (2021): “Foundations of demand estimation,” in *Handbook of Industrial Organization*, Elsevier, vol. 4, 1–62.
- CHRISTENSEN, L. R., D. W. JORGENSON, AND L. J. LAU (1975): “Transcendental logarithmic utility functions,” *The American Economic Review*, 65, 367–383.
- CONLON, C. AND J. GORTMAKER (2020): “Best practices for differentiated products demand estimation with pyblp,” *The RAND Journal of Economics*, 51, 1108–1161.
- CONLON, C., J. MORTIMER, AND P. SARKIS (2021): “Estimating Preferences and Substitution Patterns from Second Choice Data Alone,” *working paper*.
- DEATON, A. AND J. MUELLBAUER (1980): “An almost ideal demand system,” *The American Economic Review*, 70, 312–326.
- DIEWERT, W. E. (1971): “An application of the Shephard duality theorem: A generalized Leontief production function,” *Journal of political Economy*, 79, 481–507.

- GABEL, S. AND A. TIMOSHENKO (2022): “Product choice with large assortments: A scalable deep-learning model,” *Management Science*, 68, 1808–1827.
- GANDHI, A. AND J.-F. HOUDE (2020): “Measuring Firm Conduct in Differentiated Products Industries,” .
- GANDHI, A. AND A. NEVO (2021): “Empirical models of demand and supply in differentiated products industries,” in *Handbook of Industrial Organization*, Elsevier, vol. 4, 63–139.
- GOETTLER, R. L. AND R. SHACHAR (2001): “Spatial competition in the network television industry,” *RAND Journal of Economics*, 624–656.
- GORMAN, W. M. (1959): “Separable utility and aggregation,” *Econometrica*, 469–481.
- GRIECO, P. L., C. MURRY, AND A. YURUKOGLU (2021): “The evolution of market power in the US auto industry,” Tech. rep., National Bureau of Economic Research.
- HAN, S., E. H. SCHULMAN, K. GRAUMAN, AND S. RAMAKRISHNAN (2021): “Shapes as Product Differentiation: Neural Network Embedding in the Analysis of Markets for Fonts,” *arXiv preprint arXiv:2107.02739*.
- HAUSMAN, J., G. LEONARD, AND J. D. ZONA (1994): “Competitive analysis with differentiated products,” *Annales d’Economie et de Statistique*, 159–180.
- HAUSMAN, J. A. AND G. K. LEONARD (2007): “Estimation of patent licensing value using a flexible demand specification,” *Journal of Econometrics*, 139, 242–258.
- HITSCH, G. J., A. HORTACSU, AND X. LIN (2019): “Prices and promotions in us retail markets: Evidence from big data,” Tech. rep., National Bureau of Economic Research.
- HOUDE, J.-F. (2012): “Spatial differentiation and vertical mergers in retail markets for gasoline,” *American Economic Review*, 102, 2147–82.
- JAFFE, S. AND S. D. KOMINERS (2012): “Discrete choice cannot generate demand that is additively separable in own price,” *Economics Letters*, 116, 129–132.
- JAFFE, S. AND E. G. WEYL (2010): “Linear demand systems are inconsistent with discrete choice,” *The BE Journal of Theoretical Economics*.
- KUMAR, M., D. ECKLES, AND S. ARAL (2020): “Scalable bundling via dense product embeddings,” *arXiv preprint arXiv:2002.00100*.
- LANCASTER, K. J. (1966): “A new approach to consumer theory,” *Journal of political economy*, 74, 132–157.

- McFADDEN, D. (1974): “The measurement of urban travel demand,” *Journal of Public Economics*, 3, 303–328.
- McFADDEN, D. AND K. TRAIN (2000): “Mixed MNL models for discrete response,” *Journal of Applied Econometrics*, 15, 447–470.
- NEVO, A. (2001): “Measuring market power in the ready-to-eat cereal industry,” *Econometrica*, 69, 307–342.
- PETRIN, A. (2002): “Quantifying the benefits of new products: The case of the minivan,” *Journal of political Economy*, 110, 705–729.
- PINKSE, J. AND M. E. SLADE (2004): “Mergers, brand competition, and the price of a pint,” *European Economic Review*, 48, 617–643.
- PINKSE, J., M. E. SLADE, AND C. BRETT (2002): “Spatial price competition: a semiparametric approach,” *Econometrica*, 70, 1111–1153.
- REYNAERT, M. AND F. VERBOVEN (2014): “Improving the performance of random coefficients demand models: the role of optimal instruments,” *Journal of Econometrics*, 179, 83–98.
- ROSSI, P. E. (2014): “Even the rich can make themselves poor: A critical examination of IV methods in marketing applications,” *Marketing Science*, 33, 655–672.
- RUIZ, F. J., S. ATHEY, AND D. M. BLEI (2020): “Shopper: A probabilistic model of consumer choice with substitutes and complements,” *The Annals of Applied Statistics*, 14, 1–27.
- VAN DER MAATEN, L. AND G. HINTON (2008): “Visualizing data using t-SNE,” *Journal of Machine Learning Research*, 9.
- VAN DER MAATEN, L. AND K. WEINBERGER (2012): “Stochastic triplet embedding,” in *2012 IEEE International Workshop on Machine Learning for Signal Processing*, IEEE, 1–6.
- WILBER, M., I. KWAK, AND S. BELONGIE (2014): “Cost-effective hits for relative similarity comparisons,” in *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, vol. 2, 227–233.

## Appendix A IV and AIDS specifications

We discuss in this appendix two important extensions to the log-linear specification describe in the paper.

### A.1 Instrumental Variable Estimation of the Log-linear Model

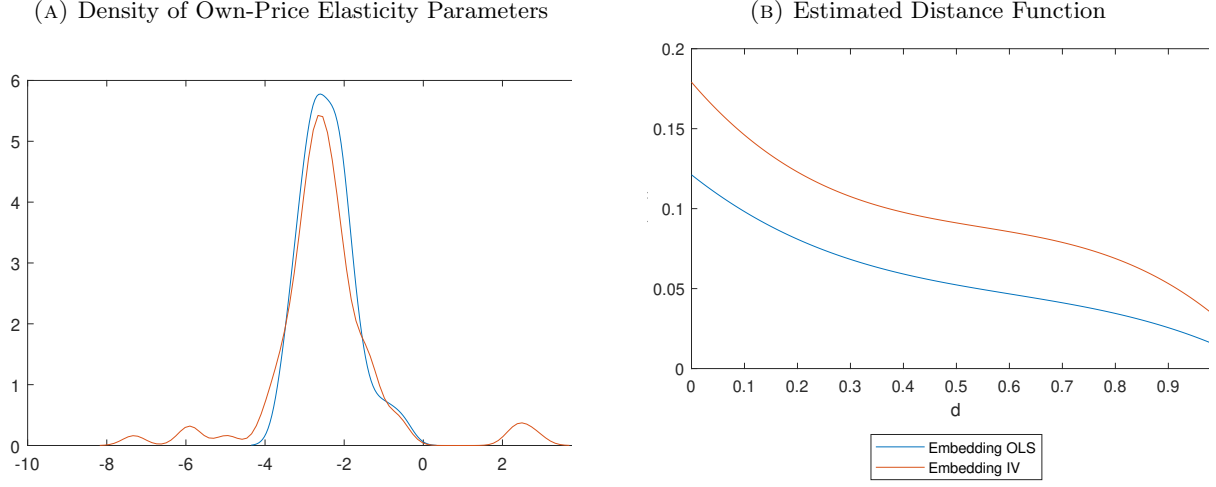
First, we present the results of a specification where we instrument for prices. Although we have argued in Section 3.4 that our particular empirical context may suggest using OLS for estimation, endogeneity of prices due to the simultaneous nature of market equilibrium outcomes, prices and quantities, is a fundamental aspect of the identification of demand systems. Hence, it is useful to discuss how to incorporate IV estimation in specifications that use embeddings data. For our empirical environment, we choose to use Hausman instruments, i.e., the prices of the same goods in other markets. Similar to [Hausman and Leonard \(2007\)](#), these instruments are valid in our context of weekly data, as factors such as national advertising campaigns – which could endanger validity – are controlled for by time fixed effects. Using instruments in unrestricted product-space demand specifications with many goods may give rise to econometric difficulties, as many instruments that vary independently are required to identify parameters ([Gandhi and Nevo, 2021](#)). The role of distances in disciplining substitutions in our log-linear specification helps substantially: we only require instruments to identify a limited number of parameters.

We present in Figure 5 the results from estimating the log-linear model in Equation (1) using Hausman instruments for log price variables. The figure shows that IV estimates of own-price elasticities are comparable to the OLS estimates, although they present more outliers (including a few products with upward sloping demand). Median price elasticity is  $-2.590$  for this specification, close to the OLS result. Despite the restrictions on the demand system made possible by the use of the embeddings data, the IV estimator may still struggle to precisely identify all parameters. The distance function implied by the IV estimates has a similar shape to the one generated by the OLS estimates, but is shifted upwards. Thus, IV estimates generate somewhat larger cross-elasticities in this application – the median cross-price elasticity is  $0.090$ . Overall, the demand system generated by the IV estimates is economically similar to the OLS demand system.

### A.2 AIDS Specification

As another important extension, we use embeddings data in a micro-founded product-space demand specification: the AIDS model of [Deaton and Muellbauer \(1980\)](#). To do so, we first transform the

FIGURE 5: Estimates for the IV Log-linear Model



Panel A shows the density of own-price elasticities  $\beta_j$  for the log-linear model (Equation (1)). Panel b shows  $f(d)$  of Equation (2) implied by the estimated  $\gamma$  parameters. OLS (IV) estimates are in blue (orange).

data to obtain products' revenue shares as  $w_j = \frac{q_j p_j}{e}$  where  $e = \sum_{k=1}^J q_k p_k$ . The demand system is:

$$w_j = \alpha_j + \sum_{k=1 \dots J} \beta_{jk} \ln(p_k) + \theta_j \ln\left(\frac{e}{p}\right) + \epsilon_j ,$$

where  $p$  is the Stone price index

$$\ln(p) = \sum_j \tilde{w}_j \ln(p_j) ,$$

and  $\tilde{w}_j$  is the average revenue share of product  $j$  across markets. This demand system is derived from an expenditure function that is a second-order approximation to any expenditure function (Diewert, 1971), and the demand system itself is a first-order approximation to any demand system (Deaton and Muellbauer, 1980).

Further economic properties that are normally imposed on this demand system include adding up, so that  $\sum_i \alpha_i = 0, \sum_i \tilde{\beta}_{ij} = 0, \forall j$ ; homogeneity, or  $\sum_i \tilde{\beta}_{ji} = 0, \forall j$ ; and symmetry, or  $\beta_{ij} = \frac{1}{2}(\tilde{\beta}_{ij} + \tilde{\beta}_{ji}) = \beta_{ji}$ . An appealing feature of the AIDS demand system is that it allows the researcher to model consumers' choice problem in a hierarchical way—that is, as a multi-stage budgeting problem (Gorman, 1959). Hence, the demand system described above can be interpreted as “conditional,” describing demand for a product in a certain category conditional on the expenditure in that category. However, that expenditure is also endogenous. To determine unconditional demand, one needs to also model the “top-level” demand equation. To do so in a scanner data context, Hausman

and Leonard (2007) propose a specification:

$$\ln(Q_t) = \delta_{0t} + Z_t\theta + \delta_1\ln(p_t) + \lambda\ln(E_t) + \eta_t$$

where  $Q_t$  is total category quantity in a certain market,  $p_t$  is the category price index, and  $E_t$  is total expenditure in market  $t$  across categories.

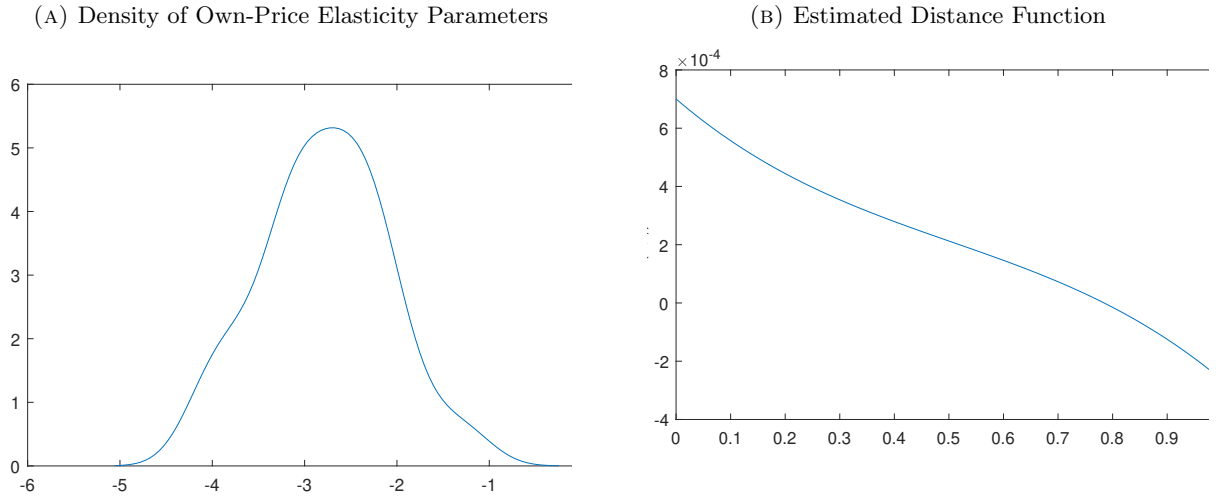
Incorporating distances from embeddings in the AIDS model enables us to restrict coefficients as  $\beta_{ij} = f(d_{ij})$ , or  $\beta_{ij} = \beta_i f(d_{ij})$ . Hence, the main equation of the demand system becomes:

$$w_j = \alpha_j + \beta_j^{own}\ln(p_j) + \beta_j^{cross} \sum_{k=1 \dots J} f(d_{jk})\ln(p_k) + \theta_j \ln\left(\frac{e}{p}\right) + \epsilon_j, \quad (4)$$

and the economic assumptions can be added to discipline parameters  $\beta_j$ .

We estimate Equation (4) using the data of our empirical application. Figure 6 reports estimates for this model. Overall, own-price elasticities are comparable to the log-linear specification, with a median value of  $-2.781$ , but with a larger variance across products. While the distance function is not comparable to the one estimated for the log-linear model due to different scale, it still comes out monotonically decreasing.

FIGURE 6: Estimates for the AIDS Model



Panel A shows the density of own-price elasticities  $\beta_j^{own}$  for the AIDS model (Equation (4)). Panel b shows  $f(d)$  of Equation (2) implied by the estimated  $\gamma$  parameters.

## Appendix B Additional figures

FIGURE 7: Survey intro page

**Cereal Survey**

We are collecting data on the similarity of different kinds of breakfast cereal sold in the United States. You will see a sequence of 20 pages asking you to indicate which two cereals are most similar to a reference brand. Even if none of the cereals are very similar to the reference brand, you will need to choose the two you think are the most similar.

Before proceeding, please answer the following questions:

**What is your 10-digit student ID?**

**How often do you eat breakfast cereal?**

- ☐ Less than once per week
- ☐ Once or twice per week
- ☐ Three or more times per week

**How many different breakfast cereals have you personally tried while living in the United States?**

- ☐ None
- ☐ Between one and ten
- ☐ More than ten

Survey respondents completed this preliminary survey before seeing the product comparison pages.