Scalable Nonparametric Price Elasticity Estimation

Jingbo Wang and Yufeng Huang *

November 1, 2022

Abstract

This paper proposes a nonparametric framework to estimate point-wise price elasticities using aggregate market data. We derive a new constructive price elasticity estimator based on the nonparametric control function framework (Newey et al., 1999), integrate the bootstrap averaged (bagged) nearest neighbors predictor (Demirkaya et al., 2022) into this framework, and implement the estimation using just-in-time compilation and parallel computation. A series of Monte Carlo simulations across a wide range of data-generating processes show that our elasticity estimator is fast to compute and achieves both precise estimates and robust inferences. In an empirical application, we demonstrate that this method can (1) flexibly estimate price response and substitution patterns and (2) directly inform optimal pricing and supply-side counterfactuals.

Key words: price elasticity; demand estimation; nonparametric control function
JEL codes: C14, L11

*Jingbo Wang is an assistant professor in the department of marketing at the Chinese University of Hong Kong; Email: jingbowang@cuhk.edu.hk. Yufeng Huang is an associate professor at the Simon Business School, University of Rochester; Email: yufeng.huang@simon.rochester.edu. We are grateful for the comments and feedback by Giovanni Compiani, Cheng Hsiao, James Powell, Chenyu Yang, Sha Yang, Guang Zeng, and seminar and conference participants at the 2021 AI and Machine Learning Conference, Chinese University of Hong Kong, University of Rochester, and University of Southern California. In particular, Jingbo Wang thanks Professor Cheng Hsiao for his generous and continuous support.

1 Introduction

Understanding the price elasticities of demand is essential for firms, regulators, and researchers. For a firm, price elasticities within a product line determine optimal prices, while cross-price elasticities with competitors' products help characterize the extent of competition. Outside of a firm, price elasticities also help regulators and researchers determine the level of markups (Berry et al., 1995), measure market power (Farrell and Shapiro, 1990), and track policy outcomes.

Estimating price elasticities entails two fundamental difficulties. The first is that price experiments are rarely available at scale, and without experiments, estimating elasticities using only observational price-quantity data faces an endogeneity problem. Endogeneity arises because unobserved demand shifters influence prices and quantity simultaneously, leading to biased elasticity estimates and potentially costly pricing or policy mistakes.[1] The second difficulty is that elasticity estimates are sensitive to parametric modeling assumptions, requiring a flexible demand model to alleviate functional-form restrictions. For decades, the literature has leveraged domain knowledge to enrich parametric demand models and to show that these extensions profoundly matter for implied elasticities.[2] However, these extensions are usually limited to their respective contexts, and implementing them can often be computationally demanding. Academics and practitioners need a flexible, generalizable, and computationally attractive approach to estimating price elasticities.

To that end, this paper proposes a nonparametric framework to estimate price elasticities using aggregate market-level observational data. We construct a point-wise estimator of price elasticities using the nonparametric control function framework (Newey et al., 1999) and demonstrate that this estimator can work together with scalable machine learning prediction methods. In particular, we show that the bootstrap averaged (bagged) nearest neighbors estimator (Demirkaya et al., 2022) applies well to this framework, given its known asymptotic properties and attractive finite-sample performance. We further boost the scalability of our estimator by utilizing its closed-form representation and recent developments in just-in-time parallel programming. The resulting price elasticity estimator demonstrates desirable performance in simulations across a broad set of data-generating processes as well

[1] Even if a firm cannot set optimal prices, price endogeneity is still a concern if the firm reacts to demand shifters unobserved to the researcher.

[2] To cite a few examples, Berry et al. (1995) capture heterogeneity using random-coefficient logit models, Erdem et al. (2003) and Hendel and Nevo (2006) capture consumer stockpiling, Kim et al. (2002) and Dubé (2004) model consumer purchases of multiple goods and multiple units, and Seiler (2013) characterizes limited consumer awareness of price discounts.

as a real-world application.

First, we adapt Newey et al. (1999)'s nonparametric triangular simultaneous equation framework to the setting of demand estimation, leading to a system of two equations. In the first equation, sales quantity is a function of prices and product characteristics. In the second equation, prices are influenced by exogenous instrumental variables. Under control function assumptions, we show that one can *constructively* identify price sensitivities by removing the impact of unobserved demand confounders, which are the source of price endogeneity. This result has intuitive interpretations that connect with the causal directed acyclic graphs (Pearl, 2009) and, more importantly, implies that one can constructively and directly estimate price elasticities as a nonlinear function of conditional expectations.

We suggest scalable machine learning methods for these conditional expectations. In particular, the bootstrap averaged (bagged) nearest neighbors estimator has two attractive features ideal for our nonparametric framework. First, one can use the jackknife to reduce finite-sample bias and avoid bias accumulation in nonlinear transformations. Second, one can use the standard bootstrap procedure for inference, because the bootstrap commutes with smooth nonlinear functions. On top of the theoretical advantages, this paper integrates bagged nearest neighbors into elasticity estimation and significantly boosts the scalability of the estimation by utilizing closed-form representation, just-in-time compilation, and parallel computation.

We estimate price elasticities in Monte Carlo simulations across a wide range of data-generating processes and demonstrate that our approach has desirable performance. These simulations mimic realistic settings and include cases where the data-generating process is a logit model (McFadden, 1973; Berry, 1994), a random-coefficient logit model (Berry et al., 1995; Rossi et al., 1996), a model of complements with bundled purchases (Gentzkow, 2007), and a model of multiple discrete-continuous choices with a budget constraint (Bhat, 2008; Kim et al., 2002). In each case, we show that our estimator recovers the point-wise own- and cross-elasticities well, and that our proposed bootstrapped standard errors deliver valid inferences. Our estimator also demonstrates robust performance in a wide range of sample sizes and levels of dimensionality—in particular, we can estimate point elasticities across many products or with small samples.

We next demonstrate how our method can be applied in an empirical setting, focusing on yogurt products from two leading national brands that come in different package sizes. Academics and practitioners can use our method at least in two ways. The first is to estimate a price elasticity profile "globally" to test a theory or to recover the shape of a

demand curve. By estimating own- and cross-price elasticities point-by-point across the data support, we find nuanced patterns in how these elasticities vary between products and over different price points. For example, both the own-demand and the own-price-elasticity curves are downward-sloping in price (Nocke and Schutz, 2018), and substitution is the strongest between products by different brands in the same package size. Recall that we obtain these results without imposing ex-ante assumptions on consumers' preference structures. As such, these findings can guide modeling assumptions and test theory predictions.

Another use of our method is to estimate price elasticities "locally" as needed, along the path of profit maximization. This use-case arises when demand is not the focus *per se* but is a means towards maximizing profits or evaluating supply-side policy counterfactuals (where the demand function is unchanged). We demonstrate that our method can be embedded in the firm's profit-maximization algorithm, which estimates point-wise derivatives of the profit function as needed. In our empirical application, such an algorithm only needs to evaluate local elasticities at a handful of points before it arrives at the optimal prices, allowing pricing decisions to be made almost in real-time. Policymakers can also use our method to compute optimal prices under different counterfactual ownership structures, allowing them to analyze a merger policy's impact on equilibrium prices.

This paper primarily contributes to the booming literature on nonparametric demand estimation. The closest work, Compiani (2022), formulates a micro-founded, differentiated-good demand system (Berry et al., 1995; Berry and Haile, 2014) into a nonparametric IV problem and approximates it using Bernstein polynomials. Our paper presents a new point-wise estimation framework based on nonparametric control functions and is an alternative to Compiani (2022). Whereas our framework does use standard control function assumptions, a crucial difference is that it does not rely on Berry and Haile's linear index assumption (i.e., utility is separable in one observed characteristic and the unobserved characteristics) to invert the demand function. In addition, our point-wise estimation routine offers considerable computational advantages and allows one to scale up the demand estimation exercise to much larger samples, and to accommodate higher dimensions.

The remainder of this paper is organized as follows. Section 2 reviews the related literature. Section 3 formally introduces our theoretical framework. Section 4 provides details on our estimation and inference strategy. We further demonstrate the performance of our method with a series of Monte Carlo simulations in Section 5. Section 6 applies our method to estimate own- and cross-price elasticities for yogurt and demonstrates how this method can aid pricing and merger policies. Section 7 concludes the paper.

2 Related literature

Our paper fits in the nonparametric demand estimation literature. Broadly speaking, this literature relaxes functional-form assumptions on demand by using regularized, nonparametric estimation methods. We contribute to the literature by achieving regularization with built-in machine learning algorithms and demonstrating a point-wise estimation route in the presence of price endogeneity.

The most closely related studies are Compiani (2022) and Compiani and Smith (2021). Compiani (2022) formulates a micro-founded differentiated-good demand system (Berry et al., 1995; Berry and Haile, 2014) into a nonparametric IV problem and approximates the system using Bernstein polynomials. In contrast, our paper formulates demand in the nonparametric control function framework. One advantage of this approach is that we do not have to assume that one characteristic is separable in the utility function in order to invert the demand function (as in Compiani 2022; Berry and Haile 2014).[3] Another advantage is that our estimator is "local" in the sense that one can use it to estimate price elasticities as needed, such as when optimizing profits or computing counterfactuals, instead of first estimating the entire "global" demand function. This feature offers a considerable computational edge.

Second, our paper is also related to the recent marketing literature on aiding managerial pricing decisions. Smith et al. (2019a) present a Bayesian shrinkage estimator to find substitutions or complementarities across product categories. Misra et al. (2019) formulate pricing into a multi-arm-bandit problem and present a framework for firms to learn the optimal price by experimentation. Cong et al. (2021) estimate demand using orthogonal random forests. Gabel and Timoshenko (2022) estimate price elasticities across a broad set of products, using deep neural networks to help managers understand the relationship between a large set of products. Dubé and Misra (2021) presents a framework that combines a micro-founded demand model with regularized machine learning. They use it to estimate demand by observed customer segments and set personalized prices. This paper contributes to previous work by presenting a machine-learning framework that (1) has valid inference, (2) can be applied to observational data with potentially endogenous prices, and (3) is fast to compute.

Third, this paper connects broadly to the demand estimation literature in marketing and

[3]Estimating the inverse demand (Compiani, 2022; Berry and Haile, 2014) does permit structural, non-separable error terms, whereas our framework assumes separable and uni-dimensional unobservables, which is standard in the control function literature.

economics. Dubé (2019) provides an excellent summary. The canonical structural model characterizes heterogeneous consumers' choices of one unit of one product from a known choice set (see, e.g., Berry et al. 1995; Rossi et al. 1996). Nevertheless, empirical studies in different industries have motivated various extensions of this canonical structure. These extensions include cases when consumers purchase multiple goods or multiple units (Hendel, 1999a; Kim et al., 2002; Dubé, 2004; Mehta et al., 2010; Chan, 2006), make decisions under limited consideration (Goeree, 2008; Joo, 2022) or search frictions (De Los Santos et al., 2012; Abaluck et al., 2022), choose among geographically-differentiated options (Houde, 2012; Magnolfi et al., 2022), and purchase complementary products (Gentzkow, 2007). Our paper offers a flexible alternative for estimating point-wise price elasticity with minimal assumptions, which can also be used to test model restrictions.

Fourth, our paper also contributes to the emerging literature on machine learning causal inference. Demirkaya et al. (2022) prove a range of statistical properties for the bagged nearest neighbors estimator (they focus on a general case of distributional nearest neighbors). In this paper, we show that point-wise price elasticity can be formulated on top of the bagged nearest neighbors algorithm under the control function framework in the presence of price endogeneity. At a high level, our framework can be seen as another case where one uses machine learning methods for first-step estimation, then establishes valid second-step inference for a low-dimensional causal parameter. Related ideas are the double machine learning framework in Chernozhukov et al. (2017), the causal forests in Wager and Athey (2018), and the deep neural networks in Farrell et al. (2021).

Finally, our framework can be used to estimate heterogeneous treatment effects (Wager and Athey, 2018). This paper offers a simple pathway to deal with continuous and endogenous treatments. Compared to the generalized random forests approach (Athey et al., 2019), our framework is fully nonparametric and allows heterogeneity on the treatment level, which is of essential interest for our price case.

3 Theoretical framework

In this section, we introduce our theoretical framework for price elasticity estimation. We first present a nonparametric demand model and list the set of assumptions. Next, we constructively identify price elasticities under endogenous prices, and we use this identification result for estimation. Finally, we intuitively interpret the identification result and connect it to the directed acyclic graph (Pearl, 2009).

3.1 Model setup

We assume that product j in market t has sales quantity (or shares) s_{jt} and price p_{jt}, for $j = 1, 2, \cdots, J$, and $t = 1, 2, \cdots, T$. Let $\mathbf{p}_t = (p_{1t}, p_{2t}, \cdots, p_{Jt})^T$ denote the price vector for products $\{1, 2, \cdots, J\}$ in market t. Here $(\cdot)^T$ is the transpose operator. Similarly, let $\mathbf{x}_{jt}$ denote the vector of the observed characteristics for product j in market t and $\mathbf{x}_t = (\mathbf{x}_{1t}^T, \mathbf{x}_{2t}^T, \cdots, \mathbf{x}_{Jt}^T)^T$ denote the stacked vector of observed product characteristics for market t. We assume that, for all j and t, the underlying relationship between sales quantity, prices, and the observed product characteristics is:

$$s_{jt} = f_j(\mathbf{p}_t, \mathbf{x}_t) + \epsilon_{jt}, \tag{1}$$

where $f_j(\cdot)$ is the demand function for product j, which is a function of prices and observed characteristics of *all* products and is stable across markets. ϵ_{jt} is the unobserved demand shock to product j in market t and has a mean of zero. In this notation, function $f_j(\mathbf{p}_t, \mathbf{x}_t)$ already absorbs market-invariant product characteristics, such as quality. Thus, $\mathbf{x}_t$ captures observed characteristics that vary across markets (such as promotion, market size, and seasonality), and ϵ_{jt} represents residual quantity shocks not captured by these observed characteristics. Our goal is to estimate the causal effect of prices on sales quantities, or the slope of demand, at point $(\mathbf{p}_t, \mathbf{x}_t)$:

$$\frac{\partial f_j(\mathbf{p}_t, \mathbf{x}_t)}{\partial p_{kt}}.$$

When $k = j$, this derivative is the own-price sensitivity. When $k \neq j$, the cross-price sensitivity measures the substitution between two products. Whereas this derivative is not unit-free, one can further normalize it by dividing the ratio between quantity and price to obtain the price elasticity.

Discussions. Before illustrating how the own- and cross-price derivatives are identified and estimated, we first highlight that these derivatives are *per se* important. To see this point, consider a multi-product monopolist firm that maximizes profits with J products in market t.[4] The firm maximizes expected profit:

$$\sum_{j=1}^{J} f_j(\mathbf{p}_t, \mathbf{x}_t) \cdot (p_{jt} - \text{mc}_{jt}),$$

[4]The illustration naturally extends to competing multi-product firms. We assume a monopolist to keep the notation simple.

where mc_{jt} is the marginal cost for product j in market t. Under regularity assumptions, which we formalize later, a necessary condition of this profit-maximization problem is to solve for the following system of first-order conditions.

$$\begin{pmatrix} p_{1t} - \text{mc}_{1t} \\ p_{2t} - \text{mc}_{2t} \\ \vdots \\ p_{Jt} - \text{mc}_{Jt} \end{pmatrix} + \begin{pmatrix} \frac{\partial f_1(\mathbf{p}_t, \mathbf{x}_t)}{\partial p_{1t}} & \frac{\partial f_2(\mathbf{p}_t, \mathbf{x}_t)}{\partial p_{1t}} & \dots & \frac{\partial f_J(\mathbf{p}_t, \mathbf{x}_t)}{\partial p_{1t}} \\ \frac{\partial f_1(\mathbf{p}_t, \mathbf{x}_t)}{\partial p_{2t}} & \frac{\partial f_2(\mathbf{p}_t, \mathbf{x}_t)}{\partial p_{2t}} & \dots & \frac{\partial f_J(\mathbf{p}_t, \mathbf{x}_t)}{\partial p_{2t}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_1(\mathbf{p}_t, \mathbf{x}_t)}{\partial p_{Jt}} & \frac{\partial f_2(\mathbf{p}_t, \mathbf{x}_t)}{\partial p_{Jt}} & \dots & \frac{\partial f_J(\mathbf{p}_t, \mathbf{x}_t)}{\partial p_{Jt}} \end{pmatrix}^{-1} \begin{pmatrix} f_1(\mathbf{p}_t, \mathbf{x}_t) \\ f_2(\mathbf{p}_t, \mathbf{x}_t) \\ \vdots \\ f_J(\mathbf{p}_t, \mathbf{x}_t) \end{pmatrix} = 0. \tag{2}$$

This system of equations defines a fixed-point solution for the price vector $\mathbf{p}_t$. Knowing the costs, the matrix of own- and cross-price derivatives allows the firm to compute the optimal price vector given a set of $(\mathbf{p}_t, \mathbf{x}_t)$. Conversely, for the policymaker who observes market outcomes $(\mathbf{s}_t, \mathbf{p}_t, \mathbf{x}_t)$ and is willing to assume optimal pricing by firms, the system directly infers markups.

3.2 Assumptions

In the demand model (1), we have not placed any assumptions on the exogeneity of prices to demand shocks ϵ_{jt}. In fact, except for rare cases with experimental variations, the price vector $\mathbf{p}_t$ often responds to unobserved demand shocks, such as unmeasured promotional activities or changes in product characteristics (e.g., packaging). This leads to price endogeneity, or $\mathbb{E}[\epsilon_{jt}|\mathbf{p}_t, \mathbf{x}_t] \neq 0$. In this paper, we build on and extend the nonparametric triangular simultaneous equations framework (Newey et al., 1999) to overcome endogeneity.

We further assume the price vector $\mathbf{p}_t$ is *related to* a vector of instrumental variables $\mathbf{z}_t$:

$$\mathbf{p}_t = \mathbf{g}(\mathbf{z}_t) + \mathbf{u}_t. \tag{3}$$

where $\mathbf{z}_t$ can include $\mathbf{x}_t$ but does not include $\mathbf{p}_t$, and function $\mathbf{g}(\cdot)$ denotes a nonparametric, *descriptive* relationship between prices and $\mathbf{z}_t$. The function $\mathbf{g}(\cdot)$ can approximate supply-side pricing behavior (such as in Berry et al., 1995), but can also come from other (potentially suboptimal) price-setting processes.[5]

We are now ready to state four identifying assumptions. Assumptions 1 and 2 are core modeling assumptions. Assumptions 3 and 4 are regularity conditions.

[5]The interpretation of $\mathbf{g}(\cdot)$ contrasts with our interpretation of the demand function $f_j(\cdot)$. The demand function is a structural object and represents the causal relationship between prices and quantities.

Assumption 1. *For the instrumental variables* $\mathbf{z}_t$*, it holds that*

$$\mathbb{E}(\mathbf{u}_t|\mathbf{z}_t) = 0.$$

Notice that $\mathbf{g}(\mathbf{z}_t)$ can be the conditional expectation of $\mathbf{p}_t$ on $\mathbf{z}_t$. Thus, this assumption can follow naturally: $\mathbf{g}(\mathbf{z}_t)$ is the projection of prices onto the space of instrumental variables, whereas $\mathbf{u}_t$ captures residual price variation orthogonal to the instruments. The residual price variation $\mathbf{u}_t$ can contain elements endogenous to demand shocks ϵ_{jt}.

Assumption 2. *Denote the excluded instruments as* $\tilde{\mathbf{z}}_t$*. For* $j = 1, 2, \cdots, J$*, it holds that*

$$\mathbb{E}(\epsilon_{jt}|\mathbf{x}_t, \tilde{\mathbf{z}}_t, \mathbf{u}_t) = \mathbb{E}(\epsilon_{jt}|\mathbf{u}_t).$$

Assumption 2 states that the residual price variation $\mathbf{u}_t$, which contains unobserved demand shocks that co-move with prices, is the source of price endogeneity. That is, conditional on $\mathbf{u}_t$, both $\mathbf{x}_t$ and $\tilde{\mathbf{z}}_t$ do not have direct links to demand shocks and thus drop out of the conditional expectation. This is the critical assumption in the triangular simultaneous equations framework by Newey et al. (1999). The essence of Assumption 2 is that the instrumental variables $\mathbf{z}_t$ do not directly impact sales quantities. In Section 3.4, we further illustrate the intuition behind this assumption using directed acyclic graphs.

Assumption 3. *The number of excluded instrumental variables* $d_{\tilde{z}}$ *is no less than the number of endogenous variables* J*, that is,* $d_{\tilde{z}} \geq J$*. Moreover, the Jacobian matrix of* $\mathbf{g}(\mathbf{z}_t)$ *with respect to* $\mathbf{z}_t$ *is of full column rank.*

Assumption 3 is a standard rank condition. It ensures sufficient variation in the instruments to drive price variation and the existence of a well-defined solution.

Assumption 4. $f_j(\mathbf{p}_t, \mathbf{x}_t)$ *for* $j = 1, 2, \cdots, J$*,* $\mathbb{E}(\epsilon_{jt}|\mathbf{u}_t)$ *for* $j = 1, 2, \cdots, J$*, and* $\mathbf{g}(\mathbf{z}_t)$*, are first-order continuously differentiable with respect to all arguments.*

Assumption 4 ensures that the partial derivatives exist and have desirable smoothness. This is a technical condition.

3.3 Identification results

We now present *constructive* identification results for the own- and cross-price sensitivities $\partial_p f_j(\mathbf{p}_t, \mathbf{x}_t)$, in the presence of endogenous prices. These results are later used directly in

our estimation. For ease of notation, we define

$$h_j(\mathbf{p}_t, \mathbf{x}_t, \tilde{\mathbf{z}}_t) := \mathbb{E}(s_{jt}|\mathbf{p}_t, \mathbf{x}_t, \tilde{\mathbf{z}}_t),$$

which is the conditional expectation of sales quantity given prices, product characteristics, and the excluded instruments. We further define the control function as the conditional expectation of demand shocks, given the residual price variation $\mathbf{u}_t$,

$$\lambda(\mathbf{u}_t) := \mathbb{E}(\epsilon_{jt}|\mathbf{u}_t),$$

Taking the conditional expectations on both sides of Equation (1), it follows that

$$\begin{aligned} h_j(\mathbf{p}_t, \mathbf{x}_t, \tilde{\mathbf{z}}_t) &= f_j(\mathbf{p}_t, \mathbf{x}_t) + \mathbb{E}(\epsilon_{jt}|\mathbf{p}_t, \mathbf{x}_t, \tilde{\mathbf{z}}_t) \\ &= f_j(\mathbf{p}_t, \mathbf{x}_t) + \lambda(\mathbf{u}_t), \end{aligned} \tag{4}$$

where the second equality comes from Assumption 2. This equality states that the conditional expectation of sales quantity, h_j, is the sum of the demand function f_j and the control function λ. The control function provides a sufficient statistic for the unobserved confounders that cause price endogeneity. Recall that $\mathbf{u}_t$ is the residual price variation orthogonal to the instruments. Therefore, controlling for $\mathbf{u}_t$ allows one to focus on price variation driven by the instruments.

We can now directly construct the own- and cross-price sensitivities as a function of components for which we can find analogs in the data. If we take the derivatives with respect to the price vector and rearrange the terms, it follows that

$$\underbrace{\partial_{\mathbf{p}_t} f_j(\mathbf{p}_t, \mathbf{x}_t)}_{J\times 1} = \underbrace{\partial_{\mathbf{p}_t} h_j(\mathbf{p}_t, \mathbf{x}_t, \tilde{\mathbf{z}}_t)}_{J\times 1} + (\underbrace{\partial_{\tilde{\mathbf{z}}_t} \mathbf{g}(\mathbf{z}_t)^T}_{J\times d_{\tilde{z}}} \underbrace{\partial_{\tilde{\mathbf{z}}_t} \mathbf{g}(\mathbf{z}_t)}_{d_{\tilde{z}}\times J})^{-1} \underbrace{\partial_{\tilde{\mathbf{z}}_t} \mathbf{g}(\mathbf{z}_t)^T}_{J\times d_{\tilde{z}}} \underbrace{\partial_{\tilde{\mathbf{z}}_t} h_j(\mathbf{p}_t, \mathbf{x}_t, \tilde{\mathbf{z}}_t)}_{d_{\tilde{z}}\times 1}, \tag{5}$$

where the terms underneath the braces indicate dimensions of the corresponding matrices. The number of excluded instruments is $d_{\tilde{z}}$ and the number of endogenous prices is J. Appendix A.1 presents the derivation details.

The left-hand side of Equation (5) is the point-wise price sensitivities. The right-hand side of Equation (5) is made up of components that can be derived directly from the data: the slope of sales quantity on price ($\partial_{\mathbf{p}_t} h_j(\mathbf{p}_t, \mathbf{x}_t, \tilde{\mathbf{z}}_t)$), price on instruments ($\partial_{\tilde{\mathbf{z}}_t} g(\mathbf{z}_t)$), and sales quantity on instruments ($\partial_{\tilde{\mathbf{z}}_t} h_j(\mathbf{p}_t, \mathbf{x}_t, \tilde{\mathbf{z}}_t)$). It directly follows that price sensitivities are identified. Formally:

Theorem 1. *Under Assumptions 1-4, $\partial_{\mathbf{p}_t} f_j(\mathbf{p}_t, \mathbf{x}_t)$ are identified.*

This constructive identification result is significant because one can use sample analogs of the right-hand side of Equation (5) to estimate price sensitivities directly. This feature is nontrivial and differentiates our framework from conventional control function approaches, where $\hat{\mathbf{u}}_t$ are explicitly estimated and then plugged back into the main equation to replace the omitted variable. As will be more apparent in Section 4, we do not need to estimate any forms of control functions. All we need for estimation is Theorem 1.

We should note that Newey et al. (1999) arrived at Equation (5) as a side theoretical result (see Equation (2.3) in their paper). However, to the best of our knowledge, neither their paper nor the subsequent literature has explored the implications of this result.

3.4 A graphical illustration

Here we demonstrate the identification results using a graphical illustration. As an aside, this illustration offers intuitive connections between our central identification results (Equation (5)), the directed acyclic graphs in the computer science literature (Pearl, 2009), and the classical instrumental variable interpretation. For ease of exposition, we further simplify Equation (5) to the case with one endogenous price and one excluded instrument. In this case, Equation (5) becomes:

$$\partial_p f(p_t, \mathbf{x}_t) = \partial_p h(p_t, \mathbf{x}_t, \tilde{z}_t) - [-\partial_{\tilde{z}_t} g(z_t)^{-1} \partial_{\tilde{z}_t} h(p_t, \mathbf{x}_t, \tilde{z}_t)]. \tag{6}$$

Figure 1 visualizes Equation (6) using a directed acyclic graph (Pearl, 2009). Our goal is to evaluate the causal effect of price on sales quantity, $\partial_p f(p_t, \mathbf{x}_t)$, which is channel (2) in Figure 1. This partial effect is a causal effect because the unobserved confounder is held fixed. However, in the observational data, the unobserved confounder affects price and sales quantity simultaneously through channels (3) and (4). When prices change in the observational data, the change in sales quantity comes from two sources. The first is movements *along* the demand function $\partial_p f(p_t, \mathbf{x}_t)$ (channel (2)), the object of our interest. The second source, channels (3) and (4), comes from price endogeneity. The data directly gives the total effect of price on sales quantity, $\partial_p h(p_t, \mathbf{x}_t, \tilde{z}_t)$.

With valid instruments and under our identifying assumptions, we can back out, in a descriptive sense, how changes in the instrumental variable relate to changes in price. That is, we can recover $\partial_z g(z_t)$ (channel (1)). Moreover, the data also identify $\partial_z h(p_t, \mathbf{x}_t, \tilde{z}_t)$, which is how the instrumental variable is correlated with sales quantity while holding the

Figure 1: An interpretation of identification

Note: This figure depicts a simplified relationship between sales quantity, price, an unobserved confounder, and instrumental variables, using directed acyclic graphs (Pearl, 2009).

price constant (channel ① → ③ ← ④).[6] The indirect effect of price on sales quantity, which is channel ③ → ④, can be worked out by using $\partial_p h(p_t, \mathbf{x}_t, \tilde{z}_t)$ to divide $\partial_{\tilde{z}_t} g(z_t)^{-1}$ and then flipping the sign. When this is done, the total effect $\partial_p h(p_t, \mathbf{x}_t, \tilde{z}_t)$ minus the unwanted indirect effect $-\partial_{\tilde{z}_t} g(z_t)^{-1} \partial_{\tilde{z}_t} h(p_t, \mathbf{x}_t, \tilde{z}_t)$ gives us the causal partial effect of price on quantity.

It is worth noting that our model, which is based on nonparametric control functions, shares the same graphical intuition as the classical instrumental variable approach. Imbens (2020) provides further discussions that compare the potential-outcome framework and direct acyclic graphs.

4 Estimation and Inference

Equation (5) provides a complete and straightforward roadmap for estimation. The left-hand side of Equation (5) is the object of interest, whereas the right-hand side involves three partial derivatives: $\partial_p \mathbf{h}_j(p_t, \mathbf{x}_t, z_t)$, $\partial_z \mathbf{g}(z_t)$, and $\partial_z \mathbf{h}_j(p_t, \mathbf{x}_t, z_t)$. By definition, both $\mathbf{h}_j(p_t, \mathbf{x}_t, z_t)$ and $\mathbf{g}(z_t)$ are conditional expectations, or equivalently, predictions.[7] Once the conditional expectation functions are known, their partial derivatives can be well-estimated

[6] Price is held constant here, thus we use the notation ③ ← ④ instead of ③ → ④. For details on these notations, see Pearl (2009).

[7] A conditional expectation is the solution of the least-squares prediction problem.

by numerical methods such as finite differences. In this sense, the estimation of price elasticities is transformed into multiple intermediate prediction tasks.

The central problem now shifts to how to make good predictions. Classical nonparametric methods, such as kernel estimation, suffer from the curse of dimensionality and are generally non-scalable to modern data sets. However, whereas modern machine-learning methods are more scalable than classical nonparametrics, not every machine-learning method can be used in our framework because price elasticity estimation places a few extra requirements on inference and implementation. First, each prediction needs to be precise and have low bias. The price sensitivity estimates are nonlinear combinations of multiple prediction terms, and the optimal prices are nonlinear functions of price sensitivities. These nonlinear transformations could magnify the initial prediction bias. Second, the firm needs to know both the recommended pricing policy and the uncertainty of this recommendation. The optimal pricing policy changes with the degree of uncertainty for the demand estimate (Dubé and Misra, 2021). This requires valid inference of the price sensitivities, i.e., the estimator should have known asymptotic properties. Third, practical model implementation requires that the model is computationally attractive and scalable to large data sets.

In this paper, we propose to use the bootstrap averaged ("bagged") nearest neighbors estimator (Biau et al., 2010) as a prediction method. In this section, we introduce the estimator and demonstrate that it satisfies the desired properties: (1) low finite-sample bias, (2) valid inference, and (3) computational scalability. We close this section by discussing other practical considerations for implementation.

4.1 The bagged nearest neighbors estimator

Here we introduce the bootstrap averaged (bagged) nearest neighbors estimator for prediction and inference. In our case, estimation requires repeated point-wise predictions of $\mathbb{E}(s_{jt}|\mathbf{p}_t, \mathbf{x}_t, \tilde{\mathbf{z}}_t)$ to obtain multiple derivatives. For exposition, we represent the general prediction goal as estimating the conditional expectation of y_i at point $\mathbf{w}_0$, $\mathbb{E}\left(y_i \,|\, \mathbf{w}_i = \mathbf{w}_0\right)$, given an i.i.d. sample of size n with $(y_i, \mathbf{w}_i)_{i=1}^n$. Here, y is a scalar, and $\mathbf{w} \in \mathbb{R}^d$ has a fixed but potentially large dimension d. For example, if the task is to predict $\mathbb{E}(s_{jt}|\mathbf{p}_t, \mathbf{x}_t, \tilde{\mathbf{z}}_t)$, then y_i is s_{jt}, and $\mathbf{w}_i$ is $(\mathbf{p}_t, \mathbf{x}_t, \tilde{\mathbf{z}}_t)$.[8]

Similar to the k nearest neighbors and the random forests estimators, the bagged nearest neighbors estimator averages across neighboring observations in the $\mathbb{R}^d$ space of $\mathbf{w}$ to pre-

[8] In this section, we use notation $i = 1, \ldots, n$ to represent an observation. This notation follows the machine learning causal inference literature.

dict $\mathbb{E}(y_i \,|\, \mathbf{w}_i = \mathbf{w}_0)$. The key difference between them is how "neighboring" observations are weighted. The k nearest neighbors estimator assigns equal weights to the nearest k observations, while the random forests estimator assigns random weights out of a splitting algorithm. For the bagged nearest neighbors estimator, the weights are analogous to drawing a subsample, pulling out the outcomes of the closest observation within this subsample, and averaging all such outcome y's from all possible subsamples.

In mathematical terms, let $\{i_1, \cdots, i_m\}$, with $i_1 < i_2 < \cdots < i_m$ and $m \leq n$, denote a subset of size m from the index set $\{1, \cdots, n\}$. With this definition, $\{(y_{i_j}, \mathbf{w}_{i_j})_{j=1}^m\}$ denotes a subsample of the data. We further denote the outcome y of the closest observation in Euclidean distance to $\mathbf{w}_0$ in this subsample as $y_{(1)}(\mathbf{w}_0; (y_{i_1}, \mathbf{w}_{i_1}), (y_{i_2}, \mathbf{w}_{i_2}), \ldots, (y_{i_m}, \mathbf{w}_{i_m}))$. Then, the bagged nearest neighbors estimator with a subsampling scale m is defined as the average of all $y_{(1)}$'s from *all* possible subsamples of size m, i.e.,

$$\tau_n(m)(\mathbf{w}_0) = \binom{n}{m}^{-1} \sum_{1 \leq i_1 < i_2 < \ldots < i_m \leq n} y_{(1)}(\mathbf{w}_0; (y_{i_1}, \mathbf{w}_{i_1}), (y_{i_2}, \mathbf{w}_{i_2}), \ldots, (y_{i_m}, \mathbf{w}_{i_m})), \tag{7}$$

where $\binom{n}{m}$, the binomial coefficient of n chooses m, is the number of all possible subsamples.

4.2 Jackknife bias reduction

We further reduce prediction bias using a generalized Jackknife approach. Biau et al. (2010); Biau and Devroye (2015) derive the theoretical form of the bias for the bagged nearest neighbors and show that the bias's first-order term is proportional to $m^{-2/d}$, where m is the subsample size and d is the dimension of $\mathbf{w}$. Specifically, the asymptotic prediction bias follows the theorem below under mild assumptions.

Theorem 2. *(Biau et al., 2010) Given* $\mathbf{w}_0 \in \text{supp}(\mathbf{w})$*,*

$$Bias(\tau_n(m)(\mathbf{w}_0)) = c\, m^{-2/d} + o(m^{-2/d}),$$

where c is a constant depending on the distribution of $\mathbf{w}$, the focal point $\mathbf{w}_0$, and the dimensionality d, but not on the choice of subsample size m. For details, see Appendix A.3.

Demirkaya et al. (2022) further proposes a generalized jackknife approach to remove the first-order prediction bias. To illustrate, consider two bagged nearest neighbors predictors with different subsampling scales m_1 and m_2 ($m_1 \neq m_2$). By the above theorem, their

associated asymptotic prediction biases have the following forms:

$$\text{Bias}(\tau_n(m_1)(\mathbf{w}_0)) = c\, m_1^{-2/d} + o(m_1^{-2/d}),$$
$$\text{Bias}(\tau_n(m_2)(\mathbf{w}_0)) = c\, m_2^{-2/d} + o(m_2^{-2/d}).$$

With the appropriate choice of weights (θ_1, θ_2), determined by m_1, m_2, and d, the weighted sum of these two predictors, $\theta_1\tau_n(m_1)(\mathbf{w}_0) + \theta_2\tau_n(m_2)(\mathbf{w}_0)$, can be free of first-order prediction bias. An example is that when $m_2 = 2m_1$ and $d = 3$, $-1.70\,\tau_n(m_1)(\mathbf{x}_0) + 2.70\,\tau_n(m_2)(\mathbf{x}_0)$ is free of first-order prediction bias. Demirkaya et al. (2022) shows that this simple procedure can substantially improve finite sample prediction performance.

4.3 Statistical inference

Various aspects of the asymptotic properties of the bagged nearest neighbors estimator have been established by Biau et al. (2010); Demirkaya et al. (2022). For statistical inference, Demirkaya et al. (2022) has shown that the bagged nearest neighbors estimator converges asymptotically to a normal distribution. Formally,

Theorem 3. *(Demirkaya et al., 2022) Given* $\mathbf{w}_0 \in \text{supp}(\mathbf{w})$*, under mild assumptions, and assuming* $m \to \infty$ *and* $m/n \to 0$*, then for some positive* σ_n *with* $\sigma_n^2 = O(\frac{m}{n})$*, as* $n \to \infty$*,*

$$\frac{\tau_n(m)(\mathbf{w}_0) - \mathbb{E}(y|\mathbf{w} = \mathbf{w}_0) - Bias(\tau_n(m)(\mathbf{w}_0))}{\sigma_n} \xrightarrow{\mathscr{D}} N(0, 1).$$

Demirkaya et al. (2022) shows that the jackknife procedure does not change the asymptotic normality property. The bless of proven inference results is an attractive feature for few machine-learning methods, except for the random forests (Wager and Athey, 2018) and the second-step parameters inference after the first-step use of neural networks (Farrell et al., 2021) and other machine models under an orthogonal structure (Chernozhukov et al., 2017).

However, our goal is not demand prediction per se. Instead, our framework uses multiple demand predictions to estimate price elasticities, and ultimately, infer optimal prices. Statistical inference in our case is even more challenging because our objects of interest involve nonlinear transformations and consist of correlated prediction components.

We address this problem by proving that the standard bootstrap procedure derives valid inferences for the bagged nearest neighbors estimator defined in Equation (7). Because the bootstrap commutes with smooth functions (Bickel and Freedman, 1981), the standard bootstrap procedure applies as long as the final estimator is a smooth function of multiple

bagged nearest neighbors predictors, which is precisely our case. By applying the bootstrap to the final estimator, its nonlinearity and the correlation between different components are handled implicitly. The key to the above convenience is the theorem below. The proof is allocated to Appendix A.3. It suggests that the bootstrap can approximate the asymptotic normal distribution of the bagged nearest neighbors estimator. Interested readers may refer to Demirkaya et al. (2022) for other forms of convergence results.

Theorem 4. *Let* $\mathbf{G}_n$ *be the empirical distribution of our sample* $(y_i, \mathbf{w}_i)_{i=1}^n$. *Given* $(y_i, \mathbf{w}_i)_{i=1}^n$, *let* $(y_i^*, \mathbf{w}_i^*)_{i=1}^n$ *be the conditionally independent bootstrap sample with common distribution* $\mathbf{G}_n$. *The bagged nearest neighbors estimator defined in this bootstrap sample is then*

$$\tau_n^*(m)(\mathbf{w}_0) = \binom{n}{m}^{-1} \sum_{1 \le i_1 < i_2 < \ldots < i_m \le n} y_{(1)}(\mathbf{w}_0; (y_{i_1}^*, \mathbf{w}_{i_1}^*), (y_{i_2}^*, \mathbf{w}_{i_2}^*), \ldots, (y_{i_m}^*, \mathbf{w}_{i_m}^*)).$$

Given $\mathbf{w}_0 \in \text{supp}(\mathbf{w})$ *and* σ_n *in Theorem 3, under mild assumptions, and assuming* $m \to \infty$ *and* $m/n \to 0$, *then for almost all sample sequences, as* $n \to \infty$,

$$\frac{\tau_n^*(m)(\mathbf{w}_0) - \mathbb{E}^* \tau_n^*(m)(\mathbf{w}_0)}{\sigma_n} \xrightarrow{\mathscr{D}} N(0,1).$$

4.4 Scalability and implementation

In this section, we discuss practical issues when implementing our estimation strategy. We first present an equivalent L-representation of the bagged nearest neighbors estimator, which provides a closed-form representation of the estimator that substantially eases computational burden. We then discuss code acceleration using just-in-time compilation and parallel computing. We streamline and implement these routines in `Python` for price elasticity estimation. This combination of L-representation and programming techniques ensures our approach's scalability and practical usability. We discuss parameter tuning at the end of the section.

L-representation. Modern firms or researchers often have large datasets. They often estimate price elasticities nonparametrically and would like the estimation procedure to be scalable. As defined in Equation (7), the bagged nearest neighbors involve repeatedly drawing subsamples and averaging between them. The heavy computational burden from subsampling is prohibitive when both n and m are large, limiting its usability in cases where flexible estimation methods are most needed. Fortunately, the bagged nearest neighbors

predictor we use has an equivalent L-statistics form (Steele, 2009), which is

$$\tau_n(m)(\mathbf{w}_0) = \binom{n}{m}^{-1}\left\{\binom{n-1}{m-1}y_{(1)} + \binom{n-2}{m-1}y_{(2)} + \cdots + \binom{m-1}{m-1}y_{(n-m+1)}\right\},$$

where $y_{(1)}$ is the outcome y of the nearest $\mathbf{w}$ to $\mathbf{w}_0$ in the *whole* sample, $y_{(2)}$ is the outcome y of the second nearest $\mathbf{w}$ to $\mathbf{w}_0$ in the *whole* sample, and so on.

The L-representation implies using closed-form weights for computation instead of drawing subsamples. Given subsampling scale m, it takes only three steps to form a bagged nearest neighbors predictor. First, compute a closed-form weight vector. Second, sort the observations in the *whole* sample based on their distances from $\mathbf{w}_0$ and the outcome vector after sorting. Third, compute the inner product of the two vectors from the previous steps. This algebraic trick substantially eases the computational burden and makes bagged nearest neighbors scalable to large data sets.

Just-in-time compilation and parallel computation. Using a just-in-time compiler and parallel computation, the above weighted sum strategy can be further integrated and streamlined for elasticity estimation in `Python`. We manage to utilize `numba`, a high-performance just-in-time (JIT) compiler, to translate `Python` functions to optimized industry-standard native machine code at runtime. For our case, the `numba`-translated `Python` functions are accelerated 100 times compared to native `Python` functions. On the other hand, the use of `numba` also helps avoid the built-in Global Interpreter Lock (GIL) in `Python` so that we can make full use of our multi-core CPUs with parallel computation by multi-threading. For our case, multi-threading works best when bootstrapping for inference. On our 10-core desktop PC, the execution time of a multi-threaded implementation is about 1/10 that of a single-threaded implementation.

Hyperparameter tuning. The bagged nearest neighbors average the 1-nearest neighbors from subsamples of scale m. The parameter m thus becomes a tuning parameter or hyperparameter. A large m means higher weights for observations near the point of interest, focusing the estimation around a local area of the sample. A small m puts flatter weight on a wider sample area, reduces the estimates' variance, but bears the risk of over-smoothing the shape of elasticity structure.

One practical convention for choosing the tuning parameter is the $k-$fold cross-validation. This procedure adds an extra layer to existing machine-learning algorithms and thus mitigates concerns about the choice of tuning parameters. Nevertheless, unlike the task of

predicting an outcome variable, we do not observe the actual price elasticities. One way to mitigate this concern is to impose an extra assumption. That is, assume that good predictions in sales quantities or some other observable proxies can lead to reasonable estimates of price elasticities. In this way, $k-$fold cross-validation can be utilized for the forecasts of sales quantities or other proxies, and we get a choice for the tuning parameter m.[9]

In our Monte Carlo simulations, we explore a range of tuning parameters and assess the robustness of our estimates across them. Our finding is robust under a range of choices for the tuning parameter m. In our empirical study, we conduct 5-fold cross-validation on the forecasts of sales quantities. The resulting loss curves in terms of mean squared prediction errors exhibit a classical U-shape, and we pick the ones with minimum losses as the tuning parameters. Nevertheless, the "best practice" of hyperparameter tuning still needs more exploration and practical experience.

5 Monte Carlo simulations

In this section, we demonstrate the performance of our nonparametric price-elasticity estimator through a broad set of Monte Carlo simulations. These simulations focus on data-generating processes (DGPs) commonly used to model demand in empirical settings. They also mimic the typical amount of data and price variation feasible to researchers. Thus, these simulations allow us to evaluate the finite-sample behavior of our estimator in realistic settings and compare its estimates to the DGPs' ground truth.

5.1 Random-coefficient logit demand

We start with the random-coefficient demand model, a framework often used in empirical demand analysis (Berry et al., 1995). The data-generating process is that each consumer has a logit demand and chooses among a fixed set of alternatives. Preferences are fixed within each consumer and differ across consumers. Similar DGPs are used to characterize micro-level choice data (Rossi et al., 1996). A special case is logit demand (McFadden, 1973; Berry, 1994), in which consumers are homogeneous apart from demand shocks.

[9] Another scenario is that, at least for a subset of the data, the price elasticities are also observable. In this case, $k-$fold cross-validation can be directly conducted on the elasticities. But the new concern for this case is the representativeness of these observed price elasticities.

Setup. Let ι index consumers. Let $j = 1, 2, ..., J$ index a fixed set of products. And let t index markets (t can also be interpreted as periods). Consumer ι buys at most one product out of the choice set and one unit of that product. If she buys product j, she derives utility

$$u_{\iota jt} = \beta_{\iota j} + \alpha_\iota \, p_{jt} + \xi_{jt} + \epsilon_{\iota jt} \ ,$$

where $\beta_{\iota j}$ characterizes product j's match value to consumer ι, α_ι is the consumer's price sensitivity, ξ_{jt} is the product's unobserved characteristics in market t following a uniform distribution between $[-0.5, 0.5)$, and $\epsilon_{\iota jt}$ is a type-I extreme value utility shock. If the consumer chooses the outside good, she gets $u_{\iota 0t} = \epsilon_{\iota 0t}$.

At the individual level, this model gives a logit choice probability, which has a closed-form derivative on price. Specifically, the probability of the consumer choosing j is

$$s_{\iota jt} = \frac{\exp(\beta_{\iota j} + \alpha_\iota \, p_{jt} + \xi_{jt})}{1 + \sum_{k=1}^{J} \exp(\beta_{\iota k} + \alpha_\iota \, p_{kt} + \xi_{kt})} \ .$$

And the derivative of individual choice probabilities on own price and other products' prices are

$$\frac{\partial s_{\iota jt}}{\partial p_{kt}} = \begin{cases} \alpha_\iota s_{\iota jt}(1 - s_{\iota jt}) & \text{if} \quad j = k, \\ -\alpha_\iota s_{\iota jt} s_{\iota kt} & \text{if} \quad j \neq k. \end{cases}$$

Researchers do not see individual-level choice probabilities. Instead, they observe the total sales quantity for each product in each market. Denote the vector of consumer heterogeneity $\Theta_\iota = (\beta_{\iota 1}, ..., \beta_{\iota J}, \alpha_\iota)$. In the model, sales quantities are the sum of choice probabilities across all individuals:

$$s_{jt} = \mathcal{M} \int_\iota s_{\iota jt} dF(\Theta_\iota),$$

Where constant $\mathcal{M}$ is the number of consumers in the market and $F(\Theta_\iota)$ is the consumer preference distribution. The derivative of sales quantity is

$$\frac{\partial s_{jt}}{\partial p_{kt}} = \mathcal{M} \int_\iota \frac{\partial s_{\iota jt}}{\partial p_{kt}} dF(\Theta_\iota).$$

We have not yet specified the distribution of consumer preferences, $F(\Theta_\iota)$. We consider three cases for this distribution. In the first case, consumer preferences are homogeneous. This case reduces to the logit model. In our simulations, we assume $\boldsymbol{\beta}$ equals 0.4 for all products and $\alpha = -3$.

In the second case, consumer preference parameters follow independent normal distri-

butions. This specification is common in empirical applications of the random-coefficient demand model. Specifically, β and α follow independent normal distributions with means of 0 and -3 and standard deviations of 0.8 and 0.5, respectively.

The third specification further allows for the correlation between demand intercepts and random coefficients. Suppose income inc_ι drives consumers' demand for product quality (part of the intercept) and price sensitivities. We have $\beta_{\iota j} = b_{0j} + b_{1j}\text{inc}_\iota$ and $\alpha_\iota = a_0 + a_1\text{inc}_\iota + a_2\nu_\iota$, where ν follows a standard normal distribution. In our simulations, we take inc_ι from a standard normal distribution (think of it as log income). In addition, we assume $\boldsymbol{b}_0 = \mathbf{0}$ and $\boldsymbol{b}_1 = \mathbf{0.8}$ for all products, $a_0 = -3$, and $a_1 = a_2 = 0.5$.

We allow prices to be endogenous to unobserved product characteristics, ξ_{jt}. For all three DGPs, we specify a reduced-form selection type equation:

$$\mathbf{p}_t = 0.5\,\mathbf{1} + 0.5\,\mathbf{z}_t + 0.5\,\xi_t + 0.5\,\mathbf{e}_t.$$

Prices are functions of (own) unobserved characteristics ξ, instruments z, and exogenous shocks e. In addition to ξ, the other two components z and e also follow uniform distributions in $[-0.5, 0.5)$. The inclusion of e reduces the power of price instruments and mimics realistic settings.

Below, we present Monte Carlo simulation results across the three cases. For the logit model, in each market $t = 1, ..., T$, we compute the closed-form choice probability for a representative consumer as the market share. In the baseline, we assume that consumers choose among $J = 4$ products and the outside option. For the two cases that involve random coefficients, we simulate 500 consumers whose choice probabilities are integrated to compute the market share. For each DGP, we simulate 100 samples, each with a sample size of $T = 20,000$. In the estimation, we hold the tuning parameter $m = 7$. We later vary the sample size, tuning parameter m, as well as the number of products, to evaluate the estimator's performance under different scenarios.

For each sample, we estimate point elasticities (with bootstrap standard errors) at the mid-point price ($p = \$0.5$ in our simulation settings). We also estimate the profile of price elasticities for a broad set of price points in a smaller set of samples.

Simulation results. Table 1 demonstrates the elasticities of product 1's and product 2's quantities to product 1's price, representing one own-elasticity and one cross-elasticity estimand. For each elasticity, we present the ground truth, the mean of our point estimates, the mean of our bootstrapped standard error, and the standard deviation of our point

Table 1: Monte Carlo simulation results: baseline

model	elasticity	true value	mean esti.	mean std. err.	std. of esti.
logit	own	-0.753	-0.774	0.067	0.068
	cross	0.247	0.240	0.065	0.075
independent RC logit	own	-1.273	-1.305	0.089	0.082
	cross	0.200	0.195	0.079	0.075
correlated RC logit	own	-1.333	-1.363	0.094	0.100
	cross	0.204	0.205	0.082	0.085

Note: Column 1 presents the true elasticity at the mid-point price. Column 2 is the mean of the nonparametric point estimates. Column 3 presents the mean of the bootstrapped standard error. Column 4 is the standard deviation of the point estimates across Monte Carlo experiments. Number of products $J = 4$. Sample size $T = 20,000$. Tuning parameter $m = 7$.

estimates, all taken across Monte Carlo experiments.

We find that the point estimates fall tightly around the true value. Across the three DGPs, the mean of own- and cross-elasticity estimates are within 0.03 and 0.01 of the true value, respectively. These gaps are statistically small, given the standard errors.

We also find that the estimated standard errors (column 3) align with the standard deviation of the point estimates (column 4). The former is the estimated uncertainty for each point estimate, which is then averaged across estimates. The latter is a direct observation of the variability of point estimates across Monte Carlo samples. We find that the two metrics align, suggesting that the bootstrapped standard errors correctly reflect the degree of uncertainty as expected by Theorem 4.

Next, we vary the sample size, tuning parameters, and the number of products to assess how the estimator performs in a broader range of cases. We first vary the sample size T by factors of four. Table 2 demonstrates that our estimator is still usable with a small sample of 5,000 observations. However, standard errors, in this case, are large relative to the elasticity values, especially for cross elasticities. With a growing sample size, the mean of the estimate remains close to the true value, and standard errors shrink by a factor that roughly equals the square root of the sample size. This finding aligns with Theorems 3 and 4, which state that the convergence rate for bagged nearest neighbors is $\sqrt{T/m}$. The point estimates are precise with large samples, such as when $T = 320,000$.

Point estimation (including computing bootstrap standard errors from 100 bootstrapped samples) is reasonably fast with only the computing power of CPUs, even with large sample

Table 2: Simulation results across sample sizes

sample size	elasticity	true value	mean esti.	mean std. err.	std. of esti.
$T = 5,000$	own	-1.333	-1.386	0.186	0.179
	cross	0.204	0.210	0.165	0.151
$T = 20,000$	own	-1.333	-1.363	0.094	0.100
	cross	0.204	0.205	0.082	0.085
$T = 80,000$	own	-1.329	-1.371	0.046	0.048
	cross	0.204	0.211	0.040	0.039
$T = 320,000$	own	-1.338	-1.374	0.023	0.025
	cross	0.205	0.208	0.020	0.015

Note: See notes in Table 1. Number of products $J = 4$. Tuning parameter $m = 7$. Varying sample sizes T.

size. On our personal computer,[10] we find that at $T = 20,000$, one point estimate takes 0.17 seconds. This time increases to 0.84 seconds for $T = 80,000$ and 4.88 seconds for $T = 320,000$. This efficiency makes the estimator scalable with very large samples.

We have so far fixed the tuning parameters at $m = 7$. Recall that the bagged nearest neighbor estimator works *as if* we draw subsamples of size m, take the nearest neighbor of each subsample, and average across them. A small m such as $m = 7$ implies that we use information from a large neighborhood when estimating each point elasticity. Using a broader part of the sample leverages more information and produces more stable estimates. However, although we use the generalized jackknife to subtract the first-order bias (Theorem A.1), it might still be prone to higher-order biases (e.g., over-smoothing demand) and more sensitive than cases where one only uses local information.

Table 3 investigates how the estimator performs with varying m. Holding the sample size T and dimensionality J fixed, we find that a very small $m = 3$ produces more precise but significantly biased estimates. On the other hand, using larger $m = 20$ leads to noisier estimates. Our main simulation exercise uses $m = 7$ because no visible bias remains, and one would only get noisier estimates for $m > 7$. Under different DGPs, one could use cross-validation to determine a sensible m, as we do in our empirical exercise. Or one could use a more conservative, larger m.

Like all nonparametric estimators, our estimator suffers from the curse of dimensionality. This occurs because of the lack of parametric functional forms to project the model across dimensions. As a result, as the size of consumers' choice set, J, grows, one potentially needs

[10]The specification of our working computer is Intel Core i9-10900K @ 3.70GHz, with 128GB RAM, and the operating systems are Ubuntu Server 20.04 LTS and Windows 10 Enterprise.

Table 3: Simulation results: tuning parameters

tuning parameter	elasticity	true value	mean esti.	mean std. err.	std. of esti.
$m = 3$	own	-1.331	-1.778	0.062	0.070
	cross	0.204	0.275	0.051	0.057
$m = 5$	own	-1.333	-1.451	0.071	0.079
	cross	0.204	0.209	0.064	0.067
$m = 7$	own	-1.333	-1.363	0.094	0.100
	cross	0.204	0.205	0.082	0.085
$m = 10$	own	-1.330	-1.357	0.115	0.115
	cross	0.204	0.187	0.101	0.116
$m = 20$	own	-1.333	-1.311	0.166	0.181
	cross	0.204	0.155	0.152	0.144

Note: See notes in Table 1. Number of products $J = 4$. Sample size $T = 20,000$. Varying tuning parameters m.

a much larger sample to estimate all point elasticities reliably. We fixed the dimensionality in our previous simulations at $J = 4$. This choice-set size can characterize a concentrated market, such as demand within a category for most consumer packaged goods (see our empirical example). Nevertheless, this choice-set size is small for less-concentrated markets. We now explore how the model performs with larger J's using a larger sample of $T = 320,000$ and focusing on more local variations, by setting $m = 30$.[11]

We can still estimate the own and cross elasticities reasonably well within $J \leq 12$. Standard errors are larger for larger J's (holding fixed T and m). This scale of choice set is comparable to many parametric exercises using random-coefficient logit models. A bonus point is that, even at this sample size, the computation time for one point estimate (including bootstrap standard errors) is still around five seconds. Our model is still usable at larger J's, provided that the researcher has access to a sizable data set.

When J goes to 16, we observe significant biases on the own-elasticity term. We therefore recommend practical caution when using our estimator for large J's without increasing the sample size T and the tuning parameter m.

Finally, thus far we have only focused on point elasticities at the mid-point price. The top panels of Figure 2 provide *one realization* of point-elasticity estimates across a wide range of prices when the DGP comes from a random-coefficient logit model. This result

[11] Our intuition behind setting a higher m is to use more local observations. This is because a higher dimensionality increases model complexity (the number of parameters at every point, here own and cross elasticities), making it more difficult to use non-local observations without introducing biases.

Table 4: Simulation results: number of products

number of products	elasticity	true value	mean esti.	mean std. err.	std. of esti.
$J = 4$	own	-1.334	-1.330	0.053	0.056
	cross	0.204	0.207	0.049	0.045
$J = 8$	own	-1.394	-1.405	0.083	0.082
	cross	0.128	0.132	0.070	0.087
$J = 12$	own	-1.428	-1.592	0.118	0.111
	cross	0.095	0.096	0.094	0.083
$J = 16$	own	-1.441	-1.896	0.152	0.150
	cross	0.075	0.081	0.115	0.104

Note: See notes in Table 1. Sample size $T = 320,000$. Tuning parameter $m = 30$. Varying number of products J.

shows that our estimator can recover the *profile* of elasticities as a flexible function of prices (and other observables).

5.2 Other data-generating processes

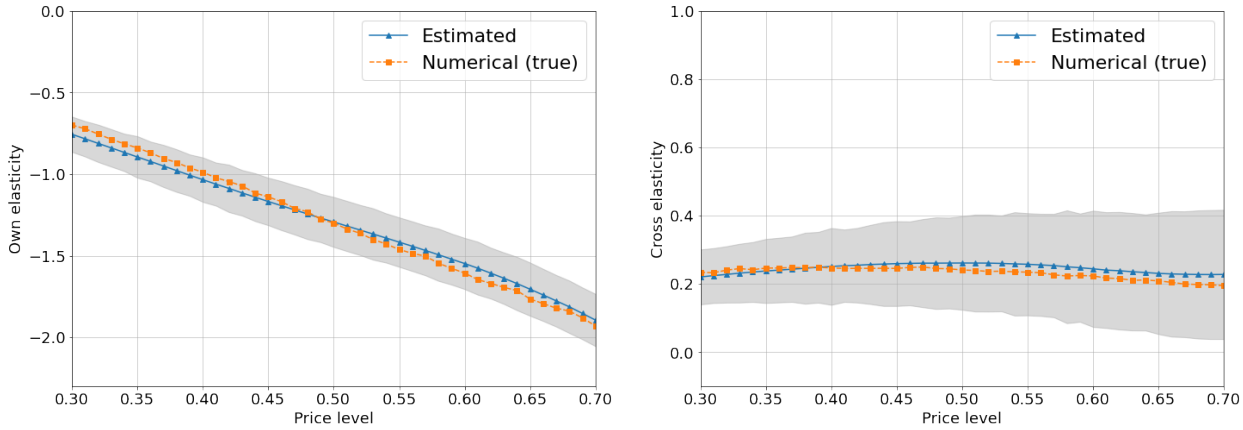
We have focused on data-generating processes that fall into the mixed logit class but have not leveraged the mixed logit structure for estimation. For example, our estimation model does not impose that products are substitutes or that consumers choose one alternative and one unit of the alternative. It is natural to imagine that our estimator works on smooth demand from other DGPs.

Here we consider two other data-generating processes. In the first case, some products are substitutes whereas others are complements. We model complements in a way similar to Gentzkow (2007), where consumers choose individual products or product bundles, and the additional utility (or disutility) from buying a bundle can generate complementarity (or additional substitutability). In the second case, consumers can choose many products and multiple quantities of each product. We model this case based on the multiple discrete-continuous demand framework from Kim et al. (2002). But different from Kim et al., we do not assume additively-separable payoff functions across varieties, and the nonseparability introduces additional substitution between varieties. Appendix B outlines these two data-generating processes in more detail.

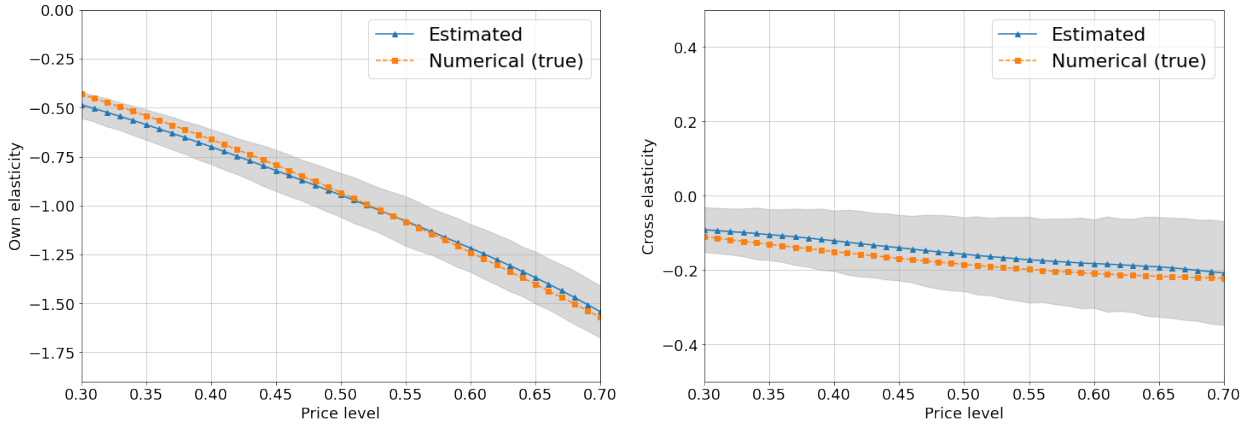
In both cases, we demonstrate that our estimator can reliably recover the own and cross elasticities. Table 5 indicates that, with the same sample size and tuning parameters

Figure 2: Own and cross elasticity profiles

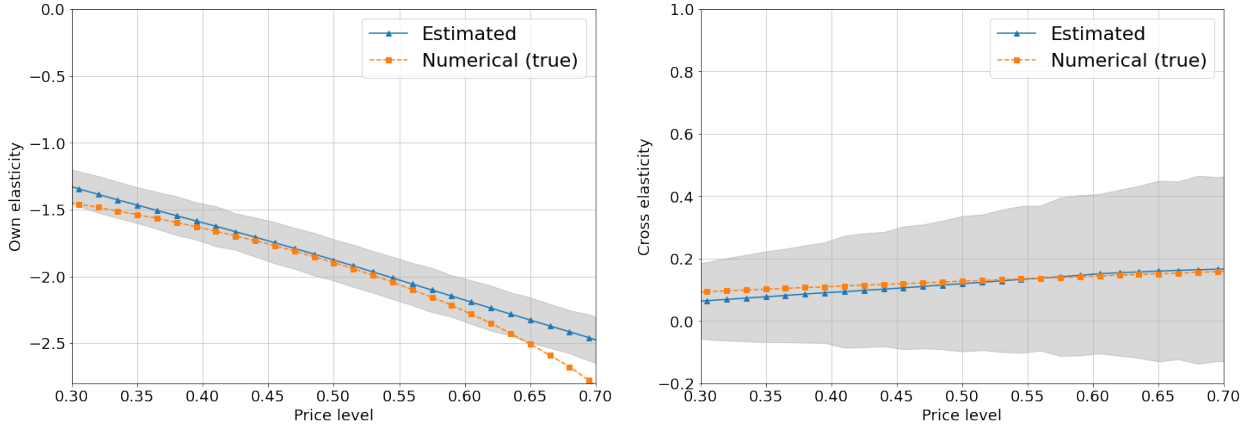
(a) Correlated BLP model



(b) Complementarity



(c) Variety and Quantity



Note: The top panels are plots of the mean own- and cross-price elasticities along own price levels from 10 Monte Carlo simulations of a random-coefficient logit model. The middle panels are a model of substitutes and complements. The bottom panels are of a model of variety and quantities. Sample size $T = 20,000$, number of products $J = 3$. Tuning parameter fixed at $m = 7$. For further model details, see Appendix B.

Table 5: Simulation results: number of products

model	elasticity	true value	mean esti.	mean std. err.	std. of esti.
complements	own	-0.934	-0.931	0.057	0.057
	cross	-0.185	-0.173	0.052	0.054
variety and quantity	own	-1.989	-1.964	0.083	0.092
	cross	0.163	0.163	0.110	0.129

Note: This table shows Monte Carlo simulation results when the data-generating processes contain complements (Gentzkow, 2007) or when the DGP allows consumers to choose from multiple products and buy multiple quantities (Kim et al., 2002). Sample size $T = 20,000$. Tuning parameter $m = 7$. Number of products $J = 3$.

as our benchmark, both own and cross elasticities under both DGPs are unbiased and precisely estimated. Note that, for reasons unrelated to our estimator, we limit the choice-set size to $J = 3$. This choice of J is because solving the optimal consumer choices in the variety-quantity model requires enumerating all possible bundles, introducing an enormous computation burden when J is large (we cannot use Kim et al.'s approach, which relies on additive separability).

In addition to Table 5, the bottom panel of Figure 2 demonstrates that we can recover the elasticities within the data support, for the case where two of the products are complements or when consumers choose a variety of products at different quantities.

6 Empirical Application

This section applies our method to estimate the price elasticities of yogurt. The yogurt category is widely studied in marketing and economics, and is often used as a touchstone for testing new methods. The literature has emphasized various aspects of consumer behavior that depart from canonical mixed logit models and has thus adopted different methods to accommodate these departures.[12] Our method estimates own- and cross-price elasticities among popular products without imposing a priori structural assumptions on consumer behavior. As such, this exercise not only serves to demonstrate how our method can be

[12]The canonical (mixed) logit model assumes that a consumer purchases at most one unit of one product among all available products. Extensive research has demonstrated departures from this assumption and has proposed solutions. For example, Kim et al. (2002) studies consumers' choices of various products in multiple quantities. Pavlidis and Ellickson (2018) examine consumer switching costs across products and brands and discuss its implications for dynamic pricing. Huang and Bronnenberg (2018) study the costly consideration-set formation in a demand system, where consumers can choose various products in multiple quantities.

applied in practice, but also presents elasticity estimates that inform the industry's markups and extent of competition.

6.1 Sample construction

We use the IRI academic dataset (Bronnenberg et al., 2008) for this exercise. The IRI data contain weekly sales quantity and prices across chains in several US states, from 2001 to 2007. Appendix Tables A1 and A2 present summary statistics.

Yoplait is the most popular brand during our sample period. Dannon is the second most popular brand, with two-thirds of the sales of Yoplait. Private labels are prevalent in this sample period but do not dominate the market. The total sales quantity of private labels, as a whole, ranks below Dannon and Yoplait and above all other brands. We group private labels together and consider them the third most popular "brand."

Like many consumer packaged goods, each yogurt product line is typically offered in multiple sizes. We divide products by size for the top two brands, Yoplait and Dannon, then report total sales quantity and average unit price by size. We find that both Yoplait and Dannon offer two vertically-different product lines. One is at around \$1.7 per pound, and the other is premium, at around \$2.4 per pound. We group products with similar sizes and unit prices. We group Dannon's 0.375 pounds (6 oz) and 0.5 pounds (8 oz) sizes together as *Dannon small*, and we group Dannon's 1.5 pounds (24 oz) and 2 pounds (32 oz) sizes together as *Dannon large*. For Yoplait, Yoplait 0.375 pounds (6 oz) is defined to be *Yoplait small* and Yoplait 1.5 pounds (24 oz) is *Yoplait large*. Since Yoplait offers more products at the premium line, we define Yoplait 0.25 pounds (4 oz), 0.6875 pounds (11 oz), and 1.125 pounds (18 oz) as *Yoplait premium*. For *Private label* yogurts, we do not further distinguish between sizes and use their distinct sum of sales in pounds to divide the total revenue and determine their prices.

With the above definitions, we eventually arrive at a sample of $156,580$ observations at the store-week level. For each observation, we have the prices and quantities for Dannon small, Dannon large, Yoplait small, Yoplait large, Yoplait premium, and Private label. Our analysis focuses on the sales quantity of Dannon small, Dannon large, Yoplait small, and Yoplait large.

We control demand shocks at the store, chain, and time levels. The nonparametric demand model in Equation (1) makes it challenging to control for fixed effects without further functional-form assumptions. Instead, we measure demand changes using observed variables and control for these factors. Specifically, we use the IRI dataset's all-commodity

volume (ACV) to control for size differences across stores. In addition, to control for chain-level differences, we construct Dannon's sales quantity and Private label's sales quantity as a fraction of the total yogurt sales quantity in the chain. Our interpretation is that different chains allocate different shelf spaces between the two top brands and their private labels, and the variables capture persistent differences in shelf space. We also include the week of the year and the year (both rescaled to between 0 and 1) as control variables, to capture seasonality and the time trend. Although our analysis uses proxy variables instead of fixed effects, we do not make functional-form restrictions, thus permitting these controls to enter the model flexibly.

We instrument for the potentially endogenous prices using Hausman instruments (Hausman et al., 1994; Nevo, 2001). If one makes the identifying assumption that products in different geographic markets have independent demand shocks, common price movements for the same product across the market reflect common cost shocks. Given this assumption, we use a product's average price in other markets to instrument its price in the focal market.

The identifying assumption fails if demand shocks are correlated between markets. Suppose demand shocks are positively correlated; one would expect that our elasticity estimates are upward-biased (that is, the true elasticities are more negative than our estimated elasticities). Our empirical exercise mainly serves to demonstrate the nonparametric demand estimation method. The reader should carefully select suitable price instrument(s) to apply our method in a given context.

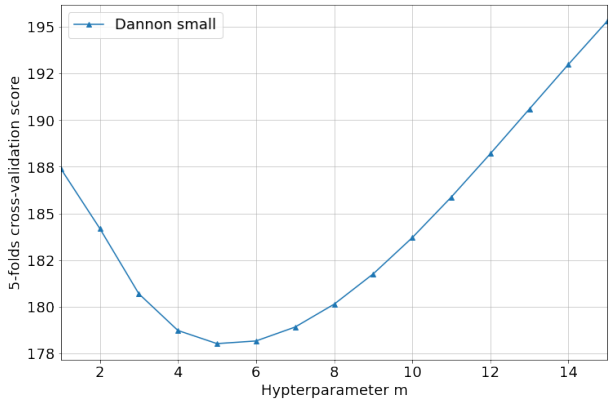
6.2 Hyperparameter tuning

To tune the hyperparameter m, We conduct five-fold cross-validation to predict sales quantity for each m. We use the sum of mean squared errors from the five validation folds as the cross-validation loss function. The results are reported in Figure 3. Except for one irregular point ($m = 1$ for Yoplait small), we have a U-shaped curve for all four products. The average loss across all samples first decreases with m and later increases with m.

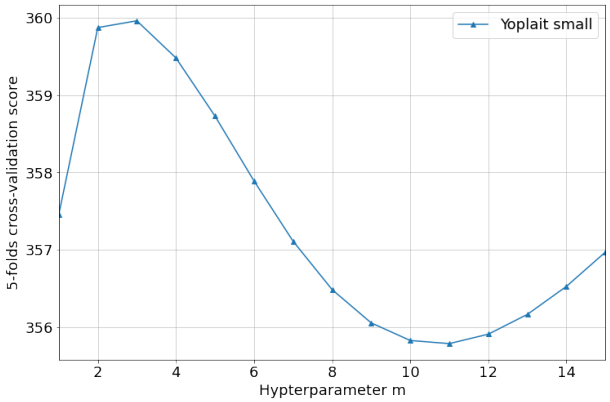
Following conventional practices, we choose the m that delivers a minor loss for each product. These are $m = 5$, 11, 5, and 2 for Dannon small, Yoplait small, Dannon large, and Yoplait large, respectively. We apply these tuning-parameter choices to estimate price elasticities. An implicit assumption is that the value of m that predicts sales quantity well also gives reasonable price elasticity estimates. Our cross-validation procedure has produced the same tuning parameters for Dannon but has chosen different tuning parameters for the two Yoplait products. In particular, m is surprisingly tiny for Yoplait large.

Figure 3: Five-fold Cross-Validation

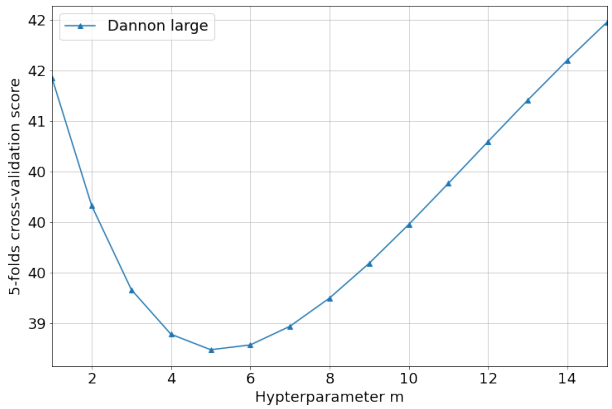
(a) Dannon small



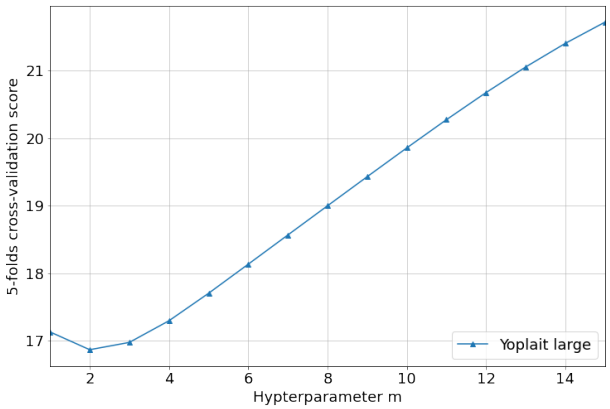
(b) Yoplait small



(c) Dannon large



(d) Yoplait large



Note: This figure reports the five-fold cross-validation score when choosing different hyperparameters m for Dannon small, Dannon large, Yoplait small, and Yoplait large. The chosen loss function is the sum of the mean squared errors in predicting sales quantity from each validation fold.

6.3 Empirical findings

Table 6 presents the average own- and cross-elasticities at the median price, and Figure 4 demonstrates how own-price elasticities vary with the price for each product. The bootstrap 95% confidence intervals for all these point estimations are reported in dashed lines, and the areas between are shaded. We highlight several findings below.

First, the estimated own-price elasticities are all negative and large in magnitude. In contrast, the estimated cross-price elasticities are primarily positive and significant in size, which suggests intense competition in the yogurt category.

Second, most products' estimated own-price elasticities' absolute values are non-decreasing with their own prices. The downward sloping pattern of own-price elasticity is sometimes called Marshall's second law of demand (Nocke and Schutz, 2018), which is commonly assumed in the theoretical industrial organization literature. We provide direct evidence that supports this assumption.

Third, for most products, the highest cross-price elasticities are with the opposite brand's product at the same size. This finding would be consistent with a model where different consumers (e.g., with varying family sizes) prefer different sizes, and each consumer chooses between brands within the preferred size. In structural demand estimation, substitution across brands or sizes is an empirical question. Still, it often requires ex ante modeling assumptions, either in the random-coefficient structure or the nesting structure. Our method can be a first step to guiding empirical modeling of micro-founded demand models.

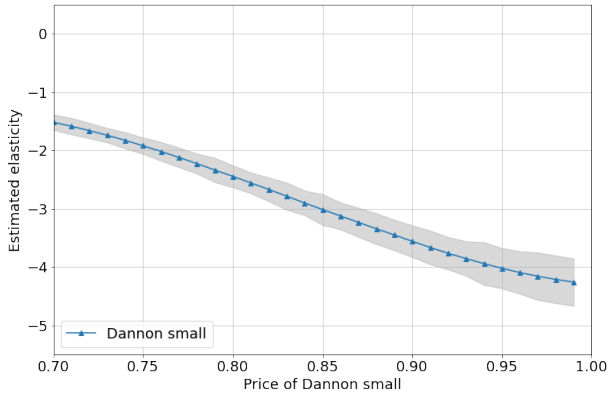
An exception to the above findings is Yoplait large's own- and cross-price elasticities. As noted before, the tuning parameter value we arrive at is $m = 2$, which looks different from the other three. For this product, we also observe wider confidence intervals and an unexpected substitution pattern. Nevertheless, we stick to the same empirical procedure for all four brand-size choices. One possible explanation is that Yoplait has a separate premium product line, which offers more package sizes and does not have much price variation. We provide more details in the tables in Appendix C.

Optimal pricing. We now revisit the managerial decision problem proposed in section 3. When one firm knows marginal costs and would like to compute the optimal static prices, it can solve for them using the standard first-order conditions in Equation (2). Estimating the point own- and cross-elasticities allows the firm to solve (2) for optimal pricing without structural assumptions about consumer demand.

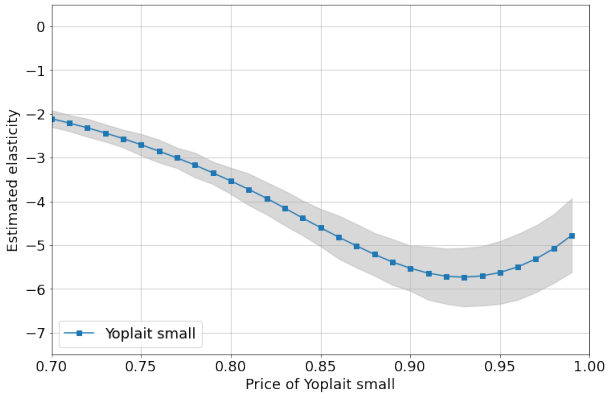
We now demonstrate this point by solving for Equation (2) for Dannon's and Yoplait's

Figure 4: The estimated own-price elasticities of yogurt

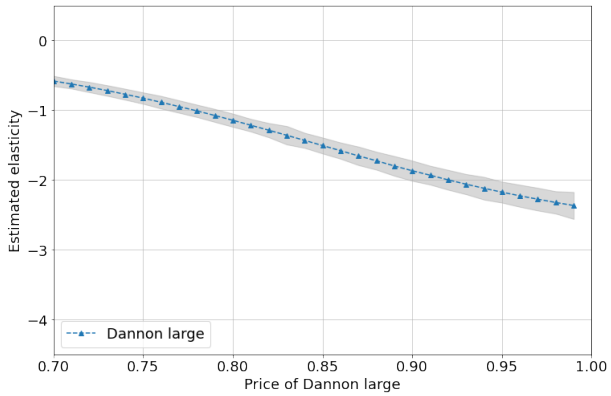
(a) Dannon small



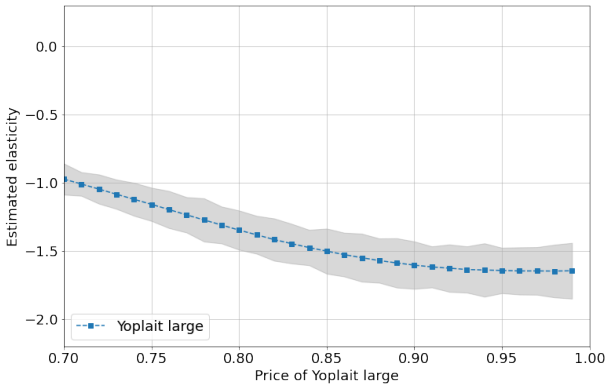
(b) Yoplait small



(c) Dannon large



(d) Yoplait large



Note: This figure provides the own-price elasticity estimates of Dannon small, Dannon large, Yoplait small, and Yoplait large. These elasticities are evaluated at 30 price levels from 0.7 to 1 dollar per 8 oz. For cross-price elasticities, see Table 6 and Appendix Figure A1.

Table 6: Price elasticities at representative points

	Danon small	Yoplait small	Dannon large	Yoplait large
(A) Dannon small ($ 0.85)	-3.044	1.728	0.218	0.726
(B) Yoplait small ($ 0.85)	4.157	-4.722	0.896	0.827
(C) Dannon large ($ 0.85)	0.137	0.450	-1.517	0.900
(D) Yoplait large ($ 0.85)	1.505	-0.556	0.289	-1.528

Note: This table presents the price elasticity estimates when the A (B, C, D) product is priced at $0.85 per 8 oz and all the other products at their median price levels.

Table 7: Pricing

	Danon small	Bootstrap S.E.	Yoplait small	Bootstrap S.E.
Competitive case	0.606	0.013	0.706	0.016
Monopolistic case	0.428	0.090	0.802	0.045

Note: Bootstrap standard errors are obtained from 10 resamplings. The marginal costs are set to be \$0.30 for Dannon small and \$0.33 for Yoplait small.

small-sized products.[13] We consider two cases. The first case focuses on the current market structure, where Dannon and Yoplait optimize their profits from the small-sized yogurt and each competes with the other firm. The equilibrium price vector is a Nash equilibrium where each price is optimal given the other firm's optimal price. The second case is that a monopolist owns Dannon's and Yoplait's small-size yogurt and maximizes joint profits. The reader could view the second case as a merger simulation. In both cases, we assume costs are \$0.30 for Dannon small and \$0.33 for Yoplait small.[14]

In both cases, we reach equilibrium prices by repeatedly iterating on the set of first-order conditions. At an arbitrary starting price vector, we estimate the point own- and cross-price elasticities once and use these elasticities to solve for the left-hand side of Equation (2). This step gives us a vector of prices we should visit next. At this new price vector, we estimate the elasticities again and arrive at a new vector of prices. We repeat these steps until the price vector is stable. In this sense, we estimate demand "locally," only at the points we visit when searching for the optimal prices.

Table 7 presents the implied optimal prices and the bootstrap standard errors. The optimal prices are intuitive: Yoplait faces a less elastic demand and has a higher equilibrium markup (\$0.38) than Dannon (\$0.31). Yoplait's higher markups are consistent with the observation that, in the data, Yoplait's products have higher average prices. In addition, if the two firms were to merge into a monopolist firm, the monopolist would set a lower price for Dannon and a higher price for Yoplait, effectively distancing its vertical product line and serving more customers.

We have demonstrated that our method can recover reasonable yet rich price elasticity patterns. One can estimate the price-elasticity profile "globally" to test a theory or to

[13] We focus on the small package size because, as we pointed out, Yoplait large's small tuning parameters lead to estimates that are difficult to interpret.

[14] We use a higher marginal cost for Yoplait small because it has higher observed prices. The reader can use different marginal costs in this simulation.

inform modeling choices. We have also demonstrated that a firm can use our estimator "locally" to compute optimal prices. In this case, one evaluates point-elasticities only as needed, without having to estimate the entire demand function. A similar use-case is to evaluate a counterfactual ownership change, under the assumption that the demand function is stable and the supply-side counterfactual only changes how the firm internalizes cross-price elasticities.

7 Conclusion

This paper proposes a scalable approach to estimating price elasticities nonparametrically. We cast the price-elasticity estimation exercise in a nonparametric control function framework, apply the bagged nearest neighbors estimator to this framework, and demonstrate the estimator's theoretical, empirical, and computational performance. We showcase how this estimator can be applied to learn the shape of a demand function, find optimal prices, and evaluate a counterfactual ownership change. We believe that our flexible price elasticity estimator can be instrumental in a wide range of marketing and economic applications.

8 Competing interests

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

References

Abaluck, J., G. Compiani, and F. Zhang (2022). A method to estimate discrete choice models that is robust to consumer search.

Athey, S., J. Tibshirani, and S. Wager (2019). Generalized random forests. *The Annals of Statistics 47*(2), 1148–1178.

Berry, S., J. Levinsohn, and A. Pakes (1995). Automobile prices in market equilibrium. *Econometrica 63*(4), 841–890.

Berry, S. T. (1994). Estimating Discrete-Choice models of product differentiation. *The RAND Journal of Economics 25*(2), 242.

Berry, S. T. and P. A. Haile (2014). Identification in differentiated products markets using market level data. *Econometrica 82*(5), 1749–1797.

Bhat, C. R. (2008). The multiple discrete-continuous extreme value (mdcev) model: role of utility function parameters, identification considerations, and model extensions. *Transportation Research Part B: Methodological 42*(3), 274–303.

Biau, G., F. Cérou, and A. Guyader (2010). On the rate of convergence of the bagged nearest neighbor estimate. *Journal of Machine Learning Research 11*(3), 687—-712.

Biau, G. and L. Devroye (2015). *Lectures on the Nearest Neighbor Method*. Springer.

Bickel, P. J. and D. A. Freedman (1981). Some asymptotic theory for the bootstrap. *The Annals of Statistics 9*(6), 1196–1217.

Bronnenberg, B. J., M. W. Kruger, and C. F. Mela (2008). Database Paper-The IRI marketing data set. *Marketing Science 27*(4), 745–748.

Chan, T. Y. (2006). Estimating a continuous hedonic-choice model with an application to demand for soft drinks. *The Rand Journal of Economics 37*(2), 466–482.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, and W. Newey (2017). Double/Debiased/Neyman machine learning of treatment effects. *American Economic Review 107*(5), 261–65.

Compiani, G. (2022). Market counterfactuals and the specification of multiproduct demand: A nonparametric approach. *Quantitative Economics 13*(2), 545–591.

Compiani, G. and A. N. Smith (2021). Boundaries of Differentiated Product Markets and Retailer Pricing. *SSRN Electronic Journal*.

Cong, Z., J. Liu, and P. Manchanda (2021). The Role of "Live" in Livestreaming Markets: Evidence Using Orthogonal Random Forest. *arXiv*.

De Los Santos, B., A. Hortaçsu, and M. R. Wildenbeest (2012). Testing models of consumer search using data on web browsing and purchasing behavior. *American Economic Review 102*(6), 2955–80.

Demirkaya, E., Y. Fan, L. Gao, J. Lv, P. Vossler, and J. Wang (2022). Optimal nonparametric inference with two-scale distributional nearest neighbors. *Journal of the American Statistical Association*.

Dubé, J. (2004). Multiple discreteness and product differentiation: Demand for carbonated soft drinks. *Marketing Science 23*(1), 66–81.

Dubé, J.-P. (2019). Microeconometric models of consumer demand. In *Handbook of the Economics of Marketing*, Volume 1, pp. 1–68. Elsevier.

Dubé, J.-P. and S. Misra (2021). Personalized pricing and consumer welfare. *SSRN Electronic Journal*.

Erdem, T., S. Imai, and M. P. Keane (2003). Brand and quantity choice dynamics under price uncertainty. *Quantitative Marketing and economics 1*(1), 5–64.

Farrell, J. and C. Shapiro (1990). Horizontal mergers: an equilibrium analysis. *The American Economic Review*, 107–126.

Farrell, M. H., T. Liang, and S. Misra (2021). Deep Neural Networks for Estimation and Inference. *Econometrica 89*(1), 181–213.

Gabel, S. and A. Timoshenko (2022). Product Choice with Large Assortments: A Scalable Deep-Learning Model. *Management Science 68*(3), 1808–1827.

Gentzkow, M. (2007). Valuing new goods in a model with complementarity: Online newspapers. *American Economic Review 97*(3), 713–744.

Goeree, M. S. (2008). Limited information and advertising in the U.S. personal computer industry. *Econometrica 76*(5), 1017–1074.

Hausman, J., G. Leonard, and D. J. Zona (1994). Competitive analysis with differenciated products. *Annales d'Économie et de Statistique* (34), 159–180.

Hendel, I. (1999a). Estimating multiple-discrete choice models: An application to computerization returns. *The Review of Economic Studies 66*(2), 423–446.

Hendel, I. (1999b). Estimating multiple-discrete choice models: An application to computerization returns. *The Review of Economic Studies 66*(2), 423–446.

Hendel, I. and A. Nevo (2006). Measuring the implications of sales and consumer inventory behavior. *Econometrica 74*(6), 1637–1673.

Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics 19*(3), 293–325.

Houde, J. (2012). Spatial differentiation and vertical mergers in retail markets for gasoline. *American Economic Review 102*(5), 2147–82.

Huang, Y. and B. J. Bronnenberg (2018). Pennies for your thoughts: Costly product consideration and purchase quantity thresholds. *Marketing Science 37*(6), 1009–1028.

Imbens, G. W. (2020). Potential Outcome and Directed Acyclic Graph Approaches to Causality: Relevance for Empirical Practice in Economics. *Journal of Economic Literature 58*(4), 1129–1179.

Joo, J. (2022). Rational Inattention as an Empirical Framework for Discrete Choice and Consumer-Welfare Evaluation. *Journal of Marketing Research*.

Kim, J., G. M. Allenby, and P. E. Rossi (2002). Modeling consumer demand for variety. *Marketing Science 21*(3), 229–250.

Magnolfi, L., J. McClure, and A. Sorensen (2022). Embeddings and Distance-based Demand for Differentiated Products. *Proceedings of the 23rd ACM Conference on Economics and Computation*, 607–607.

McFadden, D. (1973). *Conditional logit analysis of qualitative choice behavior*, pp. 105–142. Academic Press.

Mehta, N., X. J. Chen, and O. Narasimhan (2010). Examining demand elasticities in hanemann's framework: A theoretical and empirical analysis. *Marketing Science 29*(3), 422–437.

Misra, K., E. M. Schwartz, and J. Abernethy (2019). Dynamic Online Pricing with Incomplete Information Using Multiarmed Bandit Experiments. *Marketing Science 38*(2), 226–252.

Nevo, A. (2001). Measuring market power in the Ready-to-Eat cereal industry. *Econometrica 69*(2), 307–342.

Newey, W. K., J. L. Powell, and F. Vella (1999). Nonparametric estimation of triangular simultaneous equations models. *Econometrica 67*(3), 565–603.

Nocke, V. and N. Schutz (2018). Multiproduct-Firm Oligopoly: An Aggregative Games Approach. *Econometrica 86*(2), 523–557.

Pavlidis, P. and P. B. Ellickson (2018). Implications of parent brand inertia for multiproduct pricing. *Quantitative Marketing and Economics 15*(4), 369–407.

Pearl, J. (2009). Causality: Models, reasoning and inference. *Cambridge University Press*.

Rossi, P. E., R. E. McCulloch, and G. M. Allenby (1996). The value of purchase history data in target marketing. *Marketing Science 15*(4), 321–340.

Seiler, S. (2013). The impact of search costs on consumer behavior: A dynamic approach. *Quantitative Marketing and Economics 11*(2), 155–203.

Smith, A. N., P. E. Rossi, and G. M. Allenby (2019a). Inference for Product Competition and Separable Demand. *Marketing Science 38*(4), 690–710.

Smith, A. N., P. E. Rossi, and G. M. Allenby (2019b). Inference for product competition and separable demand. *Marketing Science 38*(4), 690–710.

Steele, B. M. (2009). Exact bootstrap k-nearest neighbor learners. *Machine Learning 74*(3), 235–255.

Wager, S. and S. Athey (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association 113*(523), 1228–1242.

A Technical details and proof of theorems

A.1 Derivation of Equation 5

Thanks to the regularity conditions in Assumption 3 and 4, we take partial derivatives on both sides of Equation 4 with respect to $\mathbf{p}_t$ and $\mathbf{z}_t$. By the chain rule of calculus, we have

$$\underbrace{\partial_{\mathbf{p}_t} h_j(\mathbf{p}_t, \mathbf{x}_t, \tilde{\mathbf{z}}_t)}_{J\times 1} = \underbrace{\partial_{\mathbf{p}_t} f_j(\mathbf{p}_t, \mathbf{x}_t)}_{J\times 1} + \underbrace{\partial_{\mathbf{u}_t} \lambda(\mathbf{u}_t)}_{J\times 1},$$

$$\underbrace{\partial_{\tilde{\mathbf{z}}_t} h_j(\mathbf{p}_t, \mathbf{x}_t, \tilde{\mathbf{z}}_t)}_{d_{\tilde{z}}\times 1} = -\underbrace{\partial_{\mathbf{z}_t} \mathbf{g}(\mathbf{z}_t)}_{d_{\tilde{z}}\times J} \underbrace{\partial_{\mathbf{u}_t} \lambda(\mathbf{u}_t)}_{J\times 1}.$$

Here $\partial_{\mathbf{p}_t} h_j(\mathbf{p}_t, \mathbf{x}_t, \tilde{\mathbf{z}}_t)$ is the Jacobian matrix of conditional demand function h_j with respect to price vector $\mathbf{p}_t$, evaluated at point $(\mathbf{p}_t, \mathbf{x}_t, \tilde{\mathbf{z}}_t)$. The other terms are defined in the same fashion. We explicitly list the dimensions of these Jacobian matrices underneath to avoid potential confusion on various definitions of the Jacobian matrix.

We can rearrange the above two equations and get

$$\underbrace{\partial_{\tilde{\mathbf{z}}_t} \mathbf{g}(\mathbf{z}_t)}_{d_{\tilde{z}}\times J} \underbrace{\partial_{\mathbf{p}_t} f_j(\mathbf{p}_t, \mathbf{x}_t)}_{J\times 1} = \underbrace{\partial_{\tilde{\mathbf{z}}_t} h_j(\mathbf{p}_t, \mathbf{x}_t, \tilde{\mathbf{z}}_t)}_{d_{\tilde{z}}\times 1} + \underbrace{\partial_{\tilde{\mathbf{z}}_t} \mathbf{g}(\mathbf{z}_t)}_{d_{\tilde{z}}\times J} \underbrace{\partial_{\mathbf{p}_t} h_j(\mathbf{p}_t, \mathbf{x}_t, \tilde{\mathbf{z}}_t)}_{J\times 1},$$

which gives us a system of d_z linear equations in J unknowns, whose solution is to be discussed in the following two scenarios.

When $d_{\tilde{z}} > J$, that is, when the number of excluded instrumental variables is larger than the number of endogenous prices, it is an over-identified system. Since $\partial_{\tilde{\mathbf{z}}_t} \mathbf{g}(\tilde{\mathbf{z}}_t)$ has full column rank, we are able to obtain a minimum distance solution,

$$\underbrace{\partial_{\mathbf{p}_t} f_j(\mathbf{p}_t, \mathbf{x}_t)}_{J\times 1} = \underbrace{\partial_{\mathbf{p}_t} h_j(\mathbf{p}_t, \mathbf{x}_t, \tilde{\mathbf{z}}_t)}_{J\times 1} + (\underbrace{\partial_{\tilde{\mathbf{z}}_t} \mathbf{g}(\mathbf{z}_t)^T}_{J\times d_{\tilde{z}}} \underbrace{\partial_{\tilde{\mathbf{z}}_t} \mathbf{g}(\mathbf{z}_t)}_{d_{\tilde{z}}\times J})^{-1} \underbrace{\partial_{\tilde{\mathbf{z}}_t} \mathbf{g}(\mathbf{z}_t)^T}_{J\times d_{\tilde{z}}} \underbrace{\partial_{\tilde{\mathbf{z}}_t} h_j(\mathbf{p}_t, \mathbf{x}_t, \tilde{\mathbf{z}}_t)}_{d_{\tilde{z}}\times 1}.$$

A.2 Derivation of speical cases

Case 1 When there are J endogenous prices, for each of the endogeneous price p_i there is exactly one excluded instrumental variable $\tilde{z}_i$, we get

$$\begin{pmatrix} \frac{\partial f_j(\mathbf{p}_t, \mathbf{x}_t)}{\partial p_{1t}} \\ \vdots \\ \frac{\partial f_j(\mathbf{p}_t, \mathbf{x}_t)}{\partial p_{Jt}} \end{pmatrix} = \begin{pmatrix} \frac{\partial h_j(\mathbf{p}_t, \mathbf{x}_t)}{\partial p_{1t}} \\ \vdots \\ \frac{\partial h_j(\mathbf{p}_t, \mathbf{x}_t)}{\partial p_{Jt}} \end{pmatrix} + \begin{pmatrix} (\frac{\partial g_1(z_{1t})}{\partial \tilde{z}_{1t}})^{-1} & & \\ & \ddots & \\ & & (\frac{\partial g_J(z_{Jt})}{\partial \tilde{z}_{Jt}})^{-1} \end{pmatrix} \begin{pmatrix} \frac{\partial h_j(\mathbf{p}_t, \mathbf{x}_t)}{\partial \tilde{z}_{1t}} \\ \vdots \\ \frac{\partial h_j(\mathbf{p}_t, \mathbf{x}_t)}{\partial \tilde{z}_{Jt}} \end{pmatrix}.$$

In this special case, it holds that, for $i = 1, 2, \ldots, J$,

$$\frac{\partial f_j(\mathbf{p}_t, \mathbf{x}_t)}{\partial p_{it}} = (\frac{\partial g_i(z_{it})}{\partial z_{it}})^{-1} \frac{\partial h_j(\mathbf{p}_t, \mathbf{x}_t)}{\partial \tilde{z}_{it}}.$$

Case 2 When there are one endogenous price p_t, for this endogenous price there are instrumental variables $\tilde{z}_{1t}$ and $\tilde{z}_{2t}$, we get

$$\frac{\partial f_j(p_t, \mathbf{x}_t)}{\partial p_t} = \frac{\partial h_j(p_t, \mathbf{x}_t)}{\partial p_t} + \left((\frac{\partial g(\mathbf{z}_t)}{\partial \tilde{z}_{1t}})^2 + (\frac{\partial g(\mathbf{z}_t)}{\partial \tilde{z}_{2t}})^2)^{-1}\right) \begin{pmatrix} \frac{\partial g(\mathbf{z}_t)}{\partial \tilde{z}_{1t}} & \frac{\partial g(\mathbf{z}_t)}{\partial \tilde{z}_{2t}} \end{pmatrix} \begin{pmatrix} \frac{\partial h_j(\mathbf{p}_t, \mathbf{x}_t)}{\partial \tilde{z}_{1t}} \\ \frac{\partial h_j(\mathbf{p}_t, \mathbf{x}_t)}{\partial \tilde{z}_{2t}} \end{pmatrix}.$$

In this special case, the price effect is a weighted average using both instrumental variables.

Case 3 When there are two endogenous prices p_{1t} and p_{2t}, for p_{1t} there are two excluded instruments $\tilde{z}_{1t}$ and $\tilde{z}_{2t}$, for p_{2t} there are one excluded instrument $\tilde{z}_{3t}$, we get

$$\begin{pmatrix} \frac{\partial f_j(\mathbf{p}_t, \mathbf{x}_t)}{\partial p_{1t}} \\ \frac{\partial f_j(\mathbf{p}_t, \mathbf{x}_t)}{\partial p_{2t}} \end{pmatrix} = \begin{pmatrix} \frac{\partial h_j(\mathbf{p}_t, \mathbf{x}_t)}{\partial p_{1t}} \\ \frac{\partial h_j(\mathbf{p}_t, \mathbf{x}_t)}{\partial p_{2t}} \end{pmatrix} + \begin{pmatrix} ((\frac{\partial g_1(z_{1t}, z_{2t})}{\partial \tilde{z}_{1t}})^2 + (\frac{\partial g_1(z_{1t}, z_{2t})}{\partial \tilde{z}_{2t}})^2)^{-1} & \\ & (\frac{\partial g_2(z_{3t})}{\partial \tilde{z}_{3t}})^{-2} \end{pmatrix}$$
$$\begin{pmatrix} \frac{\partial g_1(z_{1t}, z_{2t})}{\partial \tilde{z}_{1t}} & \frac{\partial g_1(z_{1t}, z_{2t})}{\partial \tilde{z}_{2t}} & \\ & & \frac{\partial g_2(z_{3t})}{\partial \tilde{z}_{3t}} \end{pmatrix} \begin{pmatrix} \frac{\partial h_j(\mathbf{p}_t, \mathbf{x}_t)}{\partial \tilde{z}_{1t}} \\ \frac{\partial h_j(\mathbf{p}_t, \mathbf{x}_t)}{\partial \tilde{z}_{2t}} \\ \frac{\partial h_j(\mathbf{p}_t, \mathbf{x}_t)}{\partial \tilde{z}_{3t}} \end{pmatrix}.$$

In this special case, it holds that,

$$\frac{\partial f_j(p_t, \mathbf{x}_t)}{\partial p_{1t}} = \frac{\partial h_j(p_t, \mathbf{x}_t)}{\partial p_{1t}} + \left((\frac{\partial g_1(z_{1t}, z_{2t})}{\partial \tilde{z}_{1t}})^2 + (\frac{\partial g_1(z_{1t}, z_{2t})}{\partial \tilde{z}_{2t}})^2)^{-1}\right) \begin{pmatrix} \frac{\partial g_1(z_{1t}, z_{2t})}{\partial \tilde{z}_{1t}} & \frac{\partial g_1(z_{1t}, z_{2t})}{\partial \tilde{z}_{2t}} \end{pmatrix} \begin{pmatrix} \frac{\partial h_j(\mathbf{p}_t, \mathbf{x}_t)}{\partial \tilde{z}_{1t}} \\ \frac{\partial h_j(\mathbf{p}_t, \mathbf{x}_t)}{\partial \tilde{z}_{2t}} \end{pmatrix},$$

$$\frac{\partial f_j(\mathbf{p}_t, \mathbf{x}_t)}{\partial p_{2t}} = (\frac{\partial g_2(z_{3t})}{\partial \tilde{z}_{3t}})^{-1} \frac{\partial h_j(\mathbf{p}_t, \mathbf{x}_t)}{\partial \tilde{z}_{3t}}.$$

Case 4 When there are two endogenous prices p_{1t} and p_{2t}, for p_{1t} there are one excluded instruments $\tilde{z}_{1t}$, for p_{2t} there are one excluded instrument $\tilde{z}_{2t}$, for both of them there is a common excluded instrument $\tilde{z}_{3t}$, we get

$$\begin{pmatrix} \frac{\partial f_j(\mathbf{p}_t,\mathbf{x}_t)}{\partial p_{1t}} \\ \frac{\partial f_j(\mathbf{p}_t,\mathbf{x}_t)}{\partial p_{2t}} \end{pmatrix} = \begin{pmatrix} \frac{\partial h_j(\mathbf{p}_t,\mathbf{x}_t)}{\partial p_{1t}} \\ \frac{\partial h_j(\mathbf{p}_t,\mathbf{x}_t)}{\partial p_{2t}} \end{pmatrix} + \begin{pmatrix} (\frac{\partial g_1(z_{1t},z_{3t})}{\partial \tilde{z}_{1t}})^2 + (\frac{\partial g_1(z_{1t},z_{3t})}{\partial \tilde{z}_{3t}})^2 & \frac{\partial g_1(z_{1t},z_{3t})}{\partial \tilde{z}_{3t}} \frac{\partial g_2(z_{2t},z_{3t})}{\partial \tilde{z}_{3t}} \\ \frac{\partial g_2(z_{2t},z_{3t})}{\partial \tilde{z}_{3t}} \frac{\partial g_1(z_{1t},z_{3t})}{\partial \tilde{z}_{3t}} & (\frac{\partial g_2(z_{2t},z_{3t})}{\partial \tilde{z}_{2t}})^2 + (\frac{\partial g_2(z_{2t},z_{3t})}{\partial \tilde{z}_{3t}})^2 \end{pmatrix}^{-1}$$
$$\begin{pmatrix} \frac{\partial g_1(z_{1t},z_{3t})}{\partial \tilde{z}_{1t}} & & \frac{\partial g_1(z_{1t},z_{3t})}{\partial \tilde{z}_{3t}} \\ & \frac{\partial g_2(z_{2t},z_{3t})}{\partial \tilde{z}_{2t}} & \frac{\partial g_2(z_{2t},z_{3t})}{\partial \tilde{z}_{3t}} \end{pmatrix} \begin{pmatrix} \frac{\partial h_j(\mathbf{p}_t,\mathbf{x}_t)}{\partial \tilde{z}_{1t}} \\ \frac{\partial h_j(\mathbf{p}_t,\mathbf{x}_t)}{\partial \tilde{z}_{2t}} \\ \frac{\partial h_j(\mathbf{p}_t,\mathbf{x}_t)}{\partial \tilde{z}_{3t}} \end{pmatrix}.$$

A.3 Proof of theorems

In this section, we provide a rigorous representation to our theorems. Without loss of generality, we are dealing with a nonparametric prediction problem.

$$y_i = \mu(\mathbf{w}_i) + \epsilon_i,$$

where y is a scalar and $\mathbf{w} \in \mathbb{R}^d$ with d fixed but potentially large.

Assumption A.1. *We have an i.i.d. sample,* $(\mathbf{w}_1, y_1), (\mathbf{w}_2, y_2), \ldots, (\mathbf{w}_n, y_n)$.

Assumption A.2. *The density* $\nu(\cdot)$ *of* $\mathbf{w}$ *is bounded away from* 0 *and* ∞, $\nu(\cdot)$ *and* $\mu(\cdot)$ *are both twice continuously differentiable with bounded second derivatives in a neighborhood of* $\mathbf{w}$, *and* y *has finite second moment,* $\mathbb{E}\, y^2 < \infty$. ϵ *is independent of* $\mathbf{w}$, *has zero mean and finite variance* $\sigma^2 > 0$.

A.3.1 Proof of Theorem 2

We rephrase Theorem 2 as the following theorem.

Theorem A.1. *Given* $\mathbf{w}_0 \in \text{supp}(\mathbf{w})$, *under Assumptions A.1 and A.2,*

$$\mathbb{E}\, \tau_n(m)(\mathbf{w}_0) = \mu(\mathbf{w}_0) + B(m),$$

$$B(m) = \Gamma(2/d+1) \frac{\nu(\mathbf{w}_0)\, \text{tr}(\mu''(\mathbf{w}_0)) + 2\, \mu'(\mathbf{w}_0)^T \nu'(\mathbf{w}_0)}{2\, d\, V_d^{2/d}\, \nu(\mathbf{w}_0)^{1+2/d}}\, m^{-2/d} + o(m^{-2/d}), \tag{8}$$

where $V_d = \frac{\pi^{d/2}}{\Gamma(1+d/2)}$, $\Gamma(\cdot)$ *denotes the Gamma function,* $\nu'(\mathbf{w}_0)$ *and* $\mu'(\mathbf{w}_0)$ *are the first order gradients at* $\mathbf{w}_0$ *for* $\nu(\mathbf{w})$ *and* $\mu(\mathbf{w})$, *respectively,* $\mu''(\mathbf{w}_0)$ *is the Hessian matrix of* $\mu(\cdot)$ *at* $\mathbf{w}_0$, *and* $\text{tr}(\cdot)$ *gives the trace.*

The proof for the above theorem can be found in Demirkaya et al. (2022). Theorem A.1 shows that the first-order asymptotic bias of the bagged nearest neighbors is of order $m^{-2/d}$. Demirkaya et al. (2022) further give the asymptotic order of the bias term after bias reduction and, in general, generalize these estimators to two-scale *distributional* nearest neighbors (TDNN). In this paper, we follow the convention and keep using the name and definition of bagged nearest neighbors.

A.3.2 Proof of Theorem 3

We rephrase Theorem 3 as the following theorem.

Theorem A.2. *Given* $\mathbf{w}_0 \in \text{supp}(\mathbf{v})$, *under Assumptions A.1 and A.2, and assuming* $m \to \infty$ *and* $m/n \to 0$, *then for some positive* σ_n *with* $\sigma_n^2 = O(\frac{m}{n})$, *as* $n \to \infty$,

$$\frac{\tau_n(m)(\mathbf{w}_0) - \mu(\mathbf{w}_0) - B(m)}{\sigma_n} \xrightarrow{\mathscr{D}} N(0,1). \tag{9}$$

The proof for the above theorem can be found in Demirkaya et al. (2022). Theorem A.2 characterizes the asymptotic distribution of the bagged nearest neighbors estimator. Demirkaya et al. (2022) further shows that asymptotic normality also holds for two-scale DNN. They also offer an upper bound for the point-wise MSE and carefully examine the optimality of convergence rate.

A.3.3 Proof of Theorem 4

This appendix shows that the bootstrap can give a valid inference for the bagged nearest neighbors estimator. We will prove Theorem A.3 from the U-statistics perspective instead of the L-statistics. Our proof makes use of the Hoeffding decomposition (Hoeffding, 1948) and the Mallow's distance or Wasserstein distance (Bickel and Freedman, 1981).

Theorem A.3. *Let* $\mathbf{G}_n$ *be the empirical distribution of our sample* $(y_i, \mathbf{w}_i)_{i=1}^n$. *Given* $(y_i, \mathbf{w}_i)_{i=1}^n$, *let* $(y_i^*, \mathbf{w}_i^*)_{i=1}^n$ *be the conditionally independent bootstrap sample with common distribution* $\mathbf{G}_n$. *The bagged nearest neighbors estimator defined on this bootstrap sample is then*

$$\tau_n^*(m)(\mathbf{w}_0) = \binom{n}{m}^{-1} \sum_{1 \le i_1 < i_2 < \ldots < i_m \le n} y_{(1)}(\mathbf{w}_0; (y_{i_1}^*, \mathbf{w}_{i_1}^*), (y_{i_2}^*, \mathbf{w}_{i_2}^*), \ldots, (y_{i_m}^*, \mathbf{w}_{i_m}^*)).$$

Given $\mathbf{w}_0 \in \text{supp}(\mathbf{w})$ *and* σ_n *in Theorem A.2, under Assumptions A.1 and A.2, and assuming* $m \to \infty$ *and* $m/n \to 0$*, then for almost all sample sequences, as* $n \to \infty$*,*

$$\frac{\tau_n^*(m)(\mathbf{w}_0) - \mathbb{E}^*\tau_n^*(m)(\mathbf{w}_0)}{\sigma_n} \xrightarrow{\mathscr{D}} N(0,1).$$

Proof: We first review some results that we will need. Hereafter, we use $\mathbf{T}_i$ as a shorthand notation for $(\mathbf{w}_i, y_i)$. From our proof in Theorem 3, we have

- The bagged nearest neighbors estimator can be decomposed,

$$\tau_n(m) - \mathbb{E}\,\tau_n(m) = \frac{m}{n}\sum_{i=1}^{n}\tilde{g}(\mathbf{T}_i) + \Delta_n(m),$$

  where $\tilde{g}(\mathbf{T}_i) = \mathbb{E}\,\Phi(\mathbf{T}_1, \mathbf{T}_2, \ldots, \mathbf{T}_n|\mathbf{T}_i) - \mathbb{E}\,\Phi(\mathbf{T}_1, \mathbf{T}_2, \ldots, \mathbf{T}_n)$, the canonical Hájek projection of kernel Φ onto $\mathbf{T}_i$. The expectation $\mathbb{E}$ is with respect to $\mathbf{G}$, the distribution of $\mathbf{T}$.

- For some finite positive variance σ^2,

$$\sigma_n^2 = \text{var}\,[\frac{m}{n}\sum_{i=1}^{n}\tilde{g}(\mathbf{T}_i)] = \frac{m^2}{n(2m-1)}\sigma^2.$$

- When $m/n \to 0$ and $n \to \infty$,

$$(\frac{\Delta_n(m)}{\sigma_n})^2 \to 0. \tag{10}$$

- By the Lindeberg–Lévy Central Limit Theorem, we can have

$$\frac{m}{n}\sum_{i=1}^{n}\frac{\tilde{g}(\mathbf{T}_i)}{\sigma_n} \xrightarrow{\mathscr{D}} N(0,1).$$

Let $\mathbf{G}_n$ be the empirical distribution of $(\mathbf{T}_1, \mathbf{T}_2 \ldots, \mathbf{T}_n)$. Given $(\mathbf{T}_1, \mathbf{T}_2, \ldots, \mathbf{T}_n)$, let $(\mathbf{T}_1^*, \ldots, \mathbf{T}_n^*)$ be conditionally independent, with common distribution $\mathbf{G}_n$. The bagged nearest neighbors estimator defined on $(\mathbf{T}_1^*, \ldots, \mathbf{T}_n^*)$ is then

$$\tau_n^*(m)(\mathbf{w}_0) = \binom{n}{m}^{-1}\sum_{1\le i_1<i_2<\ldots<i_m\le n}\Phi(\mathbf{w}_0; \mathbf{T}_{i_1}^*, \mathbf{T}_{i_2}^*, \ldots, \mathbf{T}_{i_m}^*).$$

Similarly, we can have

- For the new distribution $\mathbf{G}_n$,

$$\tau_n^*(m) - \mathbb{E}_n \tau_n^*(m) = \frac{m}{n} \sum_{i=1}^{n} \tilde{g}(\mathbf{T}_i^*) + \Delta_n^*(m),$$

  where the expectation $\mathbb{E}_n$ is with respect to $\mathbf{G}_n$.

- When $m/n \to 0$ and $n \to \infty$, for σ_n^2 defined in (10),

$$(\frac{\Delta_n^*(m)}{\sigma_n})^2 \to 0.$$

With a bit abuse of notation, let $\Rightarrow$ denote weak convergence, it can be shown that

$$\mathcal{L}(\frac{\tau_n^*(m) - \mathbb{E}_n \tau_n^*(m)}{\sigma_n}) \Rightarrow \mathcal{L}(\frac{m}{n} \sum_{i=1}^{n} \frac{\tilde{g}(\mathbf{T}_i^*, \mathbf{G}_n)}{\sigma_n}),$$

which comes from the convergence of the remainder term.

To establish Theorem 3, we still need to prove

$$\mathcal{L}(\frac{m}{n} \sum_{i=1}^{n} \frac{\tilde{g}(\mathbf{T}_i^*)}{\sigma_n}) \Rightarrow \mathcal{L}(\frac{m}{n} \sum_{i=1}^{n} \frac{\tilde{g}(\mathbf{T}_i)}{\sigma_n}).$$

We will use the Mallow's distance introduced in Bickel and Freedman (1981). Before we proceed, we list some properties of the Mallows distance we will use. Let M_p be the Mallow's distance, $p \in [1, \infty)$ and all distributions have finite p-th moments.

- If F and G are distributions on the real line, then

$$M_p(F, G) = \{\int_0^1 |F^{-1}(t) - G^{-1}(t)|^p dt\}^{1/p}.$$

- If $X_1, X_2, \ldots, X_n$ are independent observations from a distribution F, and F_n is their empirical distribution, then almost everywhere,

$$M_p(F_n, F) \to 0.$$

- For any scalar a,

$$M_p(aU, aV) = |a| \cdot M_p(U, V).$$

- If the U_i are independent, likewise for V_i, and $\mathbb{E}U_i = \mathbb{E}V_i$, then

$$M_2^2(\sum_{i=1}^n U_i, \sum_{i=1}^n V_i) \leq \sum_{i=1}^n M_2^2(U_i, V_i).$$

Now we are ready. Let $Z(\mathbf{T}_i) = \mathbb{E}\,\Phi(\mathbf{T}_1, \mathbf{T}_2, \ldots, \mathbf{T}_n|\mathbf{T}_i)$, then we have

$$\begin{aligned}\tilde{g}(\mathbf{T}_i) &= \mathbb{E}\,\Phi(\mathbf{T}_1, \mathbf{T}_2, \ldots, \mathbf{T}_n|\mathbf{T}_i) - \mathbb{E}\,\Phi(\mathbf{T}_1, \mathbf{T}_2, \ldots, \mathbf{T}_n) \\ &= Z(\mathbf{T}_i) - \mathbb{E}Z(\mathbf{T}_i).\end{aligned}$$

First, $\mathbf{G}_n$ is the empirical distribution function of $\mathbf{G}$,

$$M_2(\mathbf{T}^*, \mathbf{T}) \to 0.$$

Since $Z(\cdot)$ is continuous and bounded, we further have

$$M_2(\sqrt{m}\tilde{g}(\mathbf{T}_i), \sqrt{m}\tilde{g}(\mathbf{T}_i^*)) \to 0.$$

By the convolution property of the Mallow's distance,

$$M_2(\frac{m}{n}\sum_{i=1}^n \frac{\tilde{g}(\mathbf{T}_i^*)}{\sigma_n}, \frac{m}{n}\sum_{i=1}^n \frac{\tilde{g}(\mathbf{T}_i)}{\sigma_n}) \leq \frac{m}{n\sigma_n}\sqrt{n}M_2(\tilde{g}(\mathbf{T}_i), \tilde{g}(\mathbf{T}_i^*)),$$

where $\sqrt{n}M_2(\tilde{g}(\mathbf{T}_i), \tilde{g}(\mathbf{T}_i^*)) = O(1)$ since Z and G are both continuous and bounded.

When $m/n \to 0$,

$$M_2(\frac{m}{n}\sum_{i=1}^n \frac{\tilde{g}(\mathbf{T}_i^*)}{\sigma_n}, \frac{m}{n}\sum_{i=1}^n \frac{\tilde{g}(\mathbf{T}_i)}{\sigma_n}) \to 0,$$

which completes our proof of Theorem A.3.

Theorem A.3 is more relevant for our empirical case since this paper needs constructing confidence intervals using the bootstrap. One immediate implication of Theorem A.3 is that the bootstrap variance estimator is consistent. If we alter the above proof, uniform convergence can also be derived by the Berry-Esseen theorem. Interested readers may refer to Demirkaya et al. (2022) for extensions. These properties hold for TDNN as well.

B Additional details for Monte Carlo simulations

We now outline the two additional DGPs in our simulation: one where two products can be complements, and the other where consumers can purchase multiple product varieties in multiple quantities.

Demand when some products are complements. The random-coefficient logit model assumes that products are substitutes for each other. In reality, many products or services are complements (Smith et al., 2019b; Compiani and Smith, 2021), in which case, the random-coefficient logit model will not estimate the correct cross elasticities. In addition, standard parametric aggregate-data demand models do not have a good way to accommodate complements. We now examine how our nonparametric demand estimator recovers own and cross elasticities in markets of substitutes and complements.

We follow Gentzkow (2007) and set up a DGP where some products are substitutes and others are complements. We assume that when an individual ι consumes product j at market t, she enjoys an indirect utility

$$u_{jt} = \delta - \alpha\, p_{jt} + u_{jt} + \epsilon_{jt}.$$

where all subscripts ι have been omitted since individuals are also indifferent as in (Gentzkow, 2007). The unobserved disturbance ϵ_{jt} follows the Type-1 extreme value distribution and are independent and identically distributed across products and markets. In addition, there is an outside option of no purchase with $u_{0t} = \epsilon_{0t}$.

This model allows the existence of a bundle. To be specific, we assume that one unit of Product 1 and one unit of Product 2 can be purchased as a bundle, and for the purchase of this bundle, the indirect utility is

$$u_{bt} = \delta - \alpha\,(p_{1t} + p_{2t}) + \gamma + u_{1t} + u_{2t} + \epsilon_{bt},$$

where γ is the extra utility when purchasing Product 1 and Product 2 as bundles. Gentzkow shows that when γ is positive enough, Product 1 and Product 2 can be complements. In other words, the extra degree of complementarity overcomes the implied substitution from the logit structure.

Our simulation introduces 20,000 markets with γ set to be constant 4. The bottom panel of Figure 2 shows our estimated point-wise own and cross-price elasticities, together with

their closed-form counterparts in the true model. The closed-form derivations of the own and cross-price elasticities do not enjoy the clean forms as in the multinomial logit model. They vary across products and depend on the specification of the bundle composition, extra utility γ, and the preferences.

Demand when consumers can purchase multiple products and units. So far, all DGPs assume that consumers can only choose one product and purchase exactly one unit. The multiple discrete choice literature (Hendel, 1999b; Kim et al., 2002; Dubé, 2004; Chan, 2006; Bhat, 2008) has shown that consumers sometimes purchase multiple products and in different quantities and has demonstrated that such variety/quantity behavior might lead to a different demand function.

We follow Kim et al. (2002), who model consumers as having decreasing marginal utility on each product variety and solving an inequality constrained optimization problem when making product choices. The consumer has a random utility with a translated CES utility function with product-specific parameters governing marginal utility and satiation. She maximizes this utility function subject to nonnegative consumption constraints and budget constraints. As a consequence, variety and quantity are allowed.

In this simulation, we assume consumer ι at market t faces the following problem

$$\begin{aligned} \max_{x_{\iota 1t}, x_{\iota 2t}, x_{\iota 3t}} U(x_{\iota 1t}, x_{\iota 2t}, x_{\iota 3t}) = \sum_{k=1}^{3} \chi_k \frac{\gamma_k}{\alpha_k} \{ (\frac{x_{\iota kt}}{\gamma_k} + 1)^{\alpha_k} - 1 \} \\ \text{s.t.} \qquad x_{\iota 1t}, x_{\iota 2t}, x_{\iota 3t} \geq 0, \\ p_{1t} x_{i1t} + p_{2t} x_{i2t} + p_{3t} x_{\iota 3t} \leq e_\iota. \end{aligned}$$

where $\alpha_k, \gamma_k, \chi_k$ are preference parameters for each of the 3 existing products and are assumed to be constant across markets and individuals. e_ι is individual specific budget, p_{jt} are price for product j in market t.

Solving the above optimization problem with inequality constraints is computationally intensive, and the computation burden grows exponentially with the number of products J. Therefore, in our simulations, we limit $J = 3$.

C Additional tables and figures for the empirical application

Table A1: Yogurt brand and size

Brand	Total sales (units)	Package size (pounds)	Individual sales	Price
Yoplait	626,760,636	0.375	490,723,711	$1.66
		0.25	63,490,752	$2.55
		1.5	26,982,361	$1.73
		1.125	23,543,991	$2.37
		0.6875	11,123,974	$2.31
		0.5	6,517,007	$2.18
		3	1,884,085	$1.52
Dannon	397,986,186	0.375	199,607,994	$1.58
		0.5	71,257,840	$1.40
		1	33,874,083	$2.24
		2	23,157,036	$1.41
		1.5	14,539,104	$1.63
Private label	277,347,934	0.5	211,898,981	$0.98
		0.375	42,254,002	$1.13
		2	13,661,875	$0.97

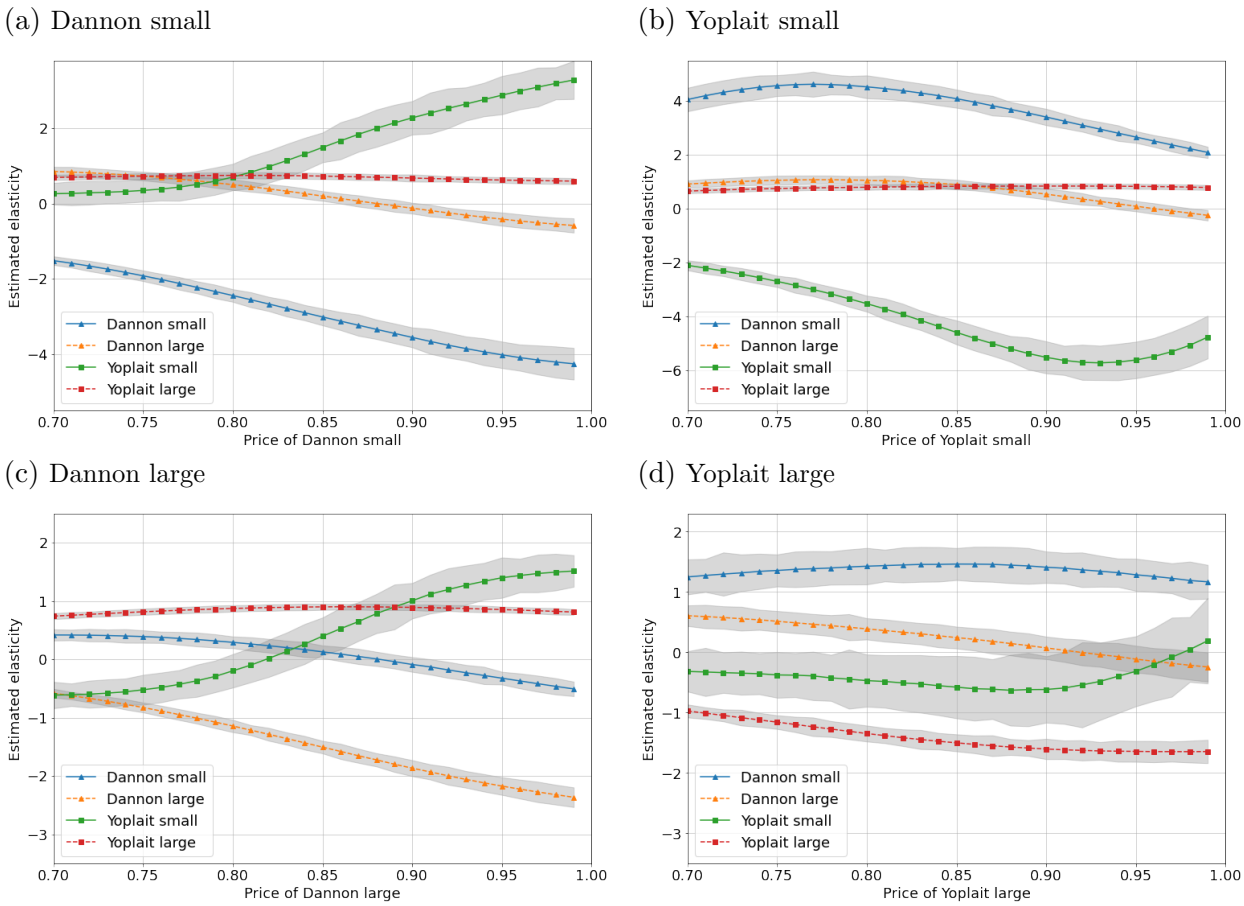
Note: This table summarizes yogurt sales quantity in popular brands and sizes from our dataset. The sales quantity are in units, the sizes are in pounds, and the price has been normalized to dollars per pound. Only most popular brands and sizes have been listed.

Table A2: Yogurt brand and size

Variables		Mean	S.D.	Min	Q1	Median	Q3	Max	Obs.
Yoplait									
	Price	0.92	0.17	0.10	0.80	0.92	1.05	1.58	156,580
small	Sales	416.13	272.49	66.00	207.75	346.13	556.88	1343.25	156,580
	IV	0.84	0.05	0.72	0.81	0.84	0.87	0.97	156,580
	Price	0.92	0.12	0.17	0.83	0.92	1.00	1.46	156,580
large	Sales	89.65	61.50	15.00	43.50	72.00	120.00	301.50	156,580
	IV	0.86	0.02	0.75	0.84	0.86	0.88	0.91	156,580
premium	Price	0.64*	0.08	0.15	0.58	0.63	0.70	0.98	156,580
Dannon									
	Price	0.83	0.18	0.11	0.69	0.80	0.94	1.36	156,580
small	Sales	216.38	180.58	20.25	83.25	160.50	291.75	915.00	156,580
	IV	0.77	0.06	0.58	0.73	0.77	0.81	0.90	156,580
	Price	0.78	0.11	0.27	0.69	0.77	0.86	1.30	156,580
large	Sales	131.13	93.37	18.00	62.00	106.00	172.50	502.00	156,580
	IV	0.73	0.02	0.64	0.72	0.73	0.74	0.80	156,580
Private label									
	Price	0.54	0.09	0.18	0.48	0.53	0.60	1.46	156,580
Other controls									
Store	ACV	0.22	0.09	0.04	0.16	0.20	0.27	1.00	156,580
Chain	Shelf.1	0.25	0.10	0.04	0.15	0.22	0.33	0.49	156,580
	Shelf.2	0.31	0.10	0.03	0.25	0.33	0.40	0.63	156,580
Time	Week	0.51	0.29	0.02	0.27	0.52	0.75	1.00	156,580
	Year	0.58	0.28	0.17	0.33	0.67	0.83	1.00	156,580

Note: This table summarizes yogurt sales in popular brands and sizes from our dataset. *Prices reported here are per 8 oz, except that the premium Yoplait price is per 4 oz.

Figure A1: The estimated own- and cross-price elasticities of yogurt



Note: This figure provides the price elasticity estimates of Dannon small, Dannon large, Yoplait small, and Yoplait large with respect to the price of Dannon small, Dannon large, Yoplait small, and Yoplait large. These elasticities are evaluated at 30 price levels from 0.7 to 1 dollar per 8 oz.