Boundaries of Differentiated Product Markets and Retailer Pricing

Giovanni Compiani and Adam N. Smith*

November 8, 2021

Abstract

This paper studies the effects of misspecified boundaries of competition on optimal retail pricing using store-level supermarket scanner data. We focus on two types of misspecification: (i) misspecification of the demand estimation problem, which can arise from either omitting relevant goods or specifying an overly restrictive model of demand; (ii) misspecification of the retailer's decision problem, which can arise from the retailer separately optimizing prices in each category but failing to account for, and thus internalize, cross-category effects. Both sources of misspecification are relevant in differentiated product markets where goods are broadly related but in ways that may not be immediately obvious a priori. Quantifying the costs associated with either form of misspecification is challenging because it requires a flexible yet valid demand system. To this end, we take a nonparametric approach to estimating a multi-category demand system that imposes minimal restrictions on the sign/magnitude of cross-price effects and also satisfies key properties required by economic theory. Our first set of empirical results is descriptive. We use data across nine diverse product groups to show that cross-category effects are appreciably large and often nuanced in their sign and magnitude. We then zoom in on refrigerated juices and estimate demand nonparametrically across five juice categories where we find empirical support for a flexible model that can accommodate both substitutes and complements. We solve for optimal prices under both sources of misspecification and estimate profit losses to be in the 4-14% range.

Keywords: nonparametric demand estimation, demand spillovers, multi-category demand, model misspecification, omitted variable bias

*Compiani: Booth School of Business, University of Chicago, Email: giovanni.compiani@chicagobooth.edu. Smith: UCL School of Management, University College London, Email: a.smith@ucl.ac.uk. We thank Jean-Pierre Dubé, Günter Hitsch, Yufeng Huang, Ilya Morozov, Stephan Seiler, Avner Strulov-Shlain, Xu Zhang, seminar participants at Chicago Booth and the European Quant Marketing Seminar, and conference participants at Marketing Science 2021 for helpful comments. We also thank Leesok Kim for his research assistance. All errors are our own.

1 Introduction

Delineating boundaries of product competition is at the center of many questions in marketing and economics. Yet in most empirical work, such boundaries already begin to take shape prior to the analysis. For example, consider two key inputs into any demand analysis: a subset of goods to represent the consumer choice set, and a stochastic specification of demand to take to the data. Decisions about the former implicitly restrict substitution between the included goods and the remaining goods excluded from the analysis, while decisions about the latter can also restrict the scope of substitution through various functional form and/or parametric assumptions. As a consequence, questions of misspecification often linger when estimating demand for differentiated products. Is the set of goods defined broadly enough in order to capture all relevant dimensions of substitution? Is the model flexible enough to capture this substitution? How will misspecification in the estimation problem impact subsequent policy prescriptions?

The goal of this paper is to quantify the costs of misspecified boundaries of competition on optimal retailer pricing policies. Our specific target is the loss in profits accrued by a retailer who solves for optimal prices under two possible forms of misspecification. The first is a misspecification of boundaries in the *estimation problem* which arises when (i) the product market is defined to be too narrow, such as when a relevant good is omitted from the demand system; or (ii) the demand model is overly restrictive and, for example, only allows for strict substitution in a market with complementary goods. In either case, the concern is that estimates of key demand derivatives will be biased. The second is a misspecification of boundaries in the retailer's *decision problem* which arises when the retailer omits relevant goods from the objective function used to set prices. The scope of a retailer's pricing problem is at the center of many questions related to category management (Basuroy and Walters, 2001) and the organization of supply (Kadiyali et al., 2000; Sudhir, 2001). For example, both retailer profits and the balance of market power have been shown to depend on whether a retailer sets prices jointly across categories (vs. separately in each category) and internalizes cross-category substitution (Smith and Thomassen, 2012; Thomassen et al., 2017; Ershov et al., 2021).

Our empirical context is pricing consumer packaged goods (CPGs) using store-level grocery retail data. We focus on estimating demand within product groups (e.g., salty snacks) spanning multiple related categories (e.g., potato chips and pretzels). Quantifying the costs of misspecification in this empirical context is challenging because we need estimates of substitution patterns that are flexible — in terms of the sign and magnitude of demand derivatives — but also obey certain properties of microeconomic theory to ensure valid counterfactual profit predictions. Workhorse demand models for differentiated products such as BLP (Berry et al., 1995) are micro-founded but relatively inflexible in the sense that functional forms constrain cross-price derivatives to be positive. In contrast, many regression or ML-based models are relatively flexible but often lack micro-foundations, cannot address the endogeneity of prices, or lack an approach to inference.

To overcome these challenges, we take a nonparametric approach to demand estimation. We first specify a demand system in which market quantities are expressed as general functions of

prices, product characteristics, and (endogenous) structural unobservables. We then adopt the nonparametric estimator of Compiani (2020) which uses Bernstein polynomials to approximate the inverse of the demand system. A nonparametric approach is attractive for our purposes because both the target demand function (i.e., the estimand) and our estimator of that demand function are flexible. In particular, we only require the underlying demand function to satisfy minimal identification restrictions (Berry and Haile, 2014) which are micro-founded but allow for a range of substitution/complementarity patterns. Moreover, the advantage of Bernstein polynomials is that they are amenable to imposing relevant constraints. That is, it is straightforward to enforce the required economic restrictions such as own-price monotonicity (i.e., downward-sloping demand) via convex inequality constraints on the Bernstein coefficients. We can also impose other constraints such as strict substitution in order to derive less flexible but nested models to take to the data.

Our empirical analysis uses store-level transaction data from one major grocery retail chain in the United States. We provide two main sets of results. The first set is descriptive: we fit a battery of log-log models to data from nine broad product groups to show that cross-category effects are appreciably large and so a more structural approach to investigating misspecification is warranted. In doing so, we also find that some cross effects are easier to sign a priori than others. For example, the majority of cross effects in the Baking Goods or Detergents/Softener groups are negative, which is consistent with our expectation and prior empirical work on cross-category complements (Manchanda et al., 1999; Song and Chintagunta, 2006). In other groups like Refrigerated Juice, estimated cross effects are much more dispersed and so many categories are neither "obvious substitutes" nor "obvious complements." This would ordinarily complicate the specification of a structural model of demand and provides further motivation for our flexible nonparametric approach.

Our second set of results is structural and zooms in on the demand for five product categories in the Refrigerated Juice product group. We take two versions of our nonparametric estimator to the data: a "flexible" specification that imposes no constraints on the sign of cross-price derivatives and a "substitutes" specification that constrains cross-price derivatives to be positive. We find empirical support, both in terms of model fit and estimated elasticities, for complementarities in this system of goods. We then use the flexible specification to quantify biases in estimated elasticities that result from omitting relevant goods from the demand system. Specifically, we estimate a sequence of smaller models in which we iteratively remove one category from the demand system, and then compare estimated elasticities to the "true" elasticities from the full model including all five categories. We find that the direction of the bias in the smaller, misspecified models is consistent with standard omitted variable bias intuition. That is, the sign of the bias is governed by both (i) the sign of the omitted elasticity; and (ii) the correlation in prices between the included price(s) and the omitted price. While most recent work on large-scale demand estimation emphasizes (i) alone, our results suggest that checking (ii) is just as important.

Finally, we solve for the optimal retail prices in Refrigerated Juice under each nonparametric specification ("flexible" and "substitutes"), with the "substitutes" version serving as the misspec-

ified model. We also consider misspecification in the retailer's decision problem and solve for a misspecified pricing policy in which the retailer separately maximizes product-specific profits ("isolated" pricing), as well as a correctly specified pricing policy, in which the retailer maximizes total profits for Refrigerated Juice ("joint" pricing). Relative to the case of joint pricing with a flexible model, profits are 3.8% lower when demand is misspecified (by incorrectly assuming substitution among all goods) but prices solved jointly; 11.6% lower when demand is correctly specified but prices are solved separately in each category; and 13.2% lower when both the model is misspecified and prices are optimized separately in each category. Thus, our results suggest that boundaries of the decision problem are first-order as the worst-case profits under joint pricing are still better than the best-case profits under isolated pricing. However, given that our misspecified demand system is still nonparametric, our results likely provide a lower bound to the losses in retailer profits due to misspecification. In other words, a more restrictive model that assumes substitution among all goods may lead to larger profit losses.

Our work contributes to a large literature on pricing differentiated products. Much existing work has focused on pricing relatively narrow assortments of imperfect substitutes, such as varieties of cereal (Nevo, 2001), ketchup (Besanko et al., 2003), milk (Handel et al., 2013), orange juice (Montgomery, 1997; Chintagunta et al., 2003; Nair et al., 2005; Dubé et al., 2008, 2010), and yogurt (Vilcassim and Chintagunta, 1995; Kim et al., 2002; Draganska and Jain, 2006). While focusing on a single category can simplify the analysis, one potential concern is that substitution patterns will be distorted if the related goods (outside of the defined category) are omitted from the system.[1] For example, when modeling the demand for orange juice alone, all remaining juice flavors would enter into an outside good typically assumed to be weakly separable from the inside goods (Chintagunta and Nair, 2011). Smith et al. (2019) show that demand for juice is generally not separable by flavor, calling into question the simple functional relationships commonly assumed between the inside and outside goods. More broadly, this raises questions of the x scope of a product market when solving for optimal prices and, in turn, what a misspecified model implies for optimal prices and profits.

Pricing wider product assortments has been made possible with the rise of multi-category models of demand (see Berry et al., 2014, for a review). Relaxing assumptions of additively separable utility and discrete choice gives rise to models with more flexible derivatives and therefore allows for richer substitution patterns. However, many empirical applications have focused on categories where the nature of complementarity is more clear a priori, such as cake mix and cake frosting (Manchanda et al., 1999; Ma et al., 2012), laundry detergent and fabric softener (Song and Chintagunta, 2006, 2007; Mehta, 2007), or milk and cereal (Lee et al., 2013). Moreover, most structural multi-category demand models are developed at the individual-level and require data on household choices (e.g.,

[1] Note that in CPG markets, academic researchers typically adopt the category definitions given by data providers like IRI and Nielsen. These categories may not align with the category definitions used by retailers, however. For example, "refrigerated orange juice" is often treated as a category in academic research, but the broader set of "refrigerated juices" could constitute a category for the retailer (in the sense that the retailer coordinates marketing decisions for all juice products, not just orange juices). Our goal is to document the extent of cross-category substitution to then comment on how broad or narrow categories should be defined.

Song and Chintagunta, 2007; Mehta, 2007; Thomassen et al., 2017; Lewbel and Nesheim, 2019). Our paper contributes to this literature by providing a fully structural yet flexible model of aggregate demand that remains suitable when the focal market spans multiple categories and consists of products that fall in between the extremes of imperfect substitutes and imperfect complements.

Our paper also contributes to a fast-growing literature on flexible demand estimation. In this literature, a few distinct approaches have emerged to accommodate more flexible substitution patterns (and thus a wider assortment of goods). One approach is to extend the framework of BLP using more flexible specifications of inverse demand functions (Fosgerau et al., 2021; Monardo, 2021) or distributions on the random coefficients (Fox et al., 2016; Wang, 2021b). A second approach is to reformulate the logit discrete choice assumption as discrete choice over product bundles (Iaria and Wang, 2020; Wang, 2021a; Ershov et al., 2021), which allows for joint consumption and product complementarities. A third approach is to estimate demand functions, or key derivatives of those functions, nonparametrically (Blundell et al., 2017; Wang and Huang, 2019; Compiani, 2020). A final approach is to exploit flexible machine learning methods, such as embedding models (Kumar et al., 2020; Chen et al., 2020), matrix factorization (Donnelly et al., 2019; Ruiz et al., 2020), and deep learning (Gabel and Timoshenko, 2021).

While each of the estimation frameworks outlined above offers some aspect of flexibility, we believe our nonparametric approach is most suitable for our research questions. Quantifying the costs of misspecification requires a "ground truth" model that guards against functional form misspecification, which is a virtue of any nonparametric approach. Our framework also makes it possible to impose theory-based restrictions that discipline the estimator and enforce key regularity properties on the profit function used in our pricing counterfactuals. Finally, our framework nests more restrictive models — such as the specification where all products are substitutes — which allows us to more cleanly compare models and quantify the incremental value of model restrictions. One drawback is that our estimator will be subject to a curse of dimensionality that will make it infeasible to estimate demand for very large assortments. That said, we believe our results provide a valuable first step in demonstrating that misspecification matters for optimal pricing, even in a just few related categories, which should in turn motivate more work on a larger scale in this area.

The remainder of this paper is organized as follows. Section 2 outlines our focal demand system and Section 3 presents the nonparametric estimator. Section 4 reports results from a variety of simulation studies to illustrate the flexibility of the proposed method as well as the scope for the effects of boundary misspecification on price elasticities. Section 5 describes the retail scanner data used in our empirical analysis. Section 6 presents descriptive results on cross-category elasticities and price correlations across a broad set of nine product groups. Section 7 presents results from our nonparametric analysis of demand and subsequent pricing counterfactuals. Section 8 concludes.

2 A Flexible Model of Demand

Consider a continuum of consumers belonging to different markets, where each market is defined by a fixed assortment of goods $j = 1, \ldots, J$ and characteristics $\chi_t = (x_t, p_t, \xi_t)$. Here $x_t = (x_{1t}, \ldots, x_{Jt})$ is a vector of observable exogenous characteristics with $x_{jt} \in \mathbb{R}^K$, $p_t = (p_{1t}, \ldots, p_{Jt})$ is a vector of prices, and $\xi_t = (\xi_{1t}, \ldots, \xi_{Jt})$ is a vector of scalar unobservable characteristics. The demand system $\sigma(\chi)$ is then given by the function mapping demand shifters $\chi \in \mathcal{X}$ into a J-dimensional vector of real-valued outcomes:

$$\sigma(\chi) = (\sigma_1(\chi), \ldots, \sigma_J(\chi)) : \mathcal{X} \to \mathbb{R}^J \tag{1}$$

The range of $\sigma(\cdot)$ can accommodate cases where the dependent variable represents market shares or quantities. In other words, the framework can be used in both discrete choice and quantity choice settings. This is important for at least two reasons. First, the assumption that each consumer buys at most one unit from the choice set is sometimes likely to be violated. For example, consumers might purchase several units of multiple different goods (Kim et al., 2002; Dubé, 2004). This is especially important in the presence of complementary products, which inherently violate the substitution patterns implied by standard discrete choice models. Second, even if the products are substitutes, the data available to researchers often only contain information about quantities as opposed to market shares. In order to apply standard models, the researcher typically needs to take a stand on the size of the market and use that to convert quantities into shares. This then raises questions about how robust the results are to the assumed market definition. In contrast, the approach discussed here does not require any such assumptions.

The model in (1) is highly general in that it places no restrictions on the interactions among elements of χ_t in the generation of demand. However, in practice it is common to assume an index structure for a subset of demand shifters. Specifically, we partition the observable characteristics $x_t = (x_t^{(1)}, x_t^{(2)})$ with a scalar $x_{jt}^{(1)}$ and define the index:

$$\delta_{jt} = \beta_j x_{jt}^{(1)} + \xi_{jt} \tag{2}$$

which imposes a weak separability condition in that the marginal rate of substitution between $x_{jt}^{(1)}$ and ξ_{jt} must be constant. In contrast, the way in which prices and the $x_t^{(2)}$ variables enter the demand system is left unrestricted. With this index structure, we can rewrite demand as:

$$s_{jt} = \sigma_j(\chi_t) = \sigma_j\left(\delta_t, p_t, x_t^{(2)}\right). \tag{3}$$

The identification of the demand system above is discussed extensively in Berry and Haile (2014). The idea is to start with the inverse of demand:

$$\delta_{jt} = \sigma_j^{-1}\left(s_t, p_t, x_t^{(2)}\right) \tag{4}$$

which is guaranteed to exist under mild assumptions (Berry et al., 2013). Crucially for our pur-

poses, these assumptions accommodate both substitutability and complementarity across goods.[2] Inversion is a useful first step because it allows the system in (1) to be rewritten as a system of equations with only one unobservable structural error per equation:

$$\beta_j x_{jt}^{(1)} = \sigma_j^{-1}\left(s_t, p_t, x_t^{(2)}\right) - \xi_{jt} \tag{5}$$

Given a set of price instruments z_t excluded from the demand system such as cost shifters or Hausman instruments, equation (5) can be used to form the moment conditions:

$$\mathbb{E}\left(\xi_{jt}|z_t, x_t\right) = \mathbb{E}\left[\sigma_j^{-1}\left(s_t, p_t, x_t^{(2)}\right) - x_{jt}^{(1)}\middle| z_t, x_t\right] = 0, \tag{6}$$

where β_j is normalized to 1 as it cannot be separately identified from the scale of ξ_j. As shown in Berry and Haile (2014), the functions σ_j^{-1} — and thus the demand system σ — are then identified under a completeness condition, the nonparametric equivalent of a standard rank condition.

3 Estimation

3.1 A Nonparametric Approach

The moment conditions in (6) lead to a natural flexible GMM estimation approach. Specifically, following Compiani (2020), we will approximate σ_j^{-1} using Bernstein polynomials and estimate the coefficients on those polynomials by minimizing a GMM objective function based on the moments in (6). The choice of Bernstein polynomials is motivated by the fact that these basis functions allow the researcher to easily impose a number of constraints, such as monotonicity and exchangeability, that are motivated by economic theory and help improve the performance of the estimator. We will discuss some of these constraints in the next subsection. Here, we provide a brief primer on Bernstein polynomials and their approximation properties.

For a positive integer m, the Bernstein basis functions of degree m are defined as

$$b_{v,m}(u) = \binom{m}{v} u^v (1-u)^{m-v}, \tag{7}$$

where $v = 0, 1, \ldots, m$ and $u \in [0, 1]$. A helpful property of Bernstein polynomials is that they can approximate well any continuous function when the coefficients on the polynomials are chosen to be the equal to the target function evaluated at a grid of points. More formally, the following result holds (see, e.g., Gal, 2008).

Result 1. *Let f be a continuous real-valued function on $[0, 1]$ and let*

$$B_m(u) = \sum_{k=0}^{m} f(k/m)\binom{m}{k} u^k (1-u)^{m-k}.$$

[2]The assumption labeled "connected substitutes" in Berry and Haile (2014) requires that there exist a transformation of the demand system exhibiting some degree of strict substitution. As shown in Example 1 of Berry et al. (2013), this condition can be satisfied by a demand system in which all goods are complementary with each other.

Then,

$$\sup_{u\in[0,1]} \left| B_m(u) - f(u) \right| \to 0$$

as $m \to \infty$.

In words, when the coefficients on the Bernstein polynomials are equal to the value of the target function f at an equally-spaced grid of points $(f(0), f(1/m), \ldots, f((m-1)/m), f(1))$, then the resulting approximation is uniformly good to an arbitrary precision as the degree m grows to infinity. A similar result holds for functions of multiple arguments, which can be approximated by the tensor product of univariate Bernstein polynomials. This is helpful because it means that for an approximation to be good it must be that, in the limit, the coefficients on the Bernstein polynomials satisfy whatever properties the target function exhibits. For instance, if the target function is monotonically increasing, it must be that the Bernstein coefficients also satisfy a monotonicity condition.

Consider a simple example where the degree m is set to 2. In this case, there are three Bernstein polynomials: $(1-u)^2, 2u(1-u)$ and u^2. Using Result 1, we obtain that if the target function f is monotonically increasing, then the coefficients on the three polynomials must also be increasing, i.e., the coefficient on $(1-u)^2$ must be less than or equal to that on $2u(1-u)$ and this in turn must be less than or equal to the coefficient on u^2.[3] Importantly, the fact that these constraints are linear in the Bernstein coefficients makes them easy to enforce computationally. This is because the GMM objective function is quadratic in the unknown functions σ_j^{-1} and the Bernstein coefficients are linear in the approximation to the unknown functions. Thus, we obtain a well-behaved problem with a convex objective function subject to linear constraints, which can be solved using off-the-shelf global solvers. By the same argument, convex constraints in the Bernstein coefficients can also be accommodated. In contrast, existing estimation procedure such as BLP typically feature non-convex programs, for which solvers that are guaranteed to converge to the global minimum are not available.[4]

3.2 Constraints: Trading off Flexibility vs. Tractability

The demand specification (i.e., the estimand) outlined above is flexible in that no restrictions are placed on the cross-price derivatives of the demand functions $\sigma_j(\cdot)$ and so the system can admit both substitutes and complements. This is in contrast to many widely used market-level demand systems such as BLP which are based on assumptions of discrete choice and additively separable utility, thus ruling out complementarity.[5] The estimator is also flexible in that it contains enough parameters to serve as a first-order approximation to any underlying Marshallian demand system

[3] More precisely, these conditions become necessary in the limit as $m \to \infty$. Letting the degree of the approximation grow to infinity with the sample size is standard in nonparametric estimation.

[4] Conlon and Gortmaker (2020) find that their implementation of the BLP estimator performs well when optimal instruments are used and restrictions from the supply side are imposed.

[5] Additively separable utility also underlies discrete-continuous demand models (Nair et al., 2005).

Table 1: Summary of Constraints

	Economic Meaning
1. Exchangeability	Demand σ_j invariant to permutations of characteristics $(\delta, p, x^{(2)})$
2. Price inside the index	Constant marginal rate of substitution between price and other characteristics in the index
3. Symmetry of $\mathbb{J}^p_\sigma(\delta, p)$	No income effects
4. Symmetry of $\mathbb{J}^\delta_\sigma(\delta, p)$	Utility of good j linear in δ_j and doesn't depend on δ_k for $k \neq j$
5. Diagonal dominance of $\mathbb{J}^\delta_\sigma(\delta, p)$	Own effects dominate cross effects
6. Own-good monotonicity	Demand σ_j increases in $(\delta_j, -p_j)$

	Necessary for invertibility?	Necessary for identification?	Reduction in # parameters?	Constraint
1. Exchangeability	No	No	Yes	Linear
2. Price inside the index	No	No	Yes	Linear
3. Symmetry of $\mathbb{J}^p_\sigma(\delta, p)$	No	No	No	Nonlinear*
4. Symmetry of $\mathbb{J}^\delta_\sigma(\delta, p)$	No	No	No	Linear
5. Diagonal dominance of $\mathbb{J}^\delta_\sigma(\delta, p)$	No	No	No	Linear
6. Own-good monotonicity	Yes	Yes	No	Linear

*The symmetry of $\mathbb{J}^p_\sigma(\delta, p)$ constraint is nonlinear only if price is not inside the index; otherwise, it is linear.

(Diewert, 1974; Pollak and Wales, 1992).[6] Although there are many other demand systems based on flexible functional forms (e.g., the almost ideal demand system of Deaton and Muellbauer, 1980), these models are typically made stochastic by appending an iid mean zero error term to the end of each estimating equation and therefore do not incorporate "structural" errors. That is, the errors cannot be interpreted as *unobserved* demand shifters (giving rise to the endogeneity of prices) that jointly enter the demand equation for each good (Berry and Haile, 2021).

Flexibility does come at a cost, however. Flexible estimators necessarily contain many parameters and can therefore suffer from a curse of dimensionality. In our case, the number of Bernstein coefficients used in the approximation of $\sigma_j^{-1}(\cdot)$ grows exponentially in the number of goods, and so a fully unrestricted estimator will be practically intractable for a moderate to large number of goods. A fully unrestricted estimator is also undesirable in that the resulting system is not guaranteed to be consistent with properties of a valid Marshallian demand system. As discussed in Compiani (2020), one of the main reasons that Bernstein polynomials are attractive is that they are amenable to a variety of microfounded constraints. In what follows, we briefly discuss different constraints that can be imposed on the nonparametric estimator in order reduce to the number of parameters and/or ensure that various restrictions required by economic theory are satisfied. A summary of restrictions is provided in Table 1.

[6]The standard definition of a "flexible functional form" is that it can provide a second-order approximation to an arbitrary twice differentiable function. While flexible functional forms are often used to approximate utility or cost functions (e.g., Deaton and Muellbauer, 1980), it is also equivalent to target first-order approximations to the demand system directly (Pollak and Wales, 1992, ch. 3). For a system of J goods, this requires at least $1 + J + J(J+1)/2$ free parameters in each demand equation.

Exchangeability A demand system is exchangeable if $\sigma_j(\cdot)$ is the same for all j and it is invariant to permutations of product characteristics $(\delta_k, p_k, x_k^{(2)})$ for all inside goods $k \neq j$. Exchangeability is satisfied in random coefficients logit models under the standard assumption that the distribution of random coefficients is the same across all goods. Moreover, systematic difference across goods can be accounted for via product fixed effects in the indices δ_{jt}. As shown in Compiani (2020), by the approximation properties of Bernstein polynomials in Result 1, exchangeability translates into equality constraints on the Bernstein coefficients. Thus, this constraint effectively reduces the number of parameters that need to be estimated, alleviating the curse of dimensionality.

Index Restriction Instead of allowing prices p_{jt} and covariates $x_t^{(2)}$ to enter the demand system fully flexibly, one could include them in the linear indices δ_{jt}. This leads to the simpler demand system $s_t = \sigma(\delta_t)$, which, as above can be inverted as follows:

$$\delta_{jt} = \sigma_j^{-1}(s_t)$$

The advantage is that the estimands σ_j^{-1} are now a function of J arguments only, instead of $2J + n_{x^{(2)}}$, where $n_{x^{(2)}}$ denotes the dimension of the covariates $x^{(2)}$. Estimating functions of fewer arguments reduces the number of parameters and thus alleviates the curse of dimensionality. Note that including p_{jt} in the linear index δ_{jt} does not imply that *demand* is linear in prices, but instead simply requires that the marginal rate of substitution between price and other characteristics in the index be constant. In BLP-type models, for example, this index restriction is consistent with price entering the indirect utility function with a non-random coefficient.[7] This highlights a trade-off between the amount of heterogeneity allowed in the model and the number of parameters that need to be estimated.

Slutsky Symmetry Slutsky symmetry refers to the symmetry of the Slutsky substitution matrix and is required for integrability. Since we observe and model uncompensated (Marshallian) demand, however, we can only impose Slutsky symmetry if income effects are zero. When prices enter the linear indices δ_{jt} (see "Index Restriction" above), symmetry can be imposed via linear constraints on the Bernstein polynomial, whereas the constraints become nonlinear — and generally non-convex — when prices are not part of the index (Compiani, 2020).

Diagonal Dominance Diagonal dominance requires the magnitude of the own-price effects to be at least as large as the sum of magnitudes of the cross-price effects. Specifically, the demand system $\sigma(\cdot)$ satisfies (column) diagonal dominance if:

$$\sum_{k \neq j} \left| \frac{\partial \sigma_k}{\partial p_j} \right| \leq \left| \frac{\partial \sigma_j}{\partial p_j} \right|. \tag{8}$$

[7] Because preference heterogeneity is only weakly identified with aggregate data (Bodapati and Gupta, 2004; Albuquerque and Bronnenberg, 2009; Dunker et al., 2017), it is common practice to only place random coefficients on a subset of variables in BLP models.

This restriction is satisfied by many common demand models (including BLP). Compiani (2020) shows that diagonal dominance can be enforced via linear inequality constraints on the Bernstein coefficients. Although not identical, a similar version of this property appears as a consequence of the homogeneity and adding-up (Cournot aggregation) restrictions required for integrability.[8] The only difference is that homogeneity and adding-up require that the *price-weighted* own-effect dominate the sum of *price-weighted* cross-effects. Also note that Anderson and Vilcassim (2001) propose a condition that amounts to diagonal dominance of the *profit function* with respect to prices in order to ensure finite optimal prices. The intuition is that the retailer's first-order conditions depend directly on derivatives of the aggregate demand function, and so optimal prices can approach infinity when cross-product substitution is left "unbounded."

Own-Good Monotonicity Monotonicity requires each the demand for each product to be increasing in its own quality index δ_{jt} and decreasing in its own price. Requiring negative own-price effects with respect to the uncompensated (Marshallian) demand function will, in turn, ensure negativity of the Slutsky substitution matrix which is required for integrability (assuming nonnegative income effects). Again, by the approximation properties in Result 1, monotonicity can be imposed easily through inequality constraints on the Bernstein coefficients.

4 Monte Carlo Simulations

In this section, we illustrate the flexibility of our approach by showing that the nonparametric estimator is able to capture complex patterns of substitution/complementarity. We also investigate the effects of misspecified demand boundaries on estimated elasticities. To start, consider a setting in which the first $J-1$ goods are substitutes to each other, but are complementary to the Jth good. We generate prices as $p_{jt} = \frac{e^{\tilde{p}_{jt}}}{1+e^{\tilde{p}_{jt}}}$, where $\tilde{p}_{jt}$ are iid standard normal across products j and across markets t. Further, we let x_{jt} be uniformly distributed on the unit interval also iid across products and markets, and let the unobservables ξ_{jt} be iid normal with mean zero and standard deviation 0.15. Then, letting $\delta_{jt} = -2p_{jt} + x_{jt} + \xi_{jt}$, we define the quantity of good j as:

$$q_{jt} = \exp\left(\delta_{jt} - 0.15 \sum_{k<J,\ k\neq j} \delta_{kt} + 0.15\delta_{Jt}\right) \tag{9}$$

for goods $j = 1, \ldots, J-1$ and

$$q_{Jt} = \exp\left(\delta_{Jt} + 0.15 \sum_{k<J} \delta_{kt}\right) \tag{10}$$

for the final Jth good. Note that the sign of the coefficients on the indices δ_{kt} $(k < J)$ and δ_{Jt} produce the desired block-wise pattern of substitutes and complements. The elasticities induced

[8] See the discussion of Propositions 2.E.1 and 2.E.2 in Mas-Colell et al. (1995).

Table 2: Elasticity Estimates

	True Value	(I) $J=3$	(II) $J=3$ Subst	(III) $J=2$ $\rho=0$	(IV) $J=2$ $\rho=0.7$	(V) $J=2$ $\rho=-0.7$
$\mathcal{E}_{11}$	-1.00	-0.98 (0.026)	-1.01 (0.026)	-0.98 (0.026)	-1.07 (0.027)	-0.89 (0.023)
$\mathcal{E}_{22}$	-1.00	-0.98 (0.024)	-1.00 (0.023)	-0.98 (0.023)	-1.06 (0.025)	-0.88 (0.022)
$\mathcal{E}_{33}$	-1.00	-1.02 (0.033)	-1.05 (0.033)			
$\mathcal{E}_{12}$	0.15	0.14 (0.005)	0.11 (0.006)	0.14 (0.005)	0.04 (0.004)	0.24 (0.007)
$\mathcal{E}_{21}$	0.15	0.14 (0.005)	0.11 (0.005)	0.14 (0.005)	0.04 (0.004)	0.24 (0.007)
$\mathcal{E}_{13}$	-0.15	-0.15 (0.006)	0.00 (0.000)			
$\mathcal{E}_{31}$	-0.15	-0.15 (0.005)	0.00 (0.000)			
$\mathcal{E}_{23}$	-0.15	-0.14 (0.005)	0.00 (0.000)			
$\mathcal{E}_{32}$	-0.15	-0.15 (0.006)	0.00 (0.000)			

Notes: In column (I), we impose substitution between the first two goods, but leave the $\{1,2\}$-$\{3\}$ effects unrestricted. In column (II), we erroneously impose substitution between all three goods. In columns (III)-(V), we fit misspecified models where the third good is omitted and under varying degrees of price correlation ρ. Standard errors are reported in parentheses.

by this system will be:

$$\mathcal{E}_{jkt} = \frac{\partial q_{jt}/q_{jt}}{\partial p_{kt}/p_{kt}} = \begin{cases} -2p_{kt} & \text{if } j = k \\ -0.15 \cdot 2p_{kt} & \text{if } j = J \text{ or } k = J, \text{ for } j \neq k \\ 0.15 \cdot 2p_{kt} & \text{otherwise.} \end{cases}$$

We generate 100 data sets from the model above, with each having $T = 10,000$ markets. We then apply the nonparametric estimator to each data set where we impose the following constraints: an index restriction (with price inside the index), diagonal dominance, weak substitutability and exchangeability among the first $J-1$ goods, and own-good monotonicity for all goods. The relationship between the first $J-1$ goods and the Jth good is left unrestricted.

Elasticity estimates from a $J = 3$ good system are reported in column (I) of Table 2. We specifically report estimates of the sampling distribution of the median. That is, for each data set we compute the median elasticity across observations, and then report the mean and standard deviation of those point estimates. Because we generate prices that are defined on the unit interval and centered at 0.5, the true own elasticities are equal to $\mathcal{E}_{jj} = -1$, the true cross elasticities between goods $j < J$ and $k < J$ are equal to $\mathcal{E}_{jk} = 0.15$, and the true cross elasticities between goods $j < J$ and $k = J$ are equal to $\mathcal{E}_{jk} = -0.15$. We find that the nonparametric estimator accurately recovers the true values of all model parameters. Elasticity estimates from larger $J = 4$ and $J = 5$ good systems, which also accurately recover true values, are reported in Appendix A.

So far, we have shown that the nonparametric estimator is flexible and can accurately recover elasticities in a demand system with a mix of substitutes and complements. Our next objective is to highlight the possible effects of misspecified boundaries in the estimation problem. We first estimate a model that erroneously assumes weak substitution between all three goods. The estimates from this misspecified model are reported in column (II). Although the own elasticities remain accurately

recovered, we find a clear bias in all cross elasticities. Specifically, the cross elasticities involving the third good are all estimated to be very close to zero because of the binding positivity constraint which, in turn, leads to a downward bias on the estimated elasticities between goods 1 and 2.[9]

Next, we estimate a misspecified model that excludes the third good from the demand system entirely. We also relax the assumption that prices are iid across goods, and define $\rho = \text{corr}(p_{jt}, p_{kt})$ to be the correlation between the price of good $j < J$ and omitted good $k = J$. We then consider three scenarios: one where $\rho = 0$ and prices remain iid across goods, a second where $\rho = 0.7$ and thus the price of the third good is positively correlated with the price of the first two goods, and a third where $\rho = -0.7$. When prices are correlated and the scope of the demand system is misspecified, a standard omitted variable bias problem arises.

Columns (III)-(V) report estimated elasticities from this smaller, misspecified demand system with $J = 2$ goods. In column (III), we find that both the own and cross elasticities can still be accurately recovered when there is no underlying correlation in prices. This suggests that, conditional on having a sufficiently flexible estimator of substitution patterns, the omission of related goods alone is not enough to generate biases in elasticity estimates. However, in columns (IV) and (V) where $\rho = \{0.7, -0.7\}$, we find clear biases in both own and cross elasticity estimates. Moreover, the sign of this bias is governed by sign of the product of ρ and the omitted cross elasticity, which is consistent with the omitted variable bias intuition. For example, we find that the estimated elasticities are biased *downwards* in column (III) where $\rho \times \mathcal{E}_{j3} = 0.7 \times -0.15 < 0$ and biased *upwards* in column (IV) where $\rho \times \mathcal{E}_{j3} = -0.7 \times -0.15 > 0$.

5 Data

Our empirical analysis uses data from one major grocery retail chain in the United States.[10] The retailer has nearly 500 stores spanning five states. Our sample consists of weekly UPC-level quantities, prices, feature promotion activity, wholesale prices, and marginal costs[11] across all departments in each store during the years 2014-2016. Ideally, we would estimate one large demand system that included products from most grocery departments. Such a system would allow us to measure all cross-category substitution and then, in some sense, define market boundaries solely from the data. However, the size of such a system would make our nonparametric approach — as well as most all structural approaches — intractable. We must therefore trade-off scope and flexibility with tractability. We do this by aggregating products to the category level and estimating a multi-category system separately for different "product groups." For example, Baking Goods is a product group with two categories (cake mix and frosting) and Refrigerated Juice is a product group with five categories (orange juice, lemonade, fruit juice, other juice, and iced tea). This unit of analysis makes it feasible to flexibly estimate cross-category substitution, while still allowing for a wider set

[9] Compiani (2020) shows that, under a different DGP with complementarity, fitting a discrete choice model which assumes substitution across goods yields substantial bias in both the own- and the cross-price elasticities.

[10] The data are provided by DecaData (http://decadata.io).

[11] Marginal cost is defined as wholesale price less any trade deals between the manufacturer and retailer.

Table 3: Product Category Descriptions

Product Group	Categories
Baking Goods	(2) Cake Mix, Frosting
Butter/Margarine/Spreads	(3) Butter, Margarine, Spreadable Butter
Canned Fish	(5) Tuna (Chunk-Light, Chunk-White, Solid-White), Sardines, Salmon
Cereal	(4) Adult, All Family, Kids, Better For You
Crackers	(3) Savory, Spray Butter, Saltines
Detergent/Softeners	(4) Laundry Detergent (Dry, Liquid), Fabric Softener (Dry, Liquid)
Hot Dogs	(3) Beef, Meat, Turkey
Jams/Jellies/Peanut Butter	(4) Peanut Butter (Regular, Organic), Jams/Jellies (Regular, Organic)
Refrigerated Juice	(5) Orange Juice, Lemonade, Fruit Juice, Other Fruit Juice, Iced Tea

of categories than typically used in market-level demand estimation.

Defining products at the category level requires us to aggregate UPC-level quantities and prices. We define quantities as the total volume (in ounces) across all UPCs in the category and define prices as revenue-weighted prices, where weights are computed for each category-store-year. We also use the same revenue weights to construct category-level promotion, wholesale price, and marginal cost variables. Estimating market demand also raises questions of price endogeneity, which we address using a battery of fixed effects and instruments. While costs serve as an ideal candidate for price instruments, there is fairly limited variation in costs over time in our sample. We therefore construct Hausman IVs — i.e., the average price of the same product in other markets (Hausman, 1996) — to instrument for price.

We construct nine different data sets from product groups that vary in the nature of product differentiation and expected substitution. A list of these nine groups and all associated categories can be found in Table 3. Groups like Butter/Margarine/Spreads, Crackers, or Hot Dogs all consist of categories that are "obvious substitutes." Other groups like Baking Goods, Detergent/Softener, and Jams/Jellies/Peanut Butter consist of categories that are "obvious complements."[12] We also include groups like Refrigerated Juice where the nature of cross-category substitution is less clear a priori. Substitution in a group like cereal, one of the classic markets for applying BLP (Nevo, 2001), is also perhaps not obvious. Households with kids, for example, may purchase multiple varieties across the adult and kids categories and so assumptions of discrete choice may be violated. Since our approach is to flexibly estimate the demand function directly, we relax standard restrictions on utility/demand and can thus let the data tell us about the nature of substitution.

[12]Some groups actually consist of both "obvious" complements and substitutes. For example, regular and organic peanut butter are likely substitutes with each other, but both complements with regular and organic jams/jellies.

6 Descriptive Evidence Across Many CPG Groups

We begin our analysis with a set of descriptive results characterizing demand and supply-side boundaries. On the demand side, our goal is to demonstrate that appreciable cross-category effects exist in our data, which then motivates the need for a flexible structural model of cross-category substitution that can be used to carry out pricing counterfactuals. To this end, we estimate a large number of log-log demand models across all nine product groups listed in Table 3. The product is defined as a category and q_{ist} represents quantity sales for category i at store s at week t. Also let $\mathcal{G}_i$ denote the product group associated with category i. For example, if i is cake mix then $\mathcal{G}_i = \{\text{cake mix, frosting}\}$ is the set of categories in Baking Goods. We follow Hitsch et al. (2019) and specify demand for good i as a function of prices and promotions for all goods within the same product group.

$$\log q_{ist} = \alpha_i + \sum_{j \in \mathcal{G}_i} \beta_{ij} \log p_{jst} + \sum_{j \in \mathcal{G}_i} \theta_{ij} d_{jst} + \phi_i(t) + \varepsilon_{ist} \tag{11}$$

Here $\log p_{jst}$ is the log price of category j, d_{jst} is the volume of feature advertising for category j, and $\phi_i(t)$ are time fixed effects (season and holiday dummmy variables). We estimate the model above separately for each category-ZIP code, thus pooling information across all stores within a ZIP code. In Table 4 we report summary statistics of the own-price and cross-price elasticities for each product group. We specifically report the mean, standard deviation, and share of positive (or negative) elasticity estimates for estimates that are significant at the 5% level. The complete distributions of own-price and cross-price elasticities, as well as the own-promotional effects are reported in Appendix B.

We find substantial heterogeneity in the sign and magnitude of cross-price elasticities. For six of the nine product groups, the average cross-category elasticity is positive. The largest effects come from the Butter/Margarine/Spreads and Crackers product groups, with average cross-elasticities of 1.20 and 1.12 respectively. Among these six groups, we still find substantial variation in the estimated cross elasticities. For example, in Refrigerated Juice the average cross elasticity is 0.22 but roughly 47% of the precisely estimated elasticities are negative. While some of the heterogeneity in these estimates can be attributed to lack of economic structure in a log-log demand system, the large mass of negatively estimated elasticities that is persistent across many groups suggest that substitution across categories may be more nuanced. The remaining three product groups (Baking Goods, Detergent/Softener, and Jams/Jellies/Peanut Butter) exhibit negative cross elasticities on average, suggesting strong complementarity across categories. This is consistent with existing work that has documented complementarities between cake mix and frosting (e.g., Manchanda et al., 1999; Ma et al., 2012), and laundry detergent and fabric softener (e.g., Song and Chintagunta, 2006; Mehta, 2007; Song and Chintagunta, 2007).

The estimated own-price elasticities are all negative, on average, for each product group. We find that demand is most elastic in the Cereal and Refrigerated Juice groups, where the average own elasticities are -2.26 and -2.07, respectively. While the share of negative own elasticities is quite high across groups, roughly 22% of these elasticities are still greater (more positive) than

Table 4: Summary of Elasticity Estimates from the Log-Log Model

	Cross-Price Elasticity				Own-Price Elasticity			
Product Group	Mean	SD	% Pos.	% Sig.	Mean	SD	% Neg.	% Sig.
Baking Goods	-1.14	0.82	8.78	40.55	-1.46	0.50	99.29	57.81
Butter/Margarine/Spreads	1.20	1.65	76.89	19.36	-1.43	1.60	79.93	79.18
Canned Fish	0.28	1.12	65.40	11.21	-0.74	0.71	91.46	44.93
Cereal	0.64	1.99	69.47	19.00	-2.26	1.15	97.14	55.00
Crackers	1.12	0.75	94.77	23.56	-1.99	0.74	99.56	83.11
Detergent/Softener	-0.50	0.86	25.08	22.40	-0.89	0.76	97.35	41.30
Hotdogs	0.28	1.29	63.61	29.86	-1.34	0.82	93.69	54.98
Jams/Jellies/Peanut Butter	-0.63	1.86	33.57	18.97	-1.80	0.79	98.39	59.45
Refrigerated Juice	0.22	2.33	52.96	25.33	-2.07	1.51	92.63	75.62

Notes: Log-log models are estimated at the category-ZIP code level. The mean, standard deviation, and share of positive (or negative) of price elasticities are conditional on estimates significant at the 5% level.

-1 indicating inelastic demand. There is also a small fraction of "incorrectly signed" estimates which are positive and statistically significant. Again, this is partly due to the lack of economic structure in the log-log system and is fairly typical when applying estimating regression models on store-level scanner data (Blattberg and George, 1991; Boatwright et al., 1999; Hitsch et al., 2019). In these regressions, we are also ignoring the potential endogeneity of prices, which would generate an upward bias in the estimated own elasticities. We address both of these limitations in our more structural, nonparametric analysis that follows, where we impose monotonicity restrictions to ensure negative own-price effects and also instrument for prices to alleviate endogeneity concerns.

An additional concern when estimating demand in CPG markets is dynamics and the storable goods demand problem. Specifically, if goods are storable, then consumers may wait for a sale and stockpile. Ignoring these consumer dynamics can in turn lead to biased estimates of price elasticities (Erdem et al., 2003; Sun et al., 2003). Hendel and Nevo (2003) propose a descriptive approach to test for the presence of stockpiling behavior using store-level data. The test is based on a regression of quantities ($\log q_{ist}$) on price ($\log p_{ist}$) as well as a variable measuring the duration since the last price promotion. The idea is that, in the presence of stockpiling, consumers will accelerate purchases and buy more during a sale. As inventories deplete weeks after a sale, quantities sold then should increase — i.e., the effect of duration on quantity sold should be positive. We also follow the empirical specifications of Hendel and Nevo (2003) and include controls for feature advertising, a duration variable for feature advertising, and a battery of fixed effects (ZIP code, season, and holiday). We estimate this regression model for each category in each product group. Estimates are reported in Appendix C. We find minimal evidence for stockpiling across the nine product groups we consider. The vast majority of estimated duration coefficients are negative and/or very close to zero and precisely estimated.

We now turn to the supply side, where our goal is to document correlations in prices across categories. Price correlations are relevant in the delineation of market boundaries to the extent that

Table 5: Correlations in Retail Prices

Product Group	Mean	SD	% Pos.
Baking Goods	0.43	0.28	91.51
Butter/Margarine/Spreads	0.19	0.33	63.84
Canned Fish	0.09	0.26	61.04
Cereal	0.51	0.41	90.76
Crackers	0.79	0.42	94.38
Detergent/Softener	0.26	0.29	80.32
Hotdogs	0.48	0.38	89.04
Jams/Jellies/Peanut Butter	0.46	0.41	84.29
Refrigerated Juice	0.05	0.36	55.14

Notes: Correlations are computed at the category-ZIP code level. The summary statistics reported above are computed separately for each product group.

they contribute to an omitted variable bias problem. That is, as discussed in Section 4, omitting a product which exhibits appreciable substitution with the "focal" good(s) is in general not enough to create a problem in estimation. Instead, problems arise when we omit goods that are related to (i.e., non-zero cross-effect) — and whose prices are correlated with — the focal good(s).

Similar to the demand-side analysis above, we estimate correlations in prices across categories (and within product groups) at the ZIP code level, and then produce summary statistics for each product group. Table 5 reports the mean, standard deviation, and share of positive correlations. The complete distributions of price correlations are shown in Appendix B. We generally find that prices are positively correlated across categories. Like the demand-side analysis, there is still some heterogeneity in the sign and magnitude of these effects. In some product groups like Baking Goods, Cereal, or Crackers, more than 90% of price correlations are positive. In other groups like Refrigerated Juice, only 55% of correlations are positive. Explaining the magnitudes of price correlations is not our primary focus, given that price correlations contribute to an omitted variable bias problem in estimation regardless of the source of the correlation. That said, it is possible that price correlations are driven by common cost shocks. We therefore include distributions of correlations in both retail prices and dollar markups in Appendix B.

Together, the descriptive analysis above highlights a few important results. First, within each broad product group, there appears to be appreciable substitution across categories. Second, the nature of substitution is nuanced and many cross-category relationships cannot be categorized as "obvious" substitutes or complements. Third, log-log models lack proper economic structure, which is evidenced by the subset of incorrectly signed own elasticities. Fourth, many categories exhibit fairly large correlations in prices which can contribute to an omitted variable bias problem in estimation. We therefore move onto a more structural model of demand which is both flexible and valid, allowing us to properly measure both the extent of cross-category substitution/complementarity as well as the cost of misspecified boundaries of competition.

7 A Structural Demand Analysis for Refrigerated Juice

In this section, we present a more structural analysis of demand using the model and nonparametric estimation approach outlined in Sections 2 and 3. For illustrative purposes, we focus on the Refrigerated Juice product group. There are a few reasons why this group is particularly attractive. First, while the orange juice category has long been used for demand estimation and pricing research (Montgomery, 1997; Chintagunta et al., 2003; Nair et al., 2005; Dubé et al., 2008, 2010), few studies have measured substitution across orange juice and other refrigerated juice categories. Quantifying these cross-category effects can help us determine whether orange juice can be treated as its own market and whether the optimal pricing function for orange juice products should consider the demand for related juices, and vice versa. Second, the nature of substitution across juice categories is complicated. This product group contains many leading CPG manufacturers with brands in multiple categories (e.g., Simply Orange and Simply Lemonade). Thus, the presence of strong brand effects or coordinated marketing efforts could lead to some form of demand spillover across categories (Smith et al., 2019). Moreover, it is not clear whether different flavors of juice should be considered strict substitutes. Our descriptive results in Section 6 showed that cross elasticities are very heterogeneous in Refrigerated Juice, with roughly half of the precisely estimated cross-price elasticities being positive and the other half negative. It is plausible that forms of brand complementarities or demand for variety leads to joint purchases spanning multiple categories. In this case, assumptions of discrete choice underlying workhorse models like BLP become less tenable. For these reasons, we view this product group as a useful laboratory to illustrate our approach.

7.1 Nonparametric Specification

We estimate a nonparametric demand system for all five categories in the Refrigerated Juice product group. Besides prices, we include the following exogenous demand shifters (x_t in the notation of Section 2): feature promotional activity, season, holiday, 5-digit ZIP code and category dummies. We use the Hausman instruments described in Section 5 as excluded IVs for prices. We impose the following five constraints from Section 3.2: (i) the linear index restriction; (ii) exchangeability between the products in the set $\mathcal{J}' = \{\text{Lemonade, Fruit Juice, Other, Iced Tea}\}$, but not across $\{\text{Orange Juice}\}$ and $\mathcal{J}'$;[13] (iii) diagonal dominance; (iv) necessary restrictions for symmetry of the Jacobian of demand with respect to prices; and (v) negative own-price effects. These restrictions lead to a flexible model which we call "NPD-Flex" (for Nonparametric Demand Model). We also estimate a second model with the five restrictions stated above as well as a positive sign constraint

[13] As discussed in Compiani (2020), exchangeability is a practically useful assumption that helps reduce the dimension of the estimation problem. We only impose exchangeability among all categories other than orange juice in order to keep the relationship between orange juice and other categories as flexible as possible. This decision is in part motivated by the fact that orange juice is a widely used category in empirical work and so we are especially interested in substitution between orange juice and other flavor categories. Note that the inclusion of category fixed effects means that the model allows for systematic unobserved differences even among the fours categories assumed to be exchangeable.

on the cross-price effects.[14] This additional sign constraint creates a nonparametric demand model for strictly substitutable goods, which we refer to as "NPD-Subst."

In summary, both NPD models are flexible models in the sense that demand is specified as a general, nonseparable function of prices, other product characteristics, and structural error terms, and this function is estimated nonparametrically. Both models are also flexible in the sense that they possess enough parameters to approximate any demand function (Diewert, 1974; Pollak and Wales, 1992). Further, the NPD-Flex model imposes no constraints on the sign of cross-price effects and so the model can accommodate both substitutable and complementary goods. In contrast, the NPD-Subst model only admits positive price effects and thus rules out any form of complementarity.

7.2 Model Comparison

Given that the NPD-Flex and NPD-Subst models imply different behavioral stories, our first task is to assess the fit of each model to the data. By comparing model fit statistics, we can get a better sense of the "bite" of the NPD-Subst model and determine whether there is any empirical support for a model allowing for cross-category complements. Each NPD model also contains a tuning parameter m which controls the complexity of the nonparametric estimator. Specifically, each unknown function is estimated using an m-degree Bernstein polynomial in each of its arguments.[15] We also want to use the data to select an optimal value of m for each model.

Our procedure for evaluating model fit and selecting tuning parameters is as follows. We first split our data into two subsamples: a *selection sample* (corresponding to 30% randomly selected weeks) and an *estimation sample* (given by the remaining 70% of the data). This initial sample-split is done to preserve the validity of any standard errors computed on elasticities.[16] To compute model fit statistics (for model selection purposes), we perform five-fold cross validation on the selection sample. Specifically, we randomly select 80% of the weeks in the selection sample, estimate the model on them and compute the root mean squared error (RMSE) on the remaining 20%. Here, RMSE refers to the square root of the mean of squared ξ errors, where the mean is across store-weeks and products. Repeating this five times and averaging yields a scalar measure of out-of-sample fit.

Table 6 reports predictive RMSEs for polynomial degrees $m \in \{2, 3\}$ for each NPD model. We find that the $m = 2$ model performs best for both NPD-Flex and NPD-Subst. Across NPD specifications, we find that NPD-Flex provides a slight improvement in fit relative to NPD-Subst. This provides some preliminary evidence that allowing for complementarities is useful, and we will see this how this plays out in the estimated elasticities reported in the next section. Lastly, because

[14] The constraint is technically a negative sign constraint on the cross-derivative of $\sigma_j(\cdot)$ with respect to the index $\delta_k(p_k, \xi_k)$. Since the index is a linear function of $-p_k$, a negative constraint with respect to the index δ_k is the same as a positive constraint with respect to the price p_k.

[15] In our system of $J = 5$ goods, the total degree of the polynomial approximation is $5m$.

[16] Estimating uncertainty in elasticity estimates with the same sample used to choose tuning parameters will lead to invalid inferences. The specific problem is that the model itself becomes stochastic when it is selected using data, and this source of uncertainty is not accounted for in standard inference procedures. This situation is now referred to as "selective inference" or "post-selection inference" and has received much attention in the literature (e.g., Belloni et al., 2014; Taylor and Tibshirani, 2015; Chernozhukov et al., 2018), particularly for high-dimensional models and regularized estimation. The most common solutions involve some form of sample splitting.

Table 6: Model Fit Statistics

	Degree of the Polynomial		Out-of-sample
Model	Each Argument (m)	Total ($5m$)	RMSE
NPD-Flex	2	10	0.020
	3	15	0.025
NPD-Subst	2	10	0.022
	3	15	0.365

the structure of NPD-Flex subsumes NPD-Subst and NPD-Flex does lead to improvements in fit, we will use the NPD-Flex model as the "ground-truth" model from which we will simulate demand for the pricing counterfactuals.

7.3 Elasticity Estimates

Table 7 reports the the full 5×5 matrix of price elasticity estimates in the refrigerated juice product group. We report estimates from both the flexible NPD model (NPD-Flex) and the NPD model imposing positive cross-price effects (NPD-Subst). We also include estimates from two additional models as a point of comparison. The first is a BLP model which, by construction, imposes substitution across all goods.[17] The second is a log-log model which imposes no economic restrictions.[18] Note that NPD and BLP are non-constant elasticity models and so elasticities vary with prices and quantities. We compute elasticities at the price-quantity values observed in the data and report medians in Table 7. We report the complete set of elasticity curves (i.e., elasticities as a function of price) for the NPD-Flex model in Appendix D.

There are a few observations to make from the elasticity estimates in Table 7. The first is with respect to the signs of the elasticities. In particular, while all own elasticities are estimated to be negative (as expected), the mix in signs of the cross elasticities looks very different across models. These differences can in part be explained by differences in each model's constraints, or lack thereof. The NPD-Flex and log-log models place no restrictions on the sign of cross-price effects while the NPD-Subst and BLP models constrain cross effects to be positive. We find that both the NPD-Flex and log-log models produce some negative (and precisely estimated) cross elasticities, suggesting that complementarity is present in this system of goods. The fact that NPD-Flex also provides superior predictive fit than NPD-Subst is also evidence that some form of complementarity exists and should be accommodated by the demand model.

The second observation is with respect to magnitudes. We find that the magnitudes of the estimated elasticities are much larger under flexible demand systems (NPD and log-log) than BLP.

[17] We estimate BLP using the pyBLP package (Conlon and Gortmaker, 2020). To translate quantities into shares, we define the market size as the maximum weekly quantity sold across both refrigerated and shelf-stable juice product groups for each store-year. Utility is specified as a linear function of price, feature promotions, and product intercepts. Random coefficients are placed on the product intercepts and price.

[18] This particular specification of the log-log model is different from the specification used in Section 6 in two ways: (i) the model is estimated using all store-weeks using ZIP code, season, and holiday fixed effects; and (ii) the model includes Hausman IVs to address the potential endogeneity of prices.

Table 7: Elasticity Estimates (Refrigerated Juice)

	%Δ in	%Δ in Price				
Model	Demand	Orange	Lemonade	Fruit	Other	Iced Tea
NPD-Flex	Orange	-7.11 (0.89)	0.09 (0.01)	0.21 (0.03)	0.28 (0.03)	0.21 (0.05)
	Lemonade	-1.56 (0.75)	-2.23 (0.24)	-0.45 (0.28)	-0.59 (0.04)	-0.43 (0.16)
	Fruit	0.32 (0.51)	0.18 (0.03)	-4.04 (0.62)	0.23 (0.10)	0.43 (0.18)
	Other	-0.08 (0.75)	0.27 (0.02)	0.93 (0.23)	-5.89 (0.63)	0.71 (0.22)
	Iced Tea	0.58 (0.53)	0.18 (0.08)	0.36 (0.14)	0.60 (0.22)	-4.23 (0.45)
NPD-Subst	Orange	-7.05 (0.31)	0.57 (0.03)	0.70 (0.03)	1.20 (0.06)	0.76 (0.03)
	Lemonade	0.67 (0.03)	-3.24 (0.03)	0.30 (0.00)	0.53 (0.01)	0.34 (0.00)
	Fruit	1.81 (0.08)	0.68 (0.01)	-3.76 (0.05)	1.44 (0.02)	0.92 (0.01)
	Other	0.91 (0.08)	0.62 (0.01)	1.10 (0.01)	-6.07 (0.10)	0.92 (0.00)
	Iced Tea	1.84 (0.07)	0.68 (0.01)	0.84 (0.01)	1.45 (0.01)	-3.73 (0.02)
BLP	Orange	-1.96 (0.00)	0.02 (0.00)	0.12 (0.01)	0.01 (0.00)	0.06 (0.00)
	Lemonade	0.18 (0.01)	-2.29 (0.01)	0.13 (0.01)	0.02 (0.00)	0.07 (0.00)
	Fruit	0.15 (0.01)	0.02 (0.00)	-0.81 (0.00)	0.01 (0.00)	0.06 (0.00)
	Other	0.18 (0.01)	0.02 (0.00)	0.13 (0.01)	-1.95 (0.01)	0.07 (0.00)
	Iced Tea	0.18 (0.01)	0.03 (0.00)	0.14 (0.01)	0.02 (0.00)	-0.96 (0.00)
Log-Log	Orange	-2.58 (0.03)	0.10 (0.01)	0.10 (0.03)	0.01 (0.01)	0.19 (0.02)
	Lemonade	3.12 (0.07)	-1.92 (0.01)	-1.05 (0.07)	-0.55 (0.03)	-0.20 (0.04)
	Fruit	-0.94 (0.04)	0.06 (0.01)	-2.14 (0.03)	0.10 (0.01)	0.04 (0.02)
	Other	1.44 (0.06)	0.05 (0.01)	0.15 (0.06)	-1.64 (0.02)	0.22 (0.04)
	Iced Tea	-0.12 (0.06)	-0.04 (0.01)	0.36 (0.05)	0.00 (0.02)	-1.70 (0.03)

Notes: (1) All models include ZIP code, season, and holiday fixed effects and Hausman IVs. (2) In the NPD and BLP models, bootstrapped standard errors are computed using 100 bootstrap replicates.

In the NPD-Flex model, the largest cross elasticities are -1.56 and 0.93; in the log-log model, the largest cross elasticities are -1.05 and 3.12; in the NPD-Subst model, the largest cross elasticity is 1.84. In contrast, the largest cross elasticity in BLP is 0.18. The attenuation of BLP elasticities is perhaps not too surprising. Logit models impose very strong restrictions on cross effects (Berry, 1994), and while BLP can in principle admit more flexible substitution patterns, this flexibility comes solely through a model of preference heterogeneity for a subset of product characteristic coefficients. In practice, it can be difficult to include a rich enough set of characteristics to allow BLP elasticities to meaningfully deviate from logit elasticities. Indeed, even in our BLP model which places random coefficients on product intercepts and price, we still find very "logit-like" elasticity estimates.

We also find that demand is estimated to be more elastic under the NPD models relative to both BLP and log-log models. Differences in the own elasticities can in part be explained by differences in the cross elasticities, as discussed above. In microfounded models, the magnitude of the own effect is tied to the magnitude of cross-effects through properties like Cournot aggregation or diagonal dominance. Therefore, the BLP own elasticities should be much smaller in magnitude if the cross

elasticities are also small, and NPD own elasticities should be much larger if the cross elasticities are also large. Even though the NPD and log-log models produce cross elasticities that are similar in magnitude, the NPD own elasticities are 2-3 times larger. This is again due to the log-log model's lack of economic structure: own effects are not tied to cross effects which is at odds with properties of a valid demand system and can ultimately lead to issues in subsequent pricing conterfactuals (Anderson and Vilcassim, 2001). In contrast, the underlying structure of NPD models imposes an "accounting rule" whereby own effects must get larger to balance off larger cross effects.

Finally, we note that even though the NPD elasticities are substantially larger in magnitude than BLP elasticities, the degree of substitution implied by these elasticities is similar. For example, we can also measure substitution using the diversion ratio, which is the ratio of the cross elasticity over the own elasticity, scaled by market shares (Conlon and Mortimer, 2021). If we consider the orange juice demand equation (i.e., the top row of the elasticity matrix) and compute the diversion ratio of orange juice demand with respect to lemonade, fruit juice, other juice, and iced tea, we find that both NPD-Flex and BLP models yield ratios in the range of (0.01, 0.06). In other words, the own and cross NPD elasticities scale up roughly by the same constant relative to BLP elasticities. Since the degree of substitution is ultimately a function of both the own and the cross elasticities, larger elasticities on their own need not imply unreasonable measures of substitution.

7.4 Demand Boundaries and Omitted Variable Bias

The Monte Carlo simulations in Section 4 demonstrated biases in the NPD own and cross-price elasticities in situations where a product is omitted from the demand system and that omitted product's price is correlated with prices of goods inside the system. We now investigate the nature of the bias in elasticity estimates when omitting goods from the refrigerated juice demand system. If demand was specified as a linear function of prices, then the sign and magnitude of the bias would follow directly from the omitted variable bias formula for linear models. For example, in a linear model $q = \alpha p + \beta\tilde{p} + \varepsilon$, the omitted variable bias in the OLS estimator $\hat{\alpha}$ when omitting $\tilde{p}$ is

$$\text{Bias}(\hat{\alpha}) = \frac{\text{Cov}(p, \tilde{p})}{\text{Var}(p)}\beta = \frac{\text{SD}(\tilde{p})}{\text{SD}(p)}\text{Corr}(p, \tilde{p})\beta. \tag{12}$$

Therefore, the sign of the bias is determined by product of the correlation between included and omitted variables and the coefficient on the omitted variables (as shown in Section 4). The challenge is that in the NPD model, demand is specified as a highly nonlinear function of prices and error terms. We also estimate the *inverse* demand system via nonparametric IV regression, and so it is not clear how the biases with respect to $\sigma^{-1}(\cdot)$ translate to biases with respect to $\sigma(\cdot)$. Although deriving the appropriate omitted variable bias formula for NPD models is beyond the scope of this paper, we show that the intuition from linear models — i.e., considering both demand-side substitution and supply-side price correlations — is still useful in our empirical setting.

We start by examining the bias for one elasticity with only one omitted good. After this simple case is made clear, then we will generalize to all elasticities with multiple omitted goods. Consider

the orange juice demand equation where the estimated own elasticity is -7.11 (from Table 7). Now consider a smaller, misspecified model which omits the index for the final good, iced tea.

$$\sigma_{\text{orange}}(\delta_{\text{orange}}, \delta_{\text{lemonade}}, \delta_{\text{fruit}}, \delta_{\text{other}}, \delta_{\text{icedtea}}) \rightarrow \sigma^{(4)}_{\text{orange}}(\delta_{\text{orange}}, \delta_{\text{lemonade}}, \delta_{\text{fruit}}, \delta_{\text{other}})$$

The key "ingredients" for signing the resulting bias on the own elasticity of orange juice are: (i) the correlation between the price of orange and price of iced tea, which is -0.13; and (ii) the cross-price elasticity corresponding to the price of iced tea and demand of orange juice, which is 0.21 (from Table 7). Based on the linear omitted variable bias formula, we would therefore predict a downward bias (i.e., more negative) in the estimated own-price elasticity of this misspecified system. After estimating the $J = 4$ model with iced tea omitted, we indeed find that the estimated elasticity of orange juice is -7.81, which is directionally consistent with our prediction.

There are many ways in which the findings above can be generalized. First, we could investigate the bias of the cross elasticities in the orange juice demand equation. Second, we could consider omitting multiple goods in the orange juice demand equation, not just iced tea. Third, we could investigate the bias each of the five total demand equations rather than focusing on orange juice alone. In what follows, we explore each one of these dimensions. In order to keep computation manageable, we focus on one specific sequence of smaller, misspecified demand systems:

$$\begin{aligned} &\sigma^{(4)}(\delta_{\text{orange}}, \delta_{\text{lemonade}}, \delta_{\text{fruit}}, \delta_{\text{other}}) \\ &\sigma^{(3)}(\delta_{\text{orange}}, \delta_{\text{lemonade}}, \delta_{\text{fruit}}) \\ &\sigma^{(2)}(\delta_{\text{orange}}, \delta_{\text{lemonade}}) \end{aligned} \tag{13}$$

where $\sigma^{(4)}(\cdot)$ is a 4-good system with $4^2 = 16$ elasticities, $\sigma^{(3)}(\cdot)$ is a 3-good system with $3^2 = 9$ elasticities, and $\sigma^{(2)}(\cdot)$ is a 2-good system with $2^2 = 4$ elasticities.

From each of the misspecified models above, we compute $\hat{\mathcal{E}}^{\mathcal{K}}_{ij}$, which is the estimated elasticity of the demand for i with respect to the price of j when the subset of goods $\mathcal{K}$ is omitted from the full five-good system. For each store-week in the data, we compute the following for each (i, j) product pair.

(i) The "observed" bias, which is the difference between the estimated (i, j) elasticity under the full model $\hat{\mathcal{E}}_{ij}$ and the estimated (i, j) elasticity under a smaller model $\hat{\mathcal{E}}^{\mathcal{K}}_{ij}$ for $|\mathcal{K}| \in \{2, 3, 4\}$ and $\mathcal{K}$ as defined by the demand systems in (13).

(ii) The "predicted" bias, which is the product of the omitted elasticity $\hat{\mathcal{E}}_{ik}$ (from the full model) and the correlation between the prices of goods i and $k \in \mathcal{K}$. When there is more than one omitted good (i.e., $|\mathcal{K}| > 1$), we average the predicted bias across k.

By comparing the "observed" bias and "predicted" bias, we can see how well the intution from (12) (using demand-side and supply-side factors) does in capturing the sign of the observed omitted variable bias.

In the context of our motivating example of orange juice under the $\sigma^{(4)}(\cdot)$ misspecified model,

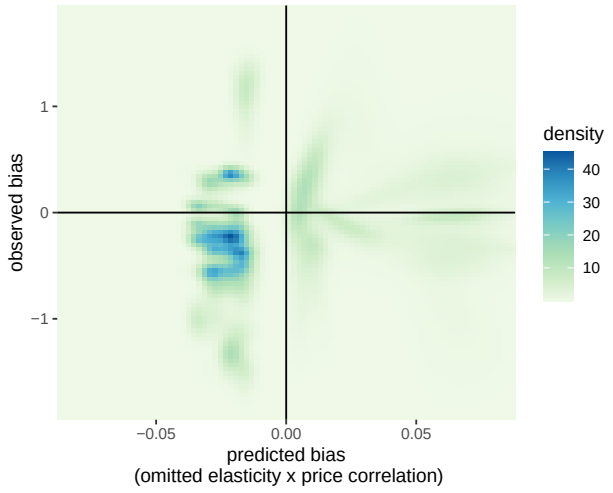

Figure 1: Heatmap of Observed Bias vs. Predicted Bias

we find that the sign of the "observed" bias and "predicted" bias agree in all store-weeks. The same is true of the cross-price elasticities from the orange juice demand equation. When we include all three types of misspecified models, we find that the biases agree in 78% of store-weeks. To generalize beyond orange juice, we produce a heatmap of the observed vs predicted biases across all store-weeks, all three misspecified models, and all demand equations in Figure 1. This plot is like a scatterplot where the points are assigned into bins and the color of a bin indicates its density (dark colors imply that the bin is filled with more observations). The heatmap is useful because it allows us to see where points are concentrated in the "observed" and "predicted" space. Of particular interest is whether points fall in the bottom-left and top-right quadrants where the sign of "predicted" matches the sign of "observed." If the points only concentrated on these two quadrants, then predictions of the sign of the bias based on (12) would always be correct.[19] We find that the largest mass of points (42%) is in the bottom-left quadrant and a smaller mass (21%) is in the top-right. So together, the sign of the "observed" bias and "predicted" bias agree in 63% of all store-weeks.

The results above suggest that the (albeit naive) predictions using the standard omitted variable bias formula can still be useful in understanding the sign of the bias when we omit relevant goods from a demand system. We specifically find that omitted variable bias predictions match the sign of the observed bias in the majority of cases, while acknowledging that the (12) is an oversimplified form of the bias present in our setting that is based on nonparametric IV estimators

[19]It is also worth noting that we are only interested in the location of the points in these four quadrants and not the exact correlation of "observed" vs. "predicted" because the magnitude of our predictions could be inaccurate. Specifically, we are not scaling predictions by the standard deviations of prices shown in (12), we are not accounting for the effect of instruments on the omitted variable bias formula, we are not accounting for the effect of omitted promotion variables, and we are using price elasticities instead of price effects. We find that the estimated promotion coefficients in the NPD model are an order of magnitude smaller than the estimated price coefficients, and so the role of promotion effects should be negligible when studying omitted variable bias. These decisions will affect the magnitude of the predicted bias (and thus the correlation between "observed" vs. "predicted") but not the sign of the predicted bias.

of an inverse demand system. At a high level, our results demonstrate that when we want to understand whether/how elasticity estimates will be affected by the omission of possibly related goods, we should consider both the demand-side substitution between those goods and also the correlation of their prices. In fact, we find that if we make predictions of the bias using only the sign of the omitted elasticity and not the price correlation, then the agreement between the sign of the "observed" and "predicted" bias drops from 63% to 32%.

7.5 Pricing Counterfactuals

Our final objective is to quantify the costs of misspecified boundaries of product competition through the lens of optimal retailer pricing. We focus on a pricing counterfactual because the first-order conditions associated with the retailer's decision problem directly depend on derivatives of the demand function, thus illuminating consequences of misspecification. Specifically, first consider the case of "joint" pricing where the retailer sets prices to maximize total profits across all juice categories.[20]

$$\max_{p_1,\ldots,p_J} \left[\Pi = \sum_{j=1}^{J} \sigma_j(\mathbf{p})(p_j - c_j) \right] \tag{14}$$

Here c_j denotes the marginal cost of category j reported in our data and, with a slight abuse of notation, we use $\sigma_j(\mathbf{p})$ to denote the demand for category j as a function of prices keeping all other demand shifters fixed. The retailer's first-order conditions associated with pricing category j are:

$$\sum_{i=1}^{J} \frac{\partial \sigma_i(\mathbf{p})}{\partial p_j}(p_i - c_i) + \sigma_j(\mathbf{p}) = 0. \tag{15}$$

and so optimal prices crucially depend on a flexible system in order to accurately capture price derivatives $\partial \sigma_i(\mathbf{p})/\partial p_j$.

The first form of misspecification we consider is misspecification of the *demand model*, whereby the retailer gets the demand function wrong and assumes that all juice categories are substitutes. This in turn implies that the price derivatives in (15) will be misspecified. For this, we use the NPD-Flex model as the true model and the NPD-Subst model as the misspecified model. Note that any differences we find in profits will likely be a conservative estimate of the costs of misspecification given that the misspecified model is still a flexible NPD system as opposed to a more restrictive parametric model.

The second form of misspecification we consider is in the boundaries of the retailer's decision problem. In contrast to the correctly specified joint pricing policy in (14), consider a misspecified "isolated" pricing policy where the retailer solves for each p_j^* separately:[21]

$$\max_{p_j} \Big[\Pi_j = \sigma_j(\mathbf{p})(p_j - c_j) \Big] \tag{16}$$

[20] We focus on optimization of the baseline prices and keep any discounts — which are observed in the data — fixed throughout the counterfactual exercises.

[21] We assume that the other prices are fixed at the values observed in the data.

with associated first-order conditions:

$$\frac{\partial \sigma_j(\mathbf{p})}{\partial p_j}(p_j - c_j) + \sigma_j(\mathbf{p}) = 0. \tag{17}$$

This decision problem is misspecified in the sense that the first-order conditions only depend on the derivatives of σ_j and not other related goods in the system. The consequence of isolated category pricing is that the retailer will not be able to internalize the externalities arising from cross-category substitution (Thomassen et al., 2017). For example, the NPD-Flex elasticity estimates suggest that a decrease in the price of orange juice will *increase* demand for lemonade, and this complementarity creates a *positive* externality. If the retailer prices jointly across all categories, then they can internalizes this externality and the price of orange juice should decrease (relative to a case of isolated pricing). Similarly, our estimates also suggest that a decrease in the price of lemonade will *decrease* demand for orange juice. This substitution creates a *negative* externality and lemonade prices would increase if internalized.

The two sources of misspecification discussed above lead to four different configurations of optimal prices: (NPD-Flex, NPD-Subst) x (isolated pricing, joint pricing). We also add a fifth configuration where demand is estimated separately in each category (NPD-Indep). Note that in this scenario, isolated and joint pricing would yield the same pricing policy since the demand specification does not allow for cross-category substitution. We solve for optimal prices at the store-week level and report the average prices and profits for a random sample of 1,000 store-weeks in Table 8. First, a few observations about optimal prices. We find that prices are on average higher under joint pricing, where the magnitude of the difference depends on the degree of substitution estimated by the NPD model. For example, we estimate lemonade to be a strict substitute with all other goods. This negative externality should increase prices when internalized, which is what we find in both NPD-Flex and NPD-Subst specifications. In contrast, consider orange juice where we find that average prices under NPD-Flex are the same irrespective of whether prices are set jointly or not. One explanation is that the NPD-Flex model estimates orange juice to be complements with lemonade and other juices and substitutes with fruit juice and iced tea. These elasticities would induce positive and negative externalities, respectively, which may offset in the pricing problem.

Moving to profits, we find that the "correctly" specified scenario of NPD-Flex plus joint pricing does in fact generate the most profitable pricing policy, with the expected per store-week profits of \$319. The differences in profits under the alternative "misspecified" demand models or decision problems allow us to quantify the cost of misspecification. For example, profits are 3.8% lower when demand is misspecified but prices are solved jointly. Profits are 11.6% lower when demand is correctly specified but prices are solved separately in each category. Profits are 13.2% lower when both the model is misspecified and prices are solved separately in each category. Finally, profits are 14.4% lower when we remove cross-category effects altogether and solve for prices independently in each category. These results suggest that the boundaries of the decision problem are a first-order issue, given that the worst case scenario under joint pricing is still better than than the best case scenario under isolated pricing. An important caveat is that our misspecified model is still a flexible

Table 8: Profit Table

			Isolated Pricing			Joint Pricing	
		Observed	NPD-Indep	NPD-Subst	NPD-Flex	NPD-Subst	NPD-Flex
Prices	Orange	0.082	0.081	0.082	0.082	0.085	0.082
	Lemonade	0.080	0.080	0.081	0.089	0.122	0.121
	Fruit	0.036	0.036	0.036	0.036	0.038	0.037
	Other	0.070	0.063	0.065	0.065	0.074	0.073
	Iced Tea	0.037	0.036	0.037	0.037	0.041	0.039
Profits	Orange	78	89	90	91	94	101
	Lemonade	20	25	24	26	11	14
	Fruit	111	106	108	110	140	138
	Other	10	12	13	13	11	11
	Iced Tea	41	41	42	43	51	55
	Total	250	273	277	282	307	319

Notes: Prices are in dollars per ounce. Profits are dollar averages at the week-store level.

NPD model and so the profit losses we estimate are likely a lower bound to true losses from pricing with more restrictive models.

Finally, we note that the observed prices somewhat deviate from the optimal prices under the NPD-Flex model and, as a result, the observed profits are lower. This could occur because the retailer fails to correctly specify the demand model and/or the pricing problem, but it could also be due to the fact that our analysis only considers five out of the many categories of products sold. In other words, the retailer could be internalizing some externalities between these five categories and products outside our demand system. Estimating larger flexible systems could help shed further light on this and we hope the analysis in this paper serves as a useful first step in that direction.

8 Conclusion

This paper quantifies the costs of misspecified boundaries of competition on optimal retailer pricing. We explore misspecification in both the estimation problem and the decision problem. Our empirical analysis uses store-level transaction data from a large grocery retailer. We first provide a set of descriptive results across nine diverse product groups to demonstrate that cross-category effects "exist." We then focus on the Refrigerated Juice product group and estimate demand non-parametrically for all five categories. We also estimate a sequence of smaller models in which we iteratively omit a good from the demand system. We find that the direction of the estimated bias from these smaller misspecified models is largely consistent with standard omitted variable bias results. This suggests that when searching for relevant goods we should consider both demand-side substitution as well as correlations in retail prices. Finally, we solve for optimal prices and find that both sources of misspecification lead to profit losses in the range of 4-14%, with misspecification of the decision problem accounting for larger losses relative to misspecification of the demand system.

References

Albuquerque, P. and Bronnenberg, B. J. (2009). Estimating demand heterogeneity using aggregated data: An application to the frozen pizza category. *Marketing Science*, 28(2):356–372.

Anderson, E. and Vilcassim, N. (2001). Structural demand models for retailer category pricing. *Working Paper*.

Basuroy, Suman, M. K. M. and Walters, R. G. (2001). The impact of category management on retailer prices and performance: Theory and evidence. *Journal of Marketing*, 65(4):16–32.

Belloni, A., Chernozhukov, V., and Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, 81(2):608–650.

Berry, S., Gandhi, A., and Haile, P. (2013). Connected substitutes and invertibility of demand. *Econometrica*, 81(5):2087–2111.

Berry, S. T. (1994). Estimating discrete-choice models of product differentiation. *The RAND Journal of Economics*, 25(2):242–262.

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

Berry, S. T. and Haile, P. A. (2021). Foundations of Demand Estimation. In Ho, K., Hortaçsu, A., and Lizzeri, A., editors, *Handbook of Industrial Organization*, Handbook of Industrial Organization.

Berry, S. T., Khwaja, A., Kumar, V., Musalem, A., Wilbur, K. C., Allenby, G. M., Anand, B., Chintagunta, P. K., Hanemann, W. M., Jeziorski, P., and Mele, A. (2014). Structural models of complementary choices. *Marketing Letters*, 25(3):245–256.

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

Besanko, D., Dubé, J.-P., and Gupta, S. (2003). Competitive price discrimination strategies in a vertical channel using aggregate retail data. *Management Science*, 49(9):1121–1138.

Blattberg, R. C. and George, E. I. (1991). Shrinkage estimation of price and promotional elasticities: Seemingly unrelated equations. *Journal of the American Statistical Association*, 86(414):304–315.

Blundell, R., Horowitz, J., and Parey, M. (2017). Nonparametric estimation of a nonseparable demand function under the slutsky inequality restriction. *Review of Economics and Statistics*, 99(2):291–304.

Boatwright, P., McCulloch, R., and Rossi, P. (1999). Account-level modeling for trade promotion: An application of a constrained parameter hierarchical model. *Journal of the American Statistical Association*, 94(448):1063–1073.

Bodapati, A. V. and Gupta, S. (2004). The recoverability of segmentation structure from store-level aggregate data. *Journal of Marketing Research*, 41(3):351–364.

Chen, F., Liu, X., Proserpio, D., and Troncoso, I. (2020). Product2Vec: Understanding product-level competition using representation learning. *Working Paper*.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68.

Chintagunta, P. K., Dubé, J. P., and Singh, V. (2003). Balancing profitability and customer welfare in a supermarket chain. *Quantitative Marketing and Economics*, 1(1):111–147.

Chintagunta, P. K. and Nair, H. S. (2011). Structural workshop paper—discrete-choice models of consumer demand in marketing. *Marketing Science*, 30(6):977–996.

Compiani, G. (2020). Market counterfactuals and the specification of multi-product demand: A nonparametric approach. *Working Paper*.

Conlon, C. and Gortmaker, J. (2020). Best practices for differentiated products demand estimation with PyBLP. *The RAND Journal of Economics*, 51(4):1108–1161.

Conlon, C. and Mortimer, J. H. (2021). Empirical properties of diversion ratios. *The RAND Journal of Economics, Forthcoming*.

Deaton, A. and Muellbauer, J. (1980). An almost ideal demand system. *American Economic Review*, 70(3):312–326.

Diewert, W. E. (1974). Applications of duality theory. In *Frontiers of Quantitative Economics*. North-Holland Publishing Company, Amsterdam.

Donnelly, R., Ruiz, F. R., Blei, D., and Athey, S. (2019). Counterfactual inference for consumer choice across many product categories. *Working Paper*.

Draganska, M. and Jain, D. C. (2006). Consumer preferences and product-line pricing strategies: An empirical analysis. *Marketing Science*, 25(2):164–174.

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

Dubé, J. P., Hitsch, G. J., and Rossi, P. E. (2010). State dependence and alternative explanations for consumer inertia. *The RAND Journal of Economics*, 41(3):417–445.

Dubé, J. P., Hitsch, G. J., Rossi, P. E., and Vitorino, M. A. (2008). Category pricing with state-dependent utility. *Marketing Science*, 27(3):417–429.

Dunker, F., Hoderlein, S., and Kaido, H. (2017). Nonparametric identification of random coefficients in endogenous and heterogeneous aggregate demand models. *Cemmap Working Paper CWP11/17*.

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

Ershov, D., Laliberté, J.-W., Marcoux, M., and Orr, S. (2021). Estimating complementarity with large choice sets: An application to mergers. *Working Paper*.

Fosgerau, M., Monardo, J., and de Palma, A. (2021). The inverse product differentiation logit model. *Working Paper*.

Fox, J. T., Kim, K. i., and Yang, C. (2016). A simple nonparametric approach to estimating the distribution of random coefficients in structural models. *Journal of Econometrics*, 195(2):236–254.

Gabel, S. and Timoshenko, A. (2021). Cross-category product choice: A scalable deep-learning model. *Management Science, Forthcoming*.

Gal, S. (2008). *Shape-Preserving Approximation by Real and Complex Polynomials*. Birkhauser, Boston.

Handel, B. R., Misra, K., and Roberts, J. W. (2013). Robust firm pricing with panel data. *Journal of Econometrics*, 174(2):165–185.

Hausman, J. A. (1996). The economics of new goods. In Bresnahan, T. F. and Gordon, R. J., editors, *Valuation of new goods under perfect and imperfect competition*, pages 207–248. Univeristy of Chicago Press.

Hendel, I. and Nevo, A. (2003). The post-promotion dip puzzle: What do the data have to say? *Quantitative Marketing and Economics*, 1(4):409–424.

Hitsch, G. J., Hortacsu, A., and Lin, X. (2019). Prices and promotions in U.S. retail markets: Evidence from big data. *NBER Working Paper 26306*.

Iaria, A. and Wang, A. (2020). Identification and estimation of demand for bundles. *Working Paper*.

Kadiyali, V., Chintagunta, P., and Vilcassim, N. (2000). Manufacturer-Retailer Channel Interactions and Implications for Channel Power: An Empirical Investigation of Pricing in a Local Market. *Marketing Science*, 19(2):127–148.

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

Kumar, M., Eckles, D., and Aral, S. (2020). Scalable bundling via dense product embeddings. *Working Paper*.

Lee, S., Kim, J., and Allenby, G. M. (2013). A direct utility model for asymmetric complements. *Marketing Science*, 32(3):454–470.

Lewbel, A. and Nesheim, L. (2019). Sparse demand systems: corners and complements. *Cemmap Working Paper CWP45/19*.

Ma, Y., Seetharaman, P., and Narasimhan, C. (2012). Modeling dependencies in brand choice outcomes across complementary categories. *Journal of Retailing*, 88(1):47–62.

Manchanda, P., Ansari, A., and Gupta, S. (1999). The "shopping basket": A model for multicategory purchase incidence decisions. *Marketing Science*, 18(2):95–114.

Mas-Colell, A., Whinston, M. D., and Green, J. R. (1995). *Microeconomic Theory*. Oxford University Press, Inc.

Mehta, N. (2007). Investigating consumers' purchase incidence and brand choice decisions across multiple product categories: A theoretical and empirical analysis. *Marketing Science*, 26(2):196–217.

Monardo, J. (2021). Measuring substitution patterns with a flexible demand model. *Working Paper*.

Montgomery, A. L. (1997). Creating micro-marketing pricing strategies using supermarket scanner data. *Marketing Science*, 16(4):315–337.

Nair, H. S., Dubé, J. P., and Chintagunta, P. (2005). Accounting for primary and secondary demand effects with aggregate data. *Marketing Science*, 24(3):444–460.

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

Pollak, R. A. and Wales, T. J. (1992). *Demand System Specification and Estimation*. Oxford University Press, Inc.

Ruiz, F. J. R., Athey, S., and Blei, D. M. (2020). SHOPPER: A probabilistic model of consumer choice with substitutes and complements. *The Annals of Applied Statistics*, 14(1):1–27.

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

Smith, H. and Thomassen, Ø. (2012). Multi-category demand and supermarket pricing. *International Journal of Industrial Organization*, 30(3):309–314.

Song, I. and Chintagunta, P. K. (2006). Measuring cross-category price effects with aggregate store data. *Management Science*, 52(10):1594–1609.

Song, I. and Chintagunta, P. K. (2007). A discrete-continuous model for multicategory purchase behavior of households. *Journal of Marketing Research*, 44(4):595–612.

Sudhir, K. (2001). Structural Analysis of Manufacturer Pricing in the Presence of a Strategic Retailer. *Marketing Science*, 20(3):244–264.

Sun, B., Neslin, S. A., and Srinivasan, K. (2003). Measuring the impact of promotions on brand switching when consumers are forward looking. *Journal of Marketing Research*, 40(4):389–405.

Taylor, J. and Tibshirani, R. J. (2015). Statistical learning and selective inference. *Proceedings of the National Academy of Sciences*, 112(25):7629–7634.

Thomassen, Ø., Smith, H., Seiler, S., and Schiraldi, P. (2017). Multi-category competition and market power: A model of supermarket pricing. *American Economic Review*, 107(8):2308–2351.

Vilcassim, N. J. and Chintagunta, P. K. (1995). Investigating retailer product category pricing from household scanner panel data. *Journal of Retailing*, 71(2):103–128.

Wang, A. (2021a). A BLP demand model of product-level market shares with complementarity. *Working Paper*.

Wang, A. (2021b). Sieve BLP: A semi-nonparametric model of demand for differentiated products. *Working Paper*.

Wang, J. and Huang, Y. (2019). Non-parametric estimation of price elasticities: A heterogeneous treatment effect approach. *Working Paper*.

APPENDIX

A Additional Simulation Results

We present additional results from the simulation studies in Section 4 with $J = 4$ and $J = 5$. In each case, the true median elasticities are:

$$\mathcal{E}_{jk} = \frac{\partial q_j / q_j}{\partial p_k / p_k} = \begin{cases} -1 & \text{if } j = k \\ -0.15 & \text{if } j = J \text{ or } k = J \\ 0.15 & \text{otherwise.} \end{cases}$$

Below we report the median elasticities across 100 data replicates. Standard errors are given in parentheses.

Table 9: Estimated Elasticities Based on Simulated Data

	%Δ in	% Δ in Price				
J	Demand	1	2	3	4	5
4	1	-0.99 (0.030)	0.14 (0.004)	0.14 (0.005)	-0.15 (0.006)	
	2	0.14 (0.005)	-0.98 (0.026)	0.14 (0.006)	-0.15 (0.006)	
	3	0.14 (0.006)	0.14 (0.005)	-0.99 (0.034)	-0.15 (0.006)	
	4	-0.16 (0.007)	-0.16 (0.005)	-0.16 (0.007)	-1.08 (0.040)	
5	1	-0.99 (0.037)	0.15 (0.006)	0.15 (0.007)	0.15 (0.006)	-0.17 (0.008)
	2	0.15 (0.006)	-0.99 (0.036)	0.15 (0.006)	0.15 (0.006)	-0.16 (0.009)
	3	0.14 (0.006)	0.14 (0.006)	-1.00 (0.036)	0.14 (0.006)	-0.15 (0.007)
	4	0.14 (0.005)	0.14 (0.005)	0.14 (0.005)	-1.00 (0.035)	-0.15 (0.009)
	5	-0.16 (0.007)	-0.16 (0.006)	-0.16 (0.007)	-0.16 (0.006)	-1.13 (0.052)

B Additional Descriptive Results

Figure B1: Price Elasticities and Promotion Effects

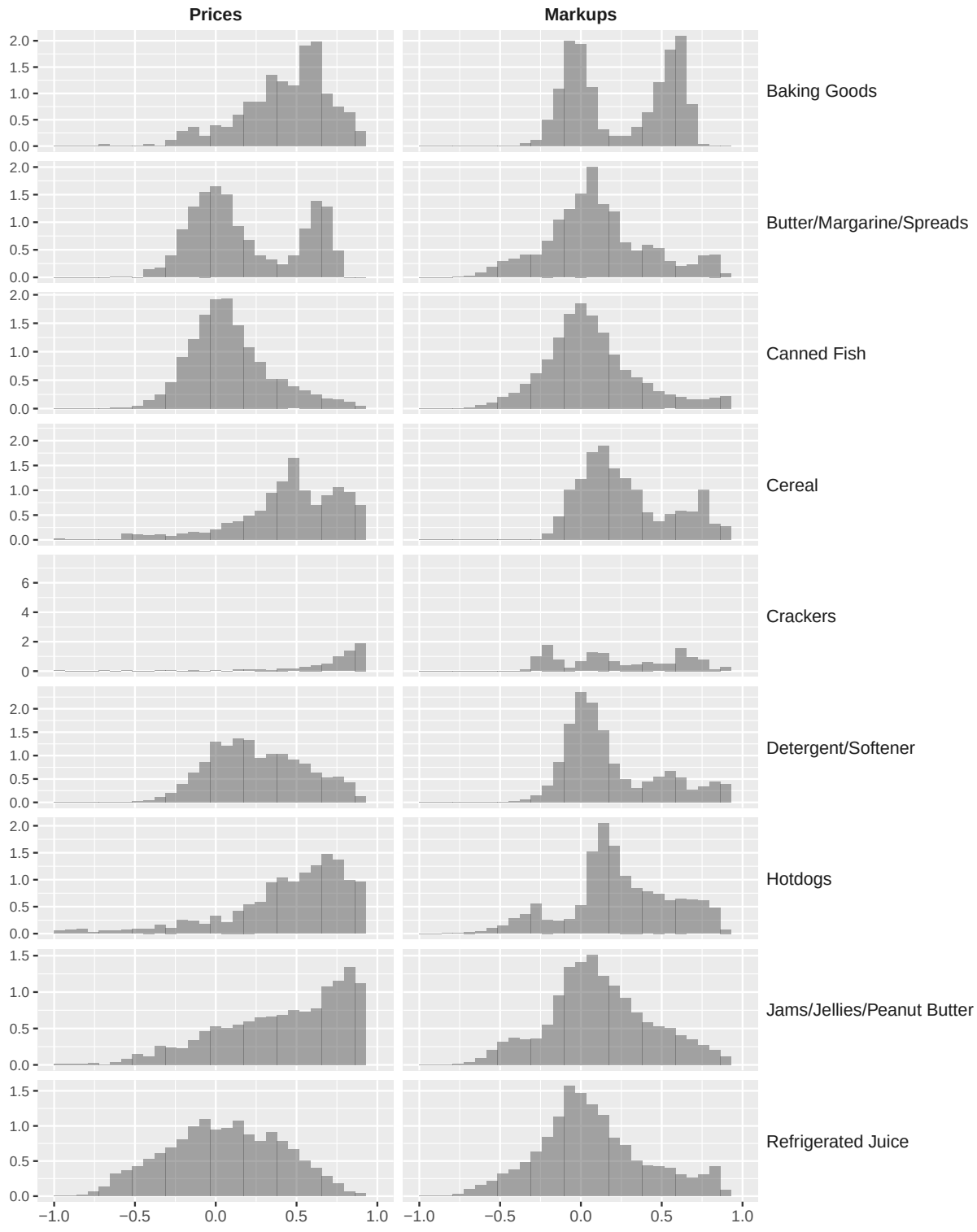

Figure B2: Correlations in Retail Prices and Dollar Markups

C Testing for Consumer Dynamics and Stockpiling Behavior

		Coefficient			
Product Group	Category	log(Price)	Duration Price	Promotion	Duration Promotion
Baking Goods	FROST	-1.15 (0.038)	0.01 (0.001)	0.33 (0.011)	-0.01 (0.004)
	MIX	-0.92 (0.042)	0.02 (0.005)	0.87 (0.021)	-0.13 (0.010)
Butter/Margarine/Spreads	BTR	-2.14 (0.029)	-0.03 (0.001)	1.37 (0.021)	0.07 (0.002)
	BTRSPRD	-1.73 (0.062)	0.00 (0.000)	0.54 (0.029)	0.01 (0.001)
	MRGRN	0.46 (0.210)	-0.01 (0.002)	1.74 (0.087)	0.04 (0.001)
Canned Fish	CHKLGHT	-0.56 (0.039)	0.02 (0.002)	1.33 (0.017)	-0.02 (0.004)
	CHKWHT	-0.75 (0.035)	-0.01 (0.002)	1.44 (0.027)	0.08 (0.010)
	SALMON	-0.24 (0.010)	-0.01 (0.001)	0.77 (0.034)	-0.04 (0.008)
	SARDINE	0.12 (0.023)	-0.01 (0.000)	0.62 (0.022)	-0.01 (0.008)
	SLDWHT	-0.55 (0.018)	-0.00 (0.003)	0.71 (0.019)	0.04 (0.005)
Cereal	ADULT	-1.54 (0.244)	0.06 (0.002)	1.30 (0.117)	-0.34 (0.010)
	BFY	-0.12 (0.012)	0.00 (0.002)	0.70 (0.012)	0.07 (0.112)
	FAMILY	-1.13 (0.037)	-0.08 (0.007)	1.46 (0.028)	-0.11 (0.240)
	KIDS	-0.30 (0.029)	0.11 (0.003)	2.52 (0.032)	0.15 (0.074)
Crackers	SALTINE	-0.47 (0.039)	0.01 (0.001)	0.56 (0.014)	0.09 (0.008)
	SAVORY	-1.69 (0.054)	0.42 (0.011)	0.39 (0.016)	
	SPRYBTR	-1.52 (0.025)	0.02 (0.003)	0.48 (0.008)	-0.30 (0.054)
Detergent/Softener	DETDRY	-0.39 (0.038)	-0.00 (0.000)		0.02 (0.007)
	DETLQD	-0.24 (0.008)	-0.00 (0.000)		0.09 (0.003)
	SFTDRY	-0.91 (0.044)	-0.01 (0.002)	0.31 (0.009)	0.04 (0.006)
	SFTLQD	-0.98 (0.049)	-0.09 (0.002)	0.49 (0.019)	0.14 (0.006)
Hotdogs	BEEF	-0.63 (0.044)	-0.02 (0.002)	1.48 (0.017)	-0.12 (0.007)
	MEAT	-1.11 (0.094)	0.00 (0.004)	1.38 (0.045)	-0.05 (0.009)
	TRKY	-0.67 (0.053)	-0.02 (0.002)	1.10 (0.027)	-0.13 (0.012)
Jams/Jellies/Peanut Butter	JAM	-1.37 (0.072)	-0.08 (0.002)	0.66 (0.022)	-0.30 (0.017)
	JAMORG	-1.17 (0.060)	-0.00 (0.001)	0.38 (0.019)	-0.08 (0.052)
	PB	-1.72 (0.046)	-0.02 (0.005)	0.86 (0.020)	-0.15 (0.006)
	PBORG	-1.58 (0.064)	0.00 (0.004)	0.81 (0.033)	0.05 (0.063)
Refrigerated Juice	FRUIT	-2.44 (0.129)	-0.00 (0.001)	0.79 (0.041)	0.02 (0.005)
	ICECOFFEE	0.81 (0.062)	0.00 (0.001)	0.55 (0.019)	0.15 (0.012)
	ICETEA	-1.72 (0.043)	0.00 (0.001)	0.36 (0.015)	-0.03 (0.011)
	LEMONADE	-1.70 (0.036)	-0.02 (0.002)	0.27 (0.028)	0.14 (0.013)
	ORANGE	-1.97 (0.049)	0.10 (0.004)	0.91 (0.016)	-0.04 (0.007)
	OTHER	-1.74 (0.055)	-0.01 (0.001)	0.73 (0.035)	0.04 (0.010)

Notes: The estimated coefficients are from a regression of log quantities for category i on its log price, the number of weeks since its last price promotion (duration price), its feature promotion activity, and the number of weeks since its last feature promotion (duration promotion). The regressions are estimated for each category separately and include ZIP code, season, and holiday fixed effects. In the presence of consumer stockpiling behavior, the coefficient on "duration price" should be positive (Hendel and Nevo, 2003).

D NPD Elasticity Curves

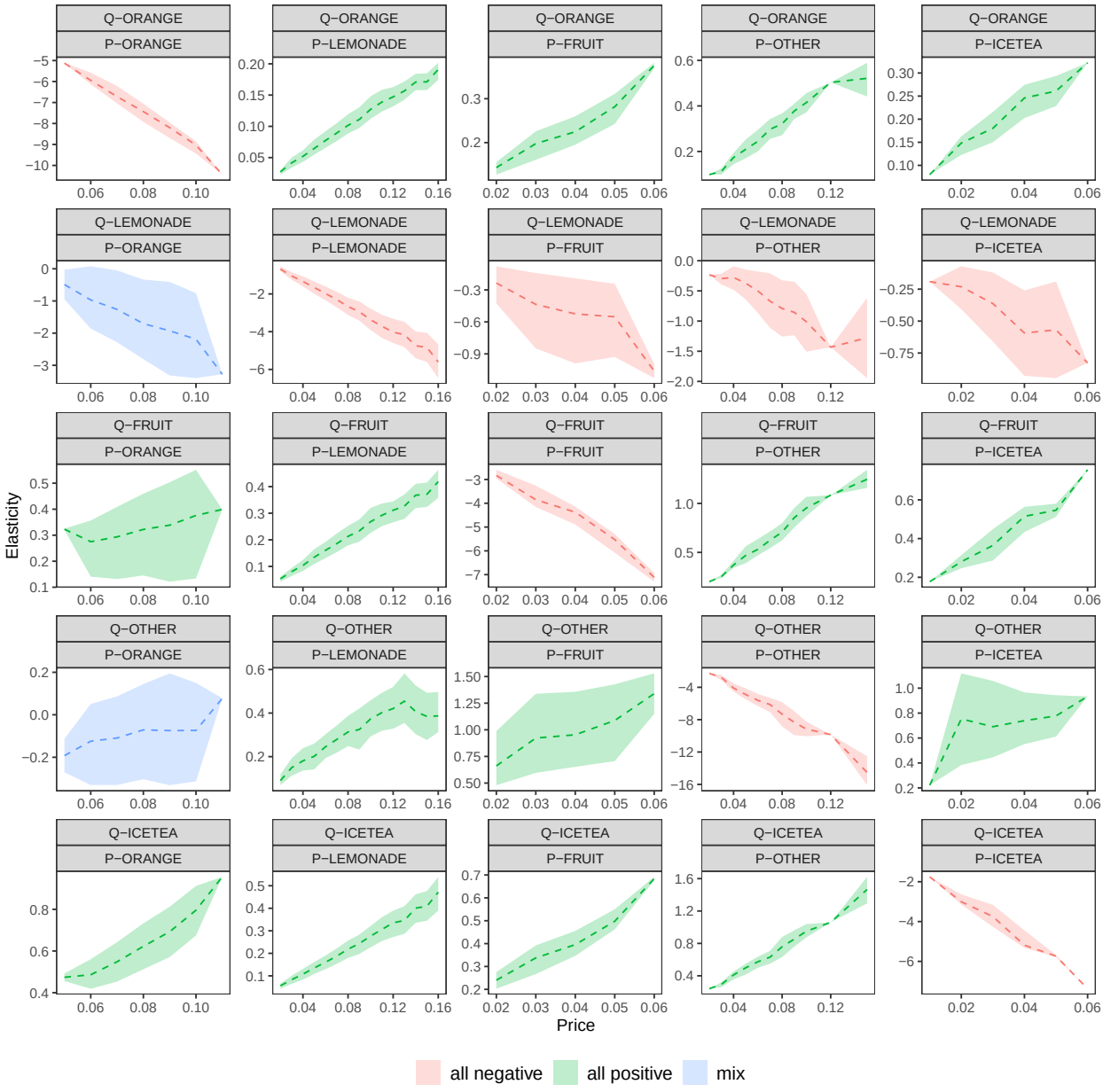

Figure D3: Elasticity Curves (Refrigerated Juice)