

# A Quasi-Bayes Approach to Nonparametric Demand Estimation with Economic Constraints\*

James Brand<sup>†</sup>

Adam N. Smith<sup>‡</sup>

March 24, 2026

## Abstract

This paper develops a new estimation framework for balancing statistical flexibility and economic regularity in multi-product nonparametric demand systems. We take a quasi-Bayes approach that transforms a sieve estimator of inverse demand into a quasi-likelihood and then uses priors to regularize and enforce economic constraints. We implement novel sampling procedures that repose the heavily constrained posterior as the limit of a sequence of softly constrained posteriors, and then utilize sequential Monte Carlo algorithms to push and filter samples through this sequence. Evidence from simulations and across several grocery retail categories shows that, relative to classical estimators, our approach more accurately recovers economic structure and improves finite sample performance. We also introduce an accompanying Julia package (`NPDemand.jl`) to help make nonparametric demand estimation more feasible in applied work.

**Keywords:** Regularization, Shape constraints, Nonparametric IV, Sequential Monte Carlo

---

\*We have benefited from discussions with Jack Collison, Giovanni Compiani, Chris Conlon, Jeff Gortmaker, Jonas Lieber, Ariel Pakes, Kenichi Shimizu, and Chris Walker. Helpful feedback was also provided by seminar and conference participants at EC'25, Marketing Science 2025, Microsoft, NUS, NYU, SBIES 2025, SMU, UCL, and the UK CMA.

<sup>†</sup>Office of the Chief Economist, Microsoft

<sup>‡</sup>UCL School of Management

# 1 Introduction

Estimating empirical models of consumer and firm behavior often involves trading off statistical flexibility and economic validity. Researchers typically seek to recover economic primitives as flexibly as possible while also ensuring that any domain-specific restrictions from economic theory are satisfied. Examples include estimating flexible product substitution patterns that satisfy the laws of demand, output elasticities subject to non-increasing returns to scale, or distributions of bidder valuations subject to monotone bidding strategies. In prescriptive work, where the goal is to inform decision-making, economic regularity often takes precedence and researchers tend to favor simpler parametric specifications. The development of alternative approaches to counterfactual measurement which are flexible yet tractable and comply with economic theory continues to be an active area of research.

In this paper, we focus on the analysis of consumer demand and offer a new quasi-Bayesian approach for estimating nonparametric demand systems for differentiated products. Demand estimation is a foundational area of applied work, and researchers have long been interested in developing flexible models that retain proper microfoundations. In simple logit specifications, the shape and curvature of demand is governed by a single parameter which makes it easier to ensure validity but comes at the cost of flexibility (Berry and Haile, 2021; Miravete et al., 2024; Birchall et al., 2024). As functional form is relaxed, the parameter space necessarily grows and it can be harder to enforce economic validity. For example, ensuring demand is downward-sloping in a more flexible mixed logit model requires constraints over a higher-dimensional distribution of utility parameters. Enforcing shape constraints on nonparametric specifications of demand is even more challenging. The main contribution of this paper is to provide an estimation framework that delivers economic regularization on flexible specifications of demand, and is especially useful when (i) theory places restrictions on nonlinear functions of model parameters, and (ii) the parameter space is large or infinite-dimensional.

The starting point of our empirical framework is the nonparametric demand (NPD) system presented in Berry and Haile (2014), in which market shares are expressed as functions of observed and unobserved product characteristics, including endogenous prices. The *inverse* demand functions, which are identified under a conditional moment restriction, are the structural objects targeted in estimation. The econometric model is thus a conditional moment restriction model that can be estimated using nonparametric instrumental variables (NPIV) methods (Newey and Powell, 2003). For example, if the inverse demand function is approximated using the method of sieves, then a sieve GMM estimator can be derived following Compiani (2022). From there, we apply quasi-Bayesian methods in the style of Chernozhukov and Hong (2003). We transform the GMM objective function into a quasi-likelihood, specify priors over model parameters (or functions thereof) to regularize and enforce constraints, and then apply Bayes rule to arrive at a quasi-posterior.

Our quasi-Bayesian NPD (QBNPD) specification offers the following advantages. First, by decoupling functional form from economic constraints, we can pare back assumptions on functional form to only those which are necessary for identification and then incorporate economic constraints

via priors. This gives researchers more flexibility in the amount and type of economic regularity imposed on their analysis, and can be easily tailored to specific empirical settings and counterfactuals of interest. For example, one need not assume demand has a logit-like functional form (or is even parametric) to ensure that estimated demand curves are downward-sloping. Second, we arrive at a posterior by way of conditional moment restrictions rather than a likelihood which is attractive because it (i) precludes the need for parametric assumptions on unobservable demand shocks, and (ii) sidesteps some of the usual challenges in deriving a valid likelihood in structural models (due to a nonlinear mapping from stochastic error terms to observed outcomes). Third, the priors in our model naturally regularize an NPIV estimator, which is notoriously high-variance and often empirically unstable due to the ill-posed nature of the problem (Darolles et al., 2011).

A final advantage is that enforcing theory via priors can ensure that the estimated demand functions satisfy all desired constraints in finite samples. As a point of comparison, Compiani (2022) imposes linear restrictions on parameters of the sieve estimator of the inverse demand function, which can be enforced as convex constraints in the GMM program. While shape restrictions on the demand function can imply shape restrictions on the inverse demand function, the converse need not be true given the multivariate nature of the target function. The linear restrictions derived in Compiani (2022) are thus necessary but not sufficient for the desired constraints on demand.<sup>1</sup> Our approach closes this gap by using the prior to directly regularize economic functionals of demand.

Despite many conceptual advantages, our QBNPD estimator is both high-dimensional and highly constrained which makes sampling difficult for even the best off-the-shelf algorithms. We overcome this challenge by using a novel variant of Sequential Monte Carlo (SMC) (Doucet et al., 2001; Chopin, 2002; Del Moral et al., 2006) introduced by Golchi and Campbell (2016). While SMC is best known for its applications in dynamic state space models, Golchi and Campbell (2016) show how the same iterative resampling procedures can be used for static models subject to the prior restriction  $\theta \in \mathcal{C}(\Theta)$  for some constrained set  $\mathcal{C}(\Theta) \subset \Theta$ . The idea is to first rewrite a hard constraint  $\mathbf{1}(\theta \in \mathcal{C}(\Theta))$  as a sequence of soft constraint functions  $\|\theta\|_{\mathcal{C}}^{\lambda}$  parameterized by a penalty  $\lambda$  which converges to the hard constraint as  $\lambda \rightarrow \infty$ . This new sequence of priors induces a corresponding sequence of posteriors that bridges the unconstrained model to the fully constrained model, and itself becomes the new target for sampling.<sup>2</sup>

Sampling then proceeds in two steps. In the first step, we sample from the unconstrained posterior using standard Markov chain Monte Carlo (MCMC) methods. The unconstrained posterior is well-behaved and can be sampled from using any off-the-shelf algorithm, such as Hamiltonian Monte Carlo. In the second step, we use SMC—i.e., a series of reweighting, resampling, and particle

---

<sup>1</sup>To enforce constraints which are both necessary and sufficient would require solving a non-convex optimization problem for which there are no theoretical convergence guarantees for off-the-shelf solvers. Practical convergence will depend on the complexity of the desired constraints and on the available data. In our own testing, we found that even the most performant nonlinear solvers in Julia’s `NLOpt.jl` library routinely failed to converge when applied to NPD problems with non-convex constraints.

<sup>2</sup>The idea of using a bridging sequence of models to facilitate sampling is analogous to tempering (Del Moral et al., 2006; Jasra et al., 2011). Note that we are tempering the prior rather than the data or likelihood, as is done in most applications (Herbst and Schorfheide, 2014, 2019; Chen et al., 2018).

rejuvenating steps—to move draws through the models in this sequence until we eventually reach the limiting posterior. Intuitively, our sampling approach works by reposing an infeasible sampling problem with a sequence of much easier sampling problems that converges to the desired problem.

We demonstrate the value of our QBNPD estimator using both Monte Carlo simulations and retail scanner data. In both contexts, we evaluate our quasi-Bayes estimator against an unconstrained GMM estimator and a constrained GMM estimator that incorporates necessary linear restrictions (Compiani, 2022). In simulations, we consider two data-generating processes (DGPs): the first is a simple logit specification and the second is a more complex system of substitutes and complements. Across both DGPs, our quasi-Bayes method provides superior finite sample performance relative to both GMM benchmarks. Specifically, we show that: (i) both constrained estimators exhibit faster rates of convergence relative to the unconstrained estimator, (ii) the quasi-Bayes approach reliably enforces all desired constraints; in contrast, a constrained GMM estimator continues to violate constraints in up to half of all markets, despite the presence of linear restrictions, and (iii) the quasi-Bayes estimator delivers lower estimation error, with the relative gains increasing with the complexity of the DGP and/or dimension of the estimation problem. Together, our results show that constraints help researchers make the most of the data they have, and often lead to the same statistical accuracy as an unconstrained estimator trained with more than 10 times the amount of data.

We then apply our methods to several consumer packaged goods (CPG) markets using two years of standard grocery retail data spanning hundreds of stores. We motivate the use of constraints by first documenting the performance of an unconstrained GMM estimator across 12 product categories. We find that whenever there are more than two products, an unconstrained estimator produces upward-sloping demand curves in more than 50% of markets. We then zoom in on four representative product categories: Ketchup, Frozen Ice Cream, Fish Canned, and Jams, Jellies, & Peanut Butter. The categories are representative in that they span a range of important dimensions for our estimation problem—namely, the number of products, the number of subcategories, and the nature of substitution across products and subcategories. We find that both GMM specifications continue to exhibit prohibitively high variance and cannot reliably learn economic structure from our data. We document differences in constraint violations, estimated elasticities, and estimated diversion ratios across all estimators, and find that our quasi-Bayes approach delivers useful and economically important regularization on the estimated demand functions.

Finally, we introduce an accompanying Julia package, `NPDemand.jl`, which implements both GMM and quasi-Bayesian estimators with minimal user overhead. Our package relies on `Turing.jl` (Ge et al., 2018), Julia’s state-of-the-art probabilistic programming language, to map our econometric problem into Julia and to automate computation in the first stage of our sampling approach. The package also includes easy-to-use functions for calculating price elasticities or predicting market shares at counterfactual prices. We use the package to produce all of the empirical results in the paper. Additional package details can be found in [Appendix B](#) and on [GitHub](#).

## Related literature

There is a long history of testing and enforcing consumer theory when estimating flexible demand functions (see, e.g., [Deaton and Muellbauer, 1980](#); [Hausman and Newey, 1995](#); [Haag et al., 2009](#); [Hoderlein and Lewbel, 2012](#); [Blundell et al., 2012, 2017](#)). Until recently, empirical work in this literature has largely focused on modeling demand for undifferentiated commodities (e.g., gasoline) or highly aggregated expenditure groups (e.g., clothing). [Berry et al. \(2013\)](#) and [Berry and Haile \(2014\)](#) develop novel nonparametric identification results for multi-product NPD systems, paving the way for more flexible analyses of demand for differentiated products. [Compiani \(2022\)](#) presents the first empirical approach to NPD estimation and is thus closest to our work—also serving as one possible starting point for our quasi-Bayesian approach.

Relative to [Compiani \(2022\)](#), we make the following contributions. First, we generalize beyond linear restrictions and use priors on functions of sieve coefficients to ensure the desired constraints hold on the implied system of demand equations. In our empirical work, we show that (i) linear restrictions alone are insufficient and often lead to appreciably large shares of violations across markets, and (ii) priors over economic functionals effectively close this gap. Second, our approach also allows us to consider more complex non-convex constraints while keeping the sampling problem tractable. For example, we extend the linear cross-good monotonicity constraints developed by [Compiani \(2022\)](#) to settings with multiple product groups to allow for within-group substitutes and across-group complements. Third, our regularization approach allows us to do more with less—we no longer require extremely large sample sizes (or prohibitively small demand systems) in order to guarantee valid estimates of demand. For example, the empirical analysis of [Compiani \(2022\)](#) estimates a demand system for two goods and 80,000 markets; in our applications, we estimate demand systems for up to six goods using only 5,500 markets. Finally, we offer a new software package for estimating NPD systems using both GMM and quasi-Bayes approaches.

We also contribute to the NPIV literature by offering a new applied framework for economic regularization. NPIV models are ill-posed inverse problems that require regularization to ensure stability in both the estimand and estimator. Several types of regularization solutions exist (see, e.g., [Carrasco et al., 2007](#)). In this paper we focus on the method of sieves, which implicitly regularizes through a finite-dimensional approximating space. Even if the sieve adds enough regularity to ensure that the problem is well-posed, the target parameter space often remains large and unconstrained estimators can exhibit high variance. Additional forms of explicit regularization, such as statistical penalties ([Blundell et al., 2007](#); [Chen and Pouzo, 2012](#)) and shape constraints ([Freyberger and Horowitz, 2015](#); [Chetverikov and Wilhelm, 2017](#); [Compiani, 2022](#); [Chernozhukov et al., 2023](#)) can improve finite sample performance. Our approach follows in the same spirit, as we also advocate for the use of shape constraints in NPIV estimation, though our operationalization of constraints is novel. We believe that enforcing constraints through priors can make it easier for researchers to bring economic intuition to bear on the analysis, and to navigate the trade-offs between statistical flexibility and economic regularity inherent in nonparametric structural models.

## Outline

The rest of the paper is organized as follows. Section 2 introduces the focal nonparametric demand system. Section 3 outlines our proposed quasi-Bayesian estimation framework. Section 4 discusses computation and our posterior sampling strategy. Section 5 presents a set of simulation results, and Section 6 presents empirical evidence of the value of our approach using retail scanner data. Section 7 discusses limitations and possible areas for future work. Section 8 concludes.

## 2 Demand

### 2.1 A Nonparametric Specification

We start by outlining a generic market-level demand model for a fixed assortment of  $J$  goods in the style of [Berry and Haile \(2014\)](#). For each product  $j = 1, \dots, J$  and market  $t = 1, \dots, T$ , we observe a market share  $s_{jt}$ , a price  $p_{jt}$ , and a vector of product-market characteristics  $\mathbf{x}_{jt}$ . We also assume there is an unobserved product characteristic  $\xi_{jt}$  that is potentially correlated with price. The demand equation for good  $j$  in market  $t$  can be written as

$$s_{jt} = \sigma_j(\mathbf{X}_t, \mathbf{p}_t, \boldsymbol{\xi}_t) \quad (1)$$

where  $\mathbf{X}_t = (\mathbf{x}_{1t}, \dots, \mathbf{x}_{Jt})$ ,  $\mathbf{p}_t = (p_{1t}, \dots, p_{Jt})$ , and  $\boldsymbol{\xi}_t = (\xi_{1t}, \dots, \xi_{Jt})$ . The model above presents two key econometric challenges. The first is that each demand function depends on the entire  $J$ -vector of unobservables, and so the model is intrinsically different from statistical regression ([Berry and Haile, 2021](#)). The second is that we allow  $\mathbb{E}(\xi_{jt} | \mathbf{p}_t, \mathbf{X}_t) \neq 0$  and thus prices are treated as endogenous. Both challenges can be addressed through an inversion of the demand system above. However, such inversion is not possible without additional assumptions which we outline below.

**Assumption 1** (Index). *There exists an additive index structure  $\delta_{jt} \equiv \delta(\mathbf{x}_{jt}, \xi_{jt}) = \mathbf{x}'_{jt}\beta_j + \xi_{jt}$  such that  $\sigma_j(\mathbf{X}_t, \mathbf{p}_t, \boldsymbol{\xi}_t) = \sigma_j(\boldsymbol{\delta}_t, \mathbf{p}_t)$  for all  $j = 1, \dots, J$ .*

The first assumption imposes an index structure which restricts the way that the unobserved and observed (non-price) characteristics affect demand. Such assumptions are prevalent in simpler parametric models, as will be seen below. Note that we have assumed the entire vector of non-price observable characteristics  $\mathbf{x}_{jt}$  enter demand via the index, which is more restrictive than what is required for identification ([Berry and Haile, 2014](#)). We maintain the assumption to simplify model exposition, but it can in principle be relaxed.

**Assumption 2** (Connected Substitutes).

- (i)  $\sigma_k(\boldsymbol{\delta}_t, \mathbf{p}_t)$  is non-increasing in  $\delta_{jt}$  for all  $j \neq k$  and any  $\boldsymbol{\delta}_t, \mathbf{p}_t \in \mathbb{R}^{2J}$ .
- (ii) For each  $\boldsymbol{\delta}_t, \mathbf{p}_t$  and any nonempty  $\mathcal{J} \subseteq \{1, \dots, J\}$ , there exists  $j \in \mathcal{J}$  and  $k \notin \mathcal{J}$  such that  $\sigma_k(\boldsymbol{\delta}_t, \mathbf{p}_t)$  is strictly decreasing in  $\delta_{jt}$ .

Table 1: Economic Shape Constraints

Constraint	Definition
1. Own-good monotonicity	$\partial\sigma_j/\partial p_j \leq 0 \forall j$
2. Diagonal dominance	$ \partial\sigma_j/\partial p_j  \geq \sum_{k \neq j}  \partial\sigma_k/\partial p_j $
3. Weak substitutes (all goods)	$\partial\sigma_j/\partial p_k \geq 0 \forall j \neq k$
4. Weak substitutes (within groups)	$\partial\sigma_j/\partial p_k \geq 0 \forall j \neq k, G_j = G_k$
5. Weak substitutes (across groups)	$\partial\sigma_j/\partial p_k \geq 0 \forall j \neq k, G_j \neq G_k$
6. Weak complements (across groups)	$\partial\sigma_j/\partial p_k \leq 0 \forall j \neq k, G_j \neq G_k$

Notes: (1) Own-good monotonicity enforces negative own-price effects and downward-sloping demand. (2) Weak substitutes (complements) enforces positive (negative) cross-price effects. (3) Diagonal dominance enforces the magnitude of the own effect to exceed the sum of the cross effects.

The second assumption is the “connected substitutes” condition of [Berry et al. \(2013\)](#), which imposes minimal shape constraints on  $\sigma_j$  required for inversion. As shown in [Berry and Haile \(2014, 2021\)](#) and [Compiani \(2022\)](#), the estimand remains highly flexible in that it subsumes many commonly used parametric models and can accommodate assortments of both substitutable and complementary goods.

**Assumption 3** (Instruments). *There exists a vector of instruments  $\mathbf{z}_t = (z_{1t}, \dots, z_{Jt})$  excluded from  $\sigma_1, \dots, \sigma_J$  such that  $\mathbb{E}(\xi_{jt} | \mathbf{X}_t, \mathbf{z}_t) = 0$  almost surely for all  $j = 1, \dots, J$ .*

The final assumption introduces an exclusion restriction which is also commonly used to aid identification in demand systems for differentiated goods. Together, Assumptions 1, 2, and 3 imply that we can invert the demand system:

$$s_{jt} = \sigma_j(\delta(\mathbf{x}_{1t}, \xi_{1t}), \dots, \delta(\mathbf{x}_{Jt}, \xi_{Jt}), \mathbf{p}_t) \iff \mathbf{x}'_{jt} \beta_j + \xi_{jt} = \sigma_j^{-1}(s_{1t}, \dots, s_{Jt}, \mathbf{p}_t) \quad (2)$$

and estimate  $\sigma^{-1} = (\sigma_1^{-1}, \dots, \sigma_J^{-1})$  using NPIV methods ([Newey and Powell, 2003](#)).

## 2.2 Economic Constraints and the Role of Functional Form

The main focus and contribution of this paper is to provide a framework for estimating the flexible NPD system above while also ensuring that estimated system satisfies restrictions required by economic theory. In particular, we are interested in shape restrictions on demand, like own-good monotonicity  $\partial\sigma_j/\partial p_j < 0$  (i.e., downward-sloping demand) and cross-good monotonicity to enforce pairs of goods to be substitutes or complements. We also consider more complex variations of cross-good monotonicity where *product groups* are pairwise substitutes or complements. [Table 1](#) provides a list of the shape constraints we consider throughout our analysis.

As discussed in [Compiani \(2022\)](#), the main technical challenge in enforcing shape constraints on NPD systems is that theory places restrictions on  $\sigma$  but  $\sigma^{-1}$  is the focal estimand. Because each component  $\sigma_j^{-1}$  is itself a multivariate function, the mapping of constraints between  $\sigma^{-1}$  and  $\sigma$  is non-trivial. Our quasi-Bayesian approach directly aids in solving this problem.

Before presenting our econometric NPD framework, we first discuss a few examples of simpler parametric demand specifications to show that (i) many economic restrictions are implicitly enforced via functional form, and (ii) as functional form is relaxed, more work is required to explicitly enforce theory restrictions in estimation.

**Example 1** (Logit) In the canonical logit model of demand, observed market shares represent discrete choice outcomes among a population of consumers. Let the indirect utility of consumer  $i$  for product  $j$  at time  $t$  be written as

$$u_{ijt} = \delta(\mathbf{x}_{jt}, p_{jt}, \xi_{jt}) + \epsilon_{ijt} = \bar{\delta}(\mathbf{x}_{jt}, p_{jt}) + \xi_{jt} + \epsilon_{ijt}. \quad (3)$$

As shown in [Berry \(1994\)](#), we can write

$$s_{jt} = \sigma_j(\delta_{1t}, \dots, \delta_{Jt}) = \frac{e^{\delta_{jt}}}{1 + \sum_{k=1}^J e^{\delta_{kt}}} \iff \bar{\delta}(\mathbf{x}_{jt}, p_{jt}) + \xi_{jt} = \log(s_{jt}/s_{0t}) \quad (4)$$

and the inverted system on the right-hand side serves as the set of linear IV estimating equations we can take to the data. Suppose we are interested in the own-price effect  $\partial s_{jt}/\partial p_{jt}$ , which based on the logit functional form can be expressed as

$$\frac{\partial s_{jt}}{\partial p_{jt}} = \frac{\partial \bar{\delta}(\mathbf{x}_{jt}, p_{jt})}{\partial p_{jt}} s_{jt}(1 - s_{jt}). \quad (5)$$

Then it immediately follows that  $\partial s_{jt}/\partial p_{jt} < 0 \iff \partial \bar{\delta}(\mathbf{x}_{jt}, p_{jt})/\partial p_j < 0$ . For example, in the usual parameterization of the index  $\bar{\delta}(\mathbf{x}_{jt}, p_{jt}) = \mathbf{x}'_{jt}\beta_j + \alpha p_{jt}$ , ensuring demand is downward-sloping simply amounts to the constraint  $\alpha < 0$ . If we specify a more flexible index structure such as  $\bar{\delta}(\mathbf{x}_{jt}, p_{jt}) = \mathbf{x}'_{jt}\beta_j + \sum_{d=0}^D \alpha_d p_{jt}^d$ , then monotonicity requires constraints on a larger set of parameters  $\boldsymbol{\alpha} \in \mathbb{R}^D$ .

**Example 2** (Mixed logit) Now consider a slightly more flexible mixed logit specification where consumer utility is given by

$$u_{ijt} = \delta(\mathbf{x}_{jt}, \xi_{jt}) + g(p_{jt}; \alpha_i) + \epsilon_{ijt} = \bar{\delta}(\mathbf{x}_{jt}) + \xi_{jt} + g(p_{jt}; \alpha_i) + \epsilon_{ijt}. \quad (6)$$

Assume the price coefficient is heterogeneous across consumers and has a distribution  $F(\alpha)$ . It follows that

$$s_{jt} = \sigma_j(\delta_{1t}, \dots, \delta_{Jt}, \mathbf{p}_t) = \int \frac{e^{\delta_{jt} + g(p_{jt}; \alpha_i)}}{1 + \sum_{k=1}^J e^{\delta_{kt} + g(p_{kt}; \alpha_i)}} dF(\alpha_i) \iff \bar{\delta}(\mathbf{x}_{jt}) + \xi_{jt} = \sigma_j^{-1}(s_{1t}, \dots, s_{Jt}, \mathbf{p}_t; F) \quad (7)$$

where the inverse  $\sigma_j^{-1}(\cdot; F)$  no longer has an analytic expression but can be characterized by an integral over  $F$ . Therefore, ensuring  $\partial s_j/\partial p_j < 0$ , for example, requires jointly constraining the functional form of  $g$  and the full distribution of heterogeneity  $F \in \mathcal{F}$ . Even in the simplest case with  $g(p_{jt}; \alpha_i) = \alpha_i p_{jt}$ , enforcing downward-sloping aggregate demand curves requires the

mean of  $\alpha_i$  to be negative and places restrictions on the variance and skew to limit the share of individual-level coefficients which are positive. As the space of candidate distributions  $\mathcal{F}$  gets larger or  $g(\cdot)$  is specified more flexibly, it becomes more challenging to specify and enforce the desired shape constraints. A common approach to enforcing such constraints is to restrict  $\mathcal{F}$  to a class of parametric distributions defined on  $\mathbb{R}^-$ , but since  $F$  serves as a mixing distribution for consumer-level logit elasticities, parametric assumptions again impact the flexibility of the model (Miravete et al., 2024). Targeting  $F$  nonparametrically, on the other hand, offers more flexibility but also makes enforcing monotonicity significantly more computationally challenging.

**Example 3** (Inverse Product Differentiation Logit) Finally, the model introduced by Fosgerau et al. (2024) allows for even more complex substitution patterns. The inverse demand function in this setting involves a function of multiple other products’ market shares. The estimating equation in this case takes the form:

$$\log\left(\frac{s_{jt}}{s_{0t}}\right) = \delta_{jt}(\mathbf{x}_{jt}, p_{jt}, \xi_{jt}) + \sum_{d=1}^D \mu_d \log\left(\frac{s_{jt}}{s_{d(j),t}}\right), \quad (8)$$

where  $d$  indexes a set of predefined product grouping characteristics. The second term in this equation allows for flexible substitution patterns between products (including complementarity), at the cost of making constraints on cross-product relationships much more complex to impose. In this model, restricting the sign of the price coefficient in the utility function is sufficient to restrict the sign of own-price elasticities, but cross-price elasticities are a function of a matrix inverse that includes all grouping parameters  $\mu_d$  and realized market shares themselves. As a result, imposing (for example) that all cross-price elasticities are positive requires imposing complex nonlinear constraints at all observed market shares.

The examples above show how parametric assumptions play an important role in enforcing economic restrictions in models of demand, and that it requires more work to enforce constraints as we start to relax functional form assumptions. While our quasi-Bayesian approach can be used to enforce constraints in simpler parametric models, we focus on NPD specifications where heavy regularization via functional form is intentionally absent. Finding ways to regularize using theory is thus first-order to the measurement problem and is a key contribution of our approach.

### 3 A Quasi-Bayesian Approach to Estimation

#### 3.1 From Conditional Moment Restrictions to a Quasi-Likelihood

Our starting point is the conditional moment restriction model induced by the system of inverse demand equations in (2) and exclusion restriction in Assumption 3:

$$m_j(\sigma_j^{-1}) = \mathbb{E}\left(\mathbf{x}'_{jt}\beta_j - \sigma_j^{-1}(\tilde{\mathbf{s}}_t) \middle| \mathbf{X}_t, \mathbf{z}_t\right) = 0 \quad \text{for all } j = 1, \dots, J \quad (9)$$

where  $\sigma_j^{-1}$  is the structural function to be estimated and  $\tilde{\mathbf{s}}_t$  is an augmented share vector which is  $\tilde{J}$ -dimensional and represents the vector of shares and any other product characteristics left outside of the index. For example, based on the demand specification in (2) we would write  $\tilde{\mathbf{s}}_t = (\mathbf{s}_t, \mathbf{p}_t)$  and  $\tilde{J} = 2J$ . This notation also nests the simpler case when prices are included inside the index and the share vector  $\mathbf{s}_t$  is the only argument of  $\sigma_j^{-1}$ .

The conditional moment restriction in (9) cannot directly serve as the basis for estimation because it depends on  $\sigma_j^{-1}$  which is unknown. We therefore approximate  $\sigma^{-1} = (\sigma_1^{-1}, \dots, \sigma_J^{-1}) \in \Sigma^{-1}$  by a finite-dimensional sieve space  $\Sigma_T^{-1}$  which becomes dense in  $\Sigma^{-1}$  as  $T \rightarrow \infty$ . Specifically, we approximate each  $\sigma_j^{-1}$  with a tensor product sieve:

$$\tilde{\sigma}_j^{-1}(\tilde{\mathbf{s}}; \theta_j) = \sum_{k_1=0}^{K_j} \cdots \sum_{k_{\tilde{J}}=0}^{K_j} \theta_{jk_1 \dots k_{\tilde{J}}} \prod_{\ell=1}^{\tilde{J}} \phi_{jk_\ell}(\tilde{s}_\ell). \quad (10)$$

Each sieve is characterized by a set of basis functions  $\phi_j(\cdot) = (\phi_{j1}(\cdot), \dots, \phi_{jM_j}(\cdot))'$  and an  $M_j$ -dimensional coefficient vector  $\theta_j$  where  $M_j = (K_j + 1)^{\tilde{J}}$ . While we omit any  $M_j$  or  $T$  superscripts on  $\theta$  to reduce notational burden, we include this more formal notation (and allow  $M_j$  to grow with  $T$ ) when presenting asymptotic results in [Appendix A](#). For simplicity, we also assume that the same basis functions  $\phi_j(\cdot)$  are used to approximate the instrument space.

Given a choice of basis functions, we can replace  $m_j(\sigma_j^{-1})$  with its finite sample counterpart  $\hat{m}_j(\theta_j) = \frac{1}{T} \sum_{t=1}^T (\mathbf{x}'_{jt} \beta_j - \tilde{\sigma}_j^{-1}(\tilde{\mathbf{s}}_t; \theta_j)) \cdot \phi_j(\mathbf{X}_t, \mathbf{z}_t)$  and define the statistical criterion function:

$$Q_T(\theta) = \sum_{j=1}^J \hat{m}_j(\theta_j)' \hat{\Omega}_j \hat{m}_j(\theta_j) \quad (11)$$

where  $\theta = (\theta'_1, \dots, \theta'_J)'$  and  $\hat{\Omega}_j$  is an  $M_j \times M_j$  weighting matrix. Instead of pursuing a frequentist optimization-based approach to estimation, we transform the criterion function into a quasi-likelihood and pursue a (quasi-)Bayesian sampling approach. Specifically, let

$$L_T(\theta) = e^{-\frac{T}{2} Q_T(\theta)} \quad (12)$$

denote the quasi-likelihood which characterizes fit to the data with respect to model parameters. Note that while the quasi-likelihood is not a valid density, the induced quasi-posterior will appropriately normalize the quasi-likelihood and integrate to one.

### 3.2 Incorporating Constraints via the Prior

We now turn to the specification of priors over sieve parameters  $\theta$  which, together with the quasi-likelihood in (12), induce a quasi-posterior. As a starting point, consider a baseline prior specification  $\bar{\pi}(\theta)$  for some choice of density  $\bar{\pi}(\cdot)$ . Absent any economic constraints, the baseline prior will provide statistical regularization and help discipline both the estimand and estimator. The induced

quasi-posterior takes the form:

$$\bar{\pi}(\theta|\mathcal{D}_T) = \frac{L_T(\theta)\bar{\pi}(\theta)}{\int_{\Theta} L_T(\theta)\bar{\pi}(\theta)d\theta} \quad (13)$$

where  $\mathcal{D}_T = \{\mathbf{s}_t, \mathbf{p}_t, \mathbf{X}_t, \mathbf{z}_t\}_{t=1}^T$  is the set of observed data across all markets.

There are two routes to incorporating the economic constraints from [Table 1](#) into the model above. The first route operates through the sieve approximation of  $\sigma^{-1}$  inside the quasi-likelihood and chooses basis functions which are amenable to shape constraints. This is the approach taken by [Compiani \(2022\)](#)—albeit in a frequentist estimation framework—who advocates for the use of Bernstein basis polynomials and shows that certain constraints on  $\sigma$  map to constraints on  $\sigma^{-1}$ , which can then be enforced via linear restrictions on polynomial coefficients  $\theta$ . For example, if demand is downward sloping ( $\partial\sigma_j/\partial p_j \leq 0$ ), then there exists an analogous restriction on  $\sigma_j^{-1}$  which can be enforced by a linear ordering of Bernstein polynomial coefficients. This approach can easily be enforced in the baseline model above via reparameterizations ([Gelman, 2004](#); [Pachali et al., 2020](#); [Gallant et al., 2022](#)).<sup>3</sup>

While operationally simple, this first route alone is unable to guarantee that the estimated demand system satisfies all desired economic constraints. Restrictions imposed on the inverse system  $\sigma^{-1}$  are generally necessary but not sufficient for the desired restrictions on  $\sigma$ . As a result, even if we impose necessary restrictions on  $\sigma^{-1}$ , the corresponding demand system  $\sigma$  may still violate constraints in some markets.

The second route incorporates economic constraints into the prior directly. To this end, we propose the following augmented prior:

$$\pi^*(\theta) \propto \bar{\pi}(\theta)\mathbf{1}(\theta \in \mathcal{C}(\Theta)) \quad (14)$$

which only places mass on the constrained set  $\mathcal{C}(\Theta) \subset \Theta$  where all desired constraints on  $\sigma$  are satisfied. The corresponding fully constrained quasi-posterior is:

$$\pi^*(\theta|\mathcal{D}_T) = \frac{L_T(\theta)\pi^*(\theta)}{\int_{\mathcal{C}(\Theta)} L_T(\theta)\pi^*(\theta)d\theta} \quad (15)$$

which inherits the support of  $\pi^*(\cdot)$  and only places mass on  $\mathcal{C}(\Theta)$ . This second route offers several advantages relative to the first. By defining the constraint set in terms of  $\sigma$ , we overcome the main limitation of the reparameterization approach above. Any posterior inferences made about demand (or functions thereof) via [\(15\)](#) are guaranteed to satisfy the desired economic constraints. Moreover, in this approach we can implement more complex constraints, including those which are nonlinear functions of model parameters. The cross-group monotonicity restrictions in [Table 1](#), for example, cannot be expressed as linear functions of model parameters. This approach also allows

---

<sup>3</sup>For example, if we want to enforce  $\theta_1 < \theta_2$ , then we can define a transformation:  $(\theta_1, \theta_2) = g(\vartheta_1, \vartheta_2) = (\vartheta_1, \vartheta_1 + e^{\vartheta_2})$  where  $\vartheta$  are the new unconstrained parameters. Then for any pair  $(\vartheta_1, \vartheta_2) \in \mathbb{R}^2$ , we are guaranteed to have  $\theta_1 < \theta_2$  when evaluating the quasi-likelihood function  $L_T(g(\vartheta))$ . While we are using the two-dimensional example for ease of illustration, we can derive a reparameterization for any set of sign and order constraints of the form  $\mathcal{C}'(\Theta) = \{\theta : \mathbf{A}\theta \leq 0\}$ .

the operationalization of constraints to be decoupled from the choice of sieve approximation. As a result, we can in principle use a wider class of sieve approximations and need not rely on linear polynomial approximations that lend themselves to shape constraints. If linear restrictions are available, then the best strategy is to use both reparameterization and prior constraint approaches. That is, we can reparameterize the model according to the necessary linear restrictions and then specify dogmatic priors to effectively close the gap and ensure that constraints are satisfied.

### 3.3 Estimation and Inference Procedures

As in any Bayesian model, our inferences about model parameters (and functions thereof) come via the (quasi-)posterior. For example, we can write the quasi-posterior mean of any target function  $h(\theta)$  as:

$$\mathbb{E}(h(\theta)|\mathcal{D}_T) = \int h(\theta)\pi(\theta|\mathcal{D}_T)d\theta \quad (16)$$

which can be approximated via Monte Carlo integration after producing a sample of draws from the posterior. Uncertainty quantification is also automatic given samples from the quasi-posterior. For example, a  $100(1 - \alpha)\%$  credible region  $C_\alpha$  satisfies

$$\mathbb{P}(h(\theta) \in C_\alpha|\mathcal{D}_T) = \int_{\{\theta:h(\theta)\in C_\alpha\}} \pi(\theta|\mathcal{D}_T) = 1 - \alpha. \quad (17)$$

Such a region  $C_\alpha$  is not unique but has many simple operationalizations. For example, a quantile-based  $100(1 - \alpha)\%$  credible interval  $(c_\alpha^L, c_\alpha^U)$  satisfies

$$\mathbb{P}(h(\theta) < c_\alpha^L|\mathcal{D}_T) = \alpha/2 \quad \text{and} \quad \mathbb{P}(h(\theta) > c_\alpha^U|\mathcal{D}_T) = \alpha/2 \quad (18)$$

which can be easily approximated by the  $\alpha/2$  and  $1 - \alpha/2$  quantiles of the empirical distribution of draws from the quasi-posterior.

### 3.4 Asymptotics

While uncertainty quantification in quasi-Bayesian models is procedurally automatic, more work is required to establish a formal link to other valid forms of inference. Here we pursue a frequentist notion of validity based on asymptotic coverage and ask if and when quasi-Bayes credible intervals coincide with frequentist confidence sets in large samples.<sup>4</sup> Bernstein-von Mises (BvM) results provide such guarantees by establishing conditions under which the target quasi-posterior converges to an appropriately centered and scaled Gaussian distribution. The seminal work of [Chernozhukov and Hong \(2003\)](#) establishes a BvM for parametric models, showing that quasi-Bayes credible intervals can exhibit asymptotically exact frequentist coverage. There are now several relevant extensions to quasi-Bayes NPIV models. [Liao and Jiang \(2011\)](#) and [Kato \(2013\)](#) both examine asymptotic behavior of the NPIV quasi-posterior based on a linear sieve approximation of the target structural function. [Liao and Jiang \(2011\)](#) proves that the quasi-posterior is consistent

---

<sup>4</sup>See [Bissiri et al. \(2016\)](#) for a rational belief-updating notion of validity for quasi-Bayesian inference.

and [Kato \(2013\)](#) derives its limiting distribution and minimax optimal contraction rates. More recently, [Kankanala \(2025\)](#) proves a nonparametric BvM for a broader class of conditional moment restriction models using Gaussian processes to flexibly model the target structural functions.

The aforementioned literature offers a precedent for establishing BvM results for models like ours. However, one notable difference is that our target posterior in (15) is heavily constrained. Developing BvMs for NPIV models subject to non-trivial constraints is challenging because the true structural function may well lie on the boundary of the feasible set—an outcome which becomes harder to rule out when the parameter space is infinite dimensional. In this case, the resulting quasi-posterior can fail to admit the local Gaussian approximation required for a standard BvM result ([Bochkina and Green, 2014](#)). Establishing a novel nonparametric BvM result for this nonstandard model would require a substantial amount of additional technical work, which is beyond the scope of this paper. Instead, we offer asymptotic guarantees for the baseline quasi-posterior in (13) and then advocate for the use of this model, alongside point estimates from the fully constrained model, to provide conservative but asymptotically valid credible intervals for quantities of interest.<sup>5</sup>

The baseline quasi-posterior in (13) closely resembles the quasi-Bayes NPIV models studied in the literature. We therefore leverage existing results, particularly those of [Kato \(2013\)](#), to establish its asymptotic behavior. For the sake of brevity, we simply state the main results here in words and leave the technical details on the setup and proof to [Appendix A](#).

**Theorem 1.** *Let  $\bar{\Pi}(\cdot|\mathcal{D}_T)$  denote the baseline posterior distribution admitting the posterior density  $\bar{\pi}(\theta|\mathcal{D}_T) \propto L_T(\theta)\bar{\pi}(\theta)$ , and suppose that all regularity conditions on the target structural functions (1, 2), data (3, A8), moments (A9), conditional expectation operator (A10), sieve approximation (A11, A12), weighting matrix (A13), and prior (A14, A15) hold. Then*

- (i)  $\bar{\Pi}(\cdot|\mathcal{D}_T)$  is consistent and approximately Gaussian in large samples;
- (ii) The credible intervals derived from  $\bar{\Pi}(\cdot|\mathcal{D}_T)$  have asymptotically exact frequentist coverage under optimal weighting of the moments in  $L_T(\theta)$ .

*Proof.* See [Appendix A](#). □

## 4 Computation

Our next objective is to devise an approach to sample from the quasi-posterior in (15). The advent of modern MCMC algorithms and probabilistic programming languages such as `RStan` ([Stan Development Team, 2024](#)), `PyMC` ([Abril-Pla et al., 2023](#)), and `Turing.jl` ([Ge et al., 2018](#)) have made it much easier to effectively sample from complex and high-dimensional Bayesian models. Examples of workhorse MCMC algorithms include random-walk Metropolis-Hastings (MH) and Hamiltonian Monte Carlo (HMC). In both frameworks, parameters are first initialized at some value and then in each subsequent iteration, a new candidate value is proposed and either accepted or rejected.

---

<sup>5</sup>[Compiani \(2022\)](#) follows a similar approach and only derives asymptotic results for an unconstrained GMM estimator.

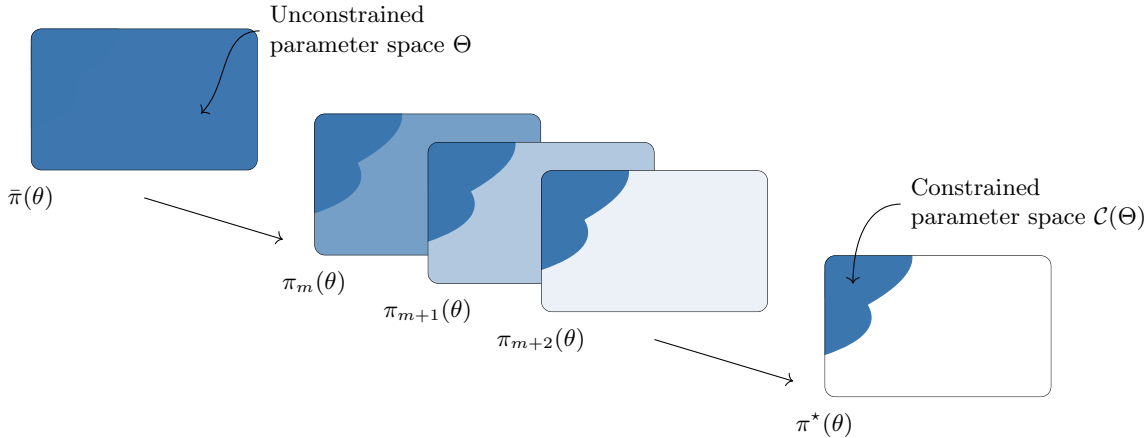


Figure 1: Illustration of the smooth prior sequence used to bridge the unconstrained model with a fully constrained model.

This accept/reject decision is made probabilistically using information from both the posterior and the proposal mechanism.<sup>6</sup>

Unfortunately, the size and complexity of the constraint space in our model render even the best off-the-shelf samplers ineffective. The proposal mechanisms inside general-purpose samplers need not move in directions informed by the constraints, which means that many proposed values will be rejected and the sampler will struggle to explore regions of high posterior probability. We overcome this challenge by using a combination of MCMC and Sequential Monte Carlo (SMC). Specifically, we first use an off-the-shelf MCMC algorithm to first sample from the baseline quasi-posterior in (13). Since this quasi-posterior is not subject to dogmatic constraints, sampling is straightforward. We then utilize the novel SMC algorithm of [Golchi and Campbell \(2016\)](#) to transform samples from  $\bar{\pi}(\theta|\mathcal{D}_T)$  into representative draws from  $\pi^*(\theta|\mathcal{D}_T)$ . This algorithm works by reposing the hard constraint prior  $\pi^*(\theta)$  as the limit of a sequence of soft constraint priors, which induces a sequence of models bridging the baseline posterior to the fully constrained posterior. SMC provides the necessary steps to push and filter draws through this sequence, allowing us to repose a hard sampling problem as the limit of a sequence of easier sampling problems. A graphical illustration of our approach is shown in [Figure 1](#), and formal details are provided below.

#### 4.1 Setup

When taking the constrained posterior in (15) to data, we must first be able to evaluate whether a candidate parameter  $\theta$  satisfies the desired constraints. Given the flexibility of the NPD estimand, this requires the specification of an evaluation domain—that is, a grid of values in the domain of  $\sigma$  for which we check the desired properties of the Jacobian of  $\sigma$ . We therefore rewrite the constrained set  $\mathcal{C}(\cdot)$  to depend on this pre-specified grid of points  $\mathcal{G}$ .

**Definition 1.** Let  $\mathcal{C}_{\mathcal{G}}(\Theta) = \{\theta \in \Theta : c_g(\theta) = 0 \forall g \in \mathcal{G}\}$  where  $\mathcal{G}$  is a grid of points in the domain

<sup>6</sup>For more details on common MCMC algorithms, see [Robert and Casella \(2004\)](#) and [Betancourt \(2017\)](#).

of  $\sigma$  and  $c_g(\theta) \geq 0$  is a measure of constraint violations at  $g \in \mathcal{G}$ .

For now we remain agnostic to the choice of  $\mathcal{G}$ . In practice, a user may want to consider the goals of the analysis. For example, if the goal is to ensure that  $\sigma$  satisfies the desired constraints globally, then we should take  $\mathcal{G}$  to be a grid of points that is sufficiently dense in the domain of  $\sigma$ . In some cases, we may be willing to define a coarser grid, trading off global guarantees with practical feasibility. For example, we may define  $\mathcal{G}$  to be the unique set of domain values observed the data, or some summary statistic of the data to represent an average market. In all such cases, our approach ensures that prior and induced quasi-posterior satisfy the desired constraints everywhere with respect to  $\mathcal{G}$ .

Next, we define the sequence of models that becomes the new target for sampling. The starting point of this sequence is the baseline quasi-posterior  $\bar{\pi}(\theta|\mathcal{D}_T) \propto L_T(\theta)\bar{\pi}(\theta)$  and the ending point is the fully constrained quasi-posterior  $\pi^*(\theta|\mathcal{D}_T) \propto L_T(\theta)\bar{\pi}(\theta)\mathbf{1}(\theta \in \mathcal{C}_{\mathcal{G}}(\Theta))$ . We bridge  $\bar{\pi}(\theta|\mathcal{D}_T)$  to  $\pi^*(\theta|\mathcal{D}_T)$  via the following sequence of penalized models.

$$\begin{aligned} \pi_1(\theta|\mathcal{D}_T) &\propto L_T(\theta)\bar{\pi}(\theta)\|\theta\|_{\mathcal{C}_{\mathcal{G}}}^{\lambda_1} \\ &\vdots \\ \pi_m(\theta|\mathcal{D}_T) &\propto L_T(\theta)\bar{\pi}(\theta)\|\theta\|_{\mathcal{C}_{\mathcal{G}}}^{\lambda_m} \\ &\vdots \\ \pi_M(\theta|\mathcal{D}_T) &\propto L_T(\theta)\bar{\pi}(\theta)\|\theta\|_{\mathcal{C}_{\mathcal{G}}}^{\lambda_M} \end{aligned} \tag{19}$$

Here  $\|\theta\|_{\mathcal{C}_{\mathcal{G}}}^{\lambda}$  is a smooth penalty function, parameterized by  $\lambda \geq 0$ , measuring the magnitude of market-level constraint violations for a given  $\theta$ . We impose the following assumptions on  $\|\theta\|_{\mathcal{C}_{\mathcal{G}}}^{\lambda}$  to ensure that the sequence above starts and ends at the appropriate targets.

**Assumption 4.** (i)  $0 \leq \|\theta\|_{\mathcal{C}_{\mathcal{G}}}^{\lambda} \leq 1$  for all  $\lambda > 0$ ; (ii)  $\|\theta\|_{\mathcal{C}_{\mathcal{G}}}^{\lambda=0} = 1$ ; (iii)  $\|\theta\|_{\mathcal{C}_{\mathcal{G}}}^{\lambda} \rightarrow \mathbf{1}(\theta \in \mathcal{C}_{\mathcal{G}}(\Theta))$  as  $\lambda \rightarrow \infty$ .

In practice, we operationalize this penalty function as

$$\|\theta\|_{\mathcal{C}_{\mathcal{G}}}^{\lambda} = \prod_{g \in \mathcal{G}} 2\Phi(-\lambda c_g(\theta)), \tag{20}$$

where  $\Phi(\cdot)$  is the standard normal CDF and  $c_g(\theta)$  is a measure of constraint violations at  $g \in \mathcal{G}$ . This functional form is bounded between zero and one, equals one when  $\lambda = 0$ , and approaches  $\mathbf{1}(\theta \in \mathcal{C}_{\mathcal{G}}(\Theta))$  as  $\lambda \rightarrow \infty$ . Hence, all conditions of Assumption 4 are satisfied.

## 4.2 Sampling Algorithm

Algorithm 1 details our proposed sampling routine. In Step 1, we use MCMC to generate a set of draws (or “particles”) from the baseline model  $\bar{\pi}(\theta|\mathcal{D}_T)$ . For this step, any performant MCMC sampler will suffice. For instance, our implementation in `NPDemand.jl` uses the No-U-Turn Sampler

---

**Algorithm 1**

---

1. (MCMC) Generate an initial sample of particles  $\{\theta_{i(0)}\}_{i=1}^N$  from  $\bar{\pi}(\theta|\mathcal{D}_T) \propto L_T(\theta)\bar{\pi}(\theta)$ .
2. (SMC) First initialize weights  $W_{i(0)} = 1/N$ . Then for each  $m = 1, \dots, M$ , do the following to sample from the sequence of  $\lambda_m$ -penalized posteriors  $\pi_m(\theta|\mathcal{D}_T) \propto L_T(\theta)\bar{\pi}(\theta)\|\theta\|_{\mathcal{C}_G}^{\lambda_m}$ .
  - a) **Reweight:** For each particle  $i = 1, \dots, N$ , set  $W_{i(m)} = W_{i(m-1)}w_{i(m)}$  where

$$w_{i(m)} = \frac{\|\theta_{i(m-1)}\|_{\mathcal{C}_G}^{\lambda_m}}{\|\theta_{i(m-1)}\|_{\mathcal{C}_G}^{\lambda_{m-1}}}$$

and then normalize weights  $W_{i(m)} \leftarrow W_{i(m)} / \sum_{i'=1}^N W_{i'(m)}$ .

- b) **Resample:** If  $\text{ESS} = (\sum_{i=1}^N W_{i(m)}^2)^{-1} < \text{ESS}_{\min}$ 
        - i. Resample  $\{\theta_{1(m-1)}, \dots, \theta_{N(m-1)}\}$  with weights  $\{W_{1(m)}, \dots, W_{N(m)}\}$ ;
        - ii. Set  $W_{i(m)} = 1/N$ .
      - c) **Rejuvenate:** Sample new particles from a  $\pi_m(\theta|\mathcal{D}_T)$ -invariant transition kernel:  $\theta_{i(m)} \sim \mathcal{K}_m(\cdot|\theta_{i(m-1)})$ .
- 

(NUTS), which is an extension of the canonical Hamiltonian Monte Carlo (HMC) algorithm. Then in Step 2, we use SMC to push draws from  $\bar{\pi}(\theta|\mathcal{D}_T)$  through the sequence of  $\lambda_m$ -penalized posteriors  $\pi_1(\theta|\mathcal{D}_T), \dots, \pi_M(\theta|\mathcal{D}_T)$  to finally produce draws from the fully constrained target  $\pi^*(\theta|\mathcal{D}_T)$ . The SMC step executes three sub-steps described below.

Step 2(a) calculates a set of weights corresponding to the relative likelihood of each particle under a slightly more penalized model  $\pi_m(\theta|\mathcal{D}_T)$  relative to  $\pi_{m-1}(\theta|\mathcal{D}_T)$ . Values of particles that are closer to satisfying all constraints will be given higher weight. Then in Step 2(b), we resample the particles according to these derived weights, which filters out parameter values that are unlikely under  $\pi_m(\theta|\mathcal{D}_T)$ . This resampling step is only applied when the variation in weights is sufficiently high (and the number of unique particles starts to shrink as a result). We monitor the concentration of weights using the effective sample size (ESS), where  $\text{ESS} = (\sum_{i=1}^N W_{i(m)}^2)^{-1} \in [1, N]$ . When  $\text{ESS} = 1$  the distribution of weights is degenerate, and when  $\text{ESS} = N$  the distribution of weights is uniform. The threshold for resampling is typically taken to be  $\text{ESS}_{\min} = N/2$ .

Finally, in Step 2(c), we sample new particles from an MCMC kernel which is  $\pi_m$ -invariant. In other words, we use the existing  $N$  particles from  $\pi_m(\theta|\mathcal{D}_T)$  to generate new particles from this same distribution. This step mitigates particle degeneracy by increasing the number of unique particles and thus “boosts” the domain for the resampling step above. Constructing an efficient transition kernel is also straightforward. In our implementation, we use a random-walk MH algorithm where new particles are proposed from a Gaussian distribution centered around the existing particles and scaled by the sample covariance matrix of the existing particles (Chopin, 2002).

### 4.3 Convergence Guarantees

We now provide formal results on the convergence of Algorithm 1. As discussed above, the algorithm first relies on a standard implementation of MCMC to sample from  $\bar{\pi}(\theta|\mathcal{D}_T)$  and then uses SMC to transform those particles into representative draws from the sequence of posteriors converging to  $\pi^*(\theta|\mathcal{D}_T)$ . While both steps are included in Algorithm 1 for completeness, we limit our discussion of convergence to the novel SMC procedure in Step 2.

We start with a minimal set of assumptions to guarantee a well-behaved sequence of target posteriors and then state the main convergence result in Proposition 1.

**Assumption 5.**  $\mathcal{C}_{\mathcal{G}}(\Theta) \neq \emptyset$ .

**Assumption 6.**  $0 < \int L_T(\theta)\bar{\pi}(\theta)\|\theta\|_{\mathcal{C}_{\mathcal{G}}}^{\lambda} d\theta < \infty$  for all  $\lambda > 0$ .

**Assumption 7.**  $\mathbb{E}(\|\theta\|_{\mathcal{C}_{\mathcal{G}}}^{\lambda_m}/\|\theta\|_{\mathcal{C}_{\mathcal{G}}}^{\lambda_{m-1}}) < \infty$  for all  $m = 1, \dots, M$ .

**Proposition 1.** *Let  $\lambda_1, \dots, \lambda_M$  be an increasing sequence of penalties indexing the sequence of  $\lambda_m$ -penalized posteriors  $\pi_1(\theta|\mathcal{D}_T), \dots, \pi_M(\theta|\mathcal{D}_T)$  targeted in Algorithm 1. Additionally let  $\tilde{\pi}_m^{(N)}(\theta|\mathcal{D}_T) = \sum_{n=1}^N W_m^{(n)}\delta_{\theta^{(n)}}$  denote the empirical distribution of  $N$  particles at stage  $m = 1, \dots, M$ . Then under Assumptions 4, 5, 6, and 7, we have*

(i)  $\tilde{\pi}_m^{(N)}(\theta|\mathcal{D}_T) \xrightarrow{d} \pi_m(\theta|\mathcal{D}_T)$  as  $N \rightarrow \infty$ ;

(ii)  $\pi_m(\theta|\mathcal{D}_T) \rightarrow \pi^*(\theta|\mathcal{D}_T)$  as  $\lambda_m \rightarrow \infty$ .

*Proof.* Under Assumptions 5, 6, and 7 we can apply a law of large numbers to  $\tilde{\pi}_m^{(N)}(\theta|\mathcal{D}_T)$  to establish convergence to the stage- $m$  target  $\pi_m(\theta|\mathcal{D}_T)$ . This is a well-established result for various implementations of SMC (see, e.g., Crisan and Doucet, 2002; Chopin, 2004; Douc and Moulines, 2008; Beskos et al., 2016). Formally, for any bounded measurable function  $\varphi(\cdot)$ , we have

$$\sum_{n=1}^N W_m^{(n)}\varphi(\theta_m^{(n)}) \xrightarrow{a.s.} \mathbb{E}_{\pi_m}[\varphi(\theta)] \text{ as } N \rightarrow \infty \quad (21)$$

which implies that the empirical distribution of particles  $\tilde{\pi}_m^{(N)}(\theta|\mathcal{D}_T)$  converges weakly in distribution to  $\pi_m(\theta|\mathcal{D}_T)$ , establishing (i) for each posterior  $m = 1, \dots, M$ . Then (ii) immediately follows from Assumption 4. Together, (i) and (ii) produce the desired result.  $\square$

Finally, note that convergence of Algorithm 1 does not require the constraints to be correctly specified. We essentially only require that the entire sequence of  $\lambda_m$ -penalized posteriors, including the limiting posterior  $\pi^*(\theta|\mathcal{D}_T)$ , has positive mass. Hence, the only time convergence will fail is when the implemented constraint yields a degenerate stage with undefined weights, in which case the particle system will collapse.

## 4.4 Practical Considerations

The main practical consideration in Algorithm 1 above is how to best choose the penalty sequence  $\lambda_1, \dots, \lambda_M$ . Intuitively, a longer sequence with shorter intervals makes the transitions between models easier in the sense of preserving the dispersion across particle weights. However, shorter intervals may also be wasteful and create an unnecessary computational bottleneck. Conversely, shorter sequences with larger intervals can pose challenges to the reweighting and resampling steps by reducing the variation among particle weights, leading to severe particle degeneracy.

We can then think of optimizing the sequence according to the ESS, which represents a measure of variation in particle weights. That is, given  $\lambda_m$ , make  $\lambda_{m+1}$  as large as possible subject to a minimum ESS threshold. This is the idea of the adaptive SMC approach popularized by [Jasra et al. \(2011\)](#), which we follow in our implementation. We also add a stopping rule which terminates the loop over models once the share of constraint violations falls below a prespecified threshold.

## 4.5 Implementation in Julia: `NPDemand.jl`

We provide an accompanying Julia package (`NPDemand.jl`) to estimate our quasi-Bayes NPD model using the sampling algorithm outlined above. The package is designed with a simple interface that executes heavy computation with minimal user overhead. For example, estimation can be carried out in only two function calls. The first function call to `define_problem()` defines a problem, similar to the PyBLP problem class ([Conlon and Gortmaker, 2020](#)), which initializes the data, model, and estimator. The second function call to `estimate!()` executes the desired estimation procedure and stores results inside the problem.

Our implementation of Algorithm 1 is built on top of `Turing.jl` ([Ge et al., 2018](#)), Julia’s state-of-the-art probabilistic programming language used for advanced Bayesian computation. We also provide functionality to estimate NPD using classical (GMM) methods relying on `JuMP.jl` ([Lubin et al., 2023](#)) as a back-end solver. More details on the package and its functionality are given in [Appendix B](#) as well as in our package documentation on [GitHub](#).

## 5 Monte Carlo Simulations

In this section, we run a series of Monte Carlo simulations to evaluate the performance of QBNPD relative to constrained and unconstrained GMM estimators. We evaluate two dimensions of model performance: (1) the statistical accuracy of estimated price elasticities, and (2) the share of markets where the estimated demand functions align with the desired economic constraints. The two dimensions are related in that they both reveal information about how well an estimator recovers the system of true demand functions, but are distinct in that only the violations reveal information about “economic fit.”

## 5.1 Setup

We simulate data from two data-generating processes (DGPs), which are differentiated by their set of admissible substitution patterns. Both assume the existence of a product-specific linear index:

$$\delta_{jt} = \alpha p_{jt} + \beta x_{jt} + \xi_{jt}. \quad (22)$$

Here  $p_{jt}$  is an endogenous price,  $x_{jt}$  is an exogenous product characteristic, and  $\xi_{jt}$  is an unobserved demand shock. Throughout our simulations, we let  $\xi_{jt} \sim N(0, 1)$ ,  $x_{jt} \sim U(0, 2)$ ,  $\beta = 1$ , and  $\alpha = -1$ . We generate prices through the equation  $p_{jt} = 2z_{jt} + w_{jt} + \xi_{jt}$ , where  $z_{jt} \sim N(0, 1)$  and  $w_{jt} \sim U(0, 0.1)$ . We use  $z_{jt}$  as an excluded instrument in estimation.

**Logit DGP** The first DGP is a simple aggregate logit model where consumers exhibit unit demand across an assortment of  $J$  substitutable goods. Indirect utility is specified as  $u_{ijt} = \delta_{jt} + \varepsilon_{ijt}$ . Under the assumption of type I extreme value errors, aggregate market shares can be expressed as:

$$s_{jt} = \frac{\exp(\delta_{jt})}{1 + \sum_{k=1}^J \exp(\delta_{kt})}. \quad (23)$$

**Complements DGP** Our second DGP is based on a flexible functional form for aggregate demand. We assume that the  $J$  goods are partitioned into two groups which we label as  $G_j$  and  $G_k$ .  $G_j$  represents the set of product indices belonging to the same group as product  $j$ , but excluding  $j$ .  $G_k$  represents the set of product indices belonging to other group. For example, if we partition  $J = 4$  goods as  $\{\{1, 2\}, \{3, 4\}\}$ , then for  $j = 1$  we have  $G_j = \{2\}$  and  $G_k = \{3, 4\}$ . Market quantities are then given by:

$$q_{jt} = \exp \left( \delta_{jt} + \frac{\gamma_{\text{own}}}{|G_j|} \sum_{j' \in G_j} \delta_{j't} + \frac{\gamma_{\text{other}}}{|G_k|} \sum_{j' \in G_k} \delta_{j't} \right) \quad (24)$$

which we normalize by an assumed market size to produce product shares. We set  $\gamma_{\text{own}} = -0.25$  and  $\gamma_{\text{other}} = 0.25$  which ensures that products within the same group are substitutes and products in different groups are complements.

## 5.2 Econometric NPD Specifications

In estimation, we specify a model of the following form:

$$s_{jt} = \sigma_j(\delta_{1t}, \dots, \delta_{Jt}) \quad (25)$$

$$\delta_{jt} = \alpha p_{jt} + \beta x_{jt} + \xi_{jt} \quad (26)$$

and approximate  $\sigma^{-1} = (\sigma_1^{-1}, \dots, \sigma_J^{-1})$  as in (10) using a tensor product of  $K$ -order univariate Bernstein basis functions:

$$\sigma_j^{-1}(\mathbf{s}; \theta_j) = \sum_{k_1=0}^K \cdots \sum_{k_J=0}^K \theta_{jk_1 \dots k_J} \prod_{\ell=1}^J \phi_{jk_\ell}(s_\ell). \quad (27)$$

Relative to (10), we make the simplifying assumption that the order of each univariate basis function is the same:  $K_j = K$  for all  $j = 1, \dots, J$ .

We estimate a total of three specifications: (1) unconstrained GMM, (2) constrained GMM, and (3) constrained quasi-Bayes. In both unconstrained and constrained specifications, we impose some symmetry of the demand functions. In the logit DGP, we impose full exchangeability across all  $J$  goods while in the complements DGP, we only impose exchangeability within, but not across, each group of products. For the constrained logit specifications, we impose own-good monotonicity (negative own-price effects), cross-good substitution, and diagonal dominance. For the constrained complements specifications, we impose own-good monotonicity, within-group cross-good substitution, and across-group cross-good complements. The latter is meant to mimic a setting of multicategory demand, where products are substitutes within a category but complements across categories. We choose the grid over which constraints are enforced to be the set of unique product-market shares observed in the data. We execute the analysis using our Julia package `NPDemand.jl`.

### 5.3 Results

In our first set of simulations, we examine the finite sample performance of different NPD estimators. For this exercise, we only use the logit DGP with  $J = 2$  products and consider  $T \in \{50, 100, 500, 1000, 5000\}$ . We simulate 100 data sets for each value of  $T$ . In estimation, we also vary the order of  $K \in \{2, 3, 4\}$ . The results are shown across Figures 2 and 3. Figure 2 reports the statistical accuracy of the estimated price elasticities and plots the median absolute deviation (MAD) as a function of  $T$ . Each line represents the MAD averaged across 100 simulations, and the shaded regions correspond to the 10th and 90th percentiles of the MAD distribution. We report estimation accuracy for all elasticities combined (top row), as well as the accuracy for own and cross elasticities separately (middle and bottom rows).

The main takeaway from Figure 2 is that constraints improve finite sample performance. With enough data, unconstrained estimators can learn the shape of the true demand functions, but the rate of convergence is prohibitively slow. This is evidenced in our simulations by the flattening of the solid blue line as the dimension of the sieve grows (moving across columns). We find that both constrained estimators exhibit lower average MAD for every  $T$  and  $K$ —often by a sizable margin. This result is consistent with the previous literature showing improved rates from monotonicity restrictions in NPIV (e.g., Chetverikov and Wilhelm, 2017). The practical implication is that constraints allow you to do more with less. For instance, in the case of  $K = 2$ , a constrained estimator achieves the same statistical performance as an unconstrained estimator that uses 10 times the amount of data. We find that quasi-Bayes provides further finite sample improvements over the constrained GMM specification, with the largest gains coming for large  $K$  and small  $T$ . Quasi-Bayes also reduces the variability of the estimates and produces a tighter range of MAD values across simulations.

Figure 3 offers some economic rationale for the improved statistical performance offered by our quasi-Bayes approach. We plot the distribution of market-level violations across data realizations.

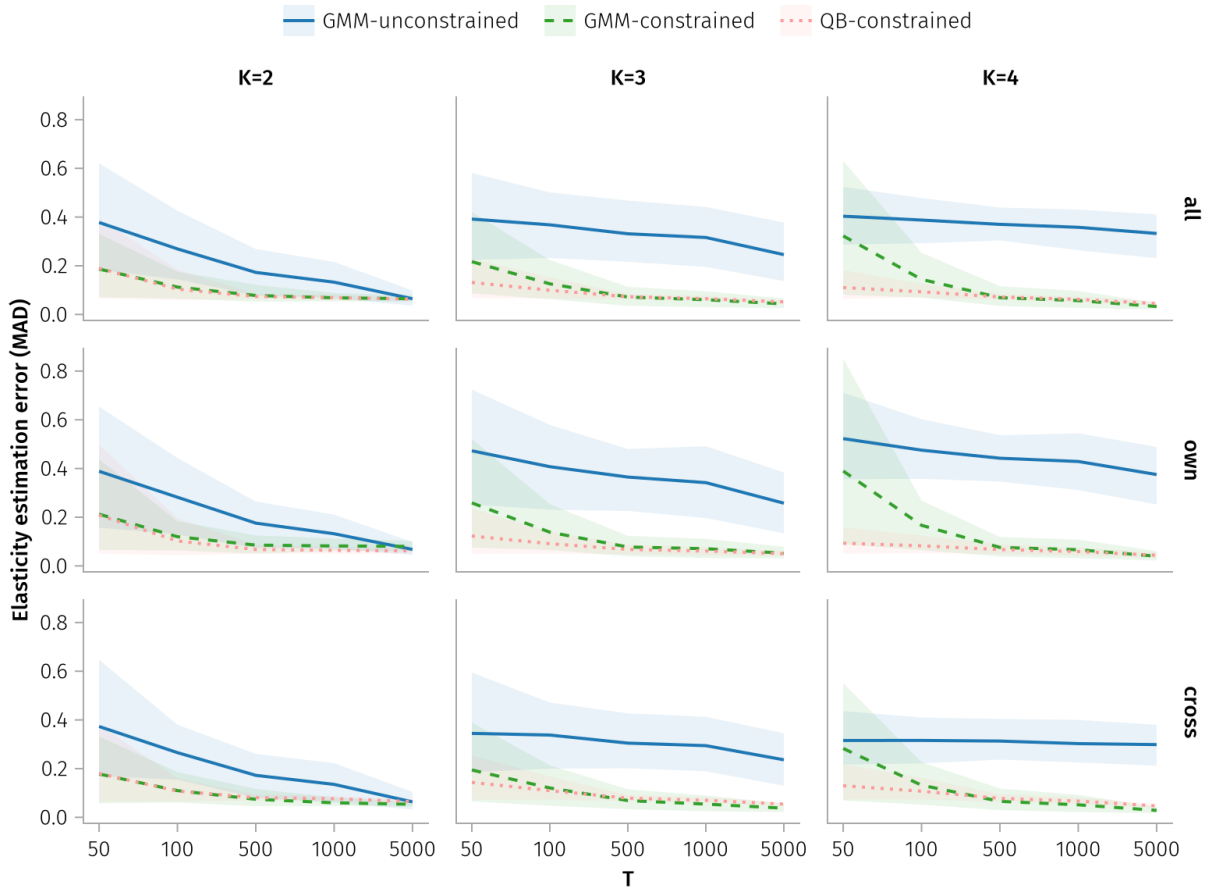


Figure 2: An example showing that constraints improve the finite sample performance of NPD estimation. Each line represents the median absolute deviation (MAD) of price elasticities averaged across 100 simulations for a logit DGP with  $J = 2$  products. The shaded regions represent the 10th and 90th percentiles of the distribution of MADs across simulation runs. The columns correspond to different orders  $K$  of each univariate Bernstein polynomial in the tensor product sieve approximation and the rows correspond to different sets of price elasticities.

The columns again correspond to the order of each univariate Bernstein polynomial ( $K$ ) in the overall tensor product sieve approximation. The rows now correspond to the different constraints imposed in estimation: own-good monotonicity, cross-good substitution, and diagonal dominance. The last row checks for violations of any of the three constraints. We find a strong correlation between the accuracy of estimated price elasticities in Figure 2 and the alignment of the economic properties of the true model shown in Figure 3. With more data, an unconstrained estimator will eventually learn that demand functions are monotonically decreasing in the own price and increasing in the cross price, and this will lead to better estimated elasticities. However, the rate at which these shape constraints are learned is very slow. For example, even the own-good monotonicity constraint—which is the constraint for which there is the strongest signal in the data—is violated more than a quarter of the time on average whenever  $K > 2$ . We therefore stand to gain by

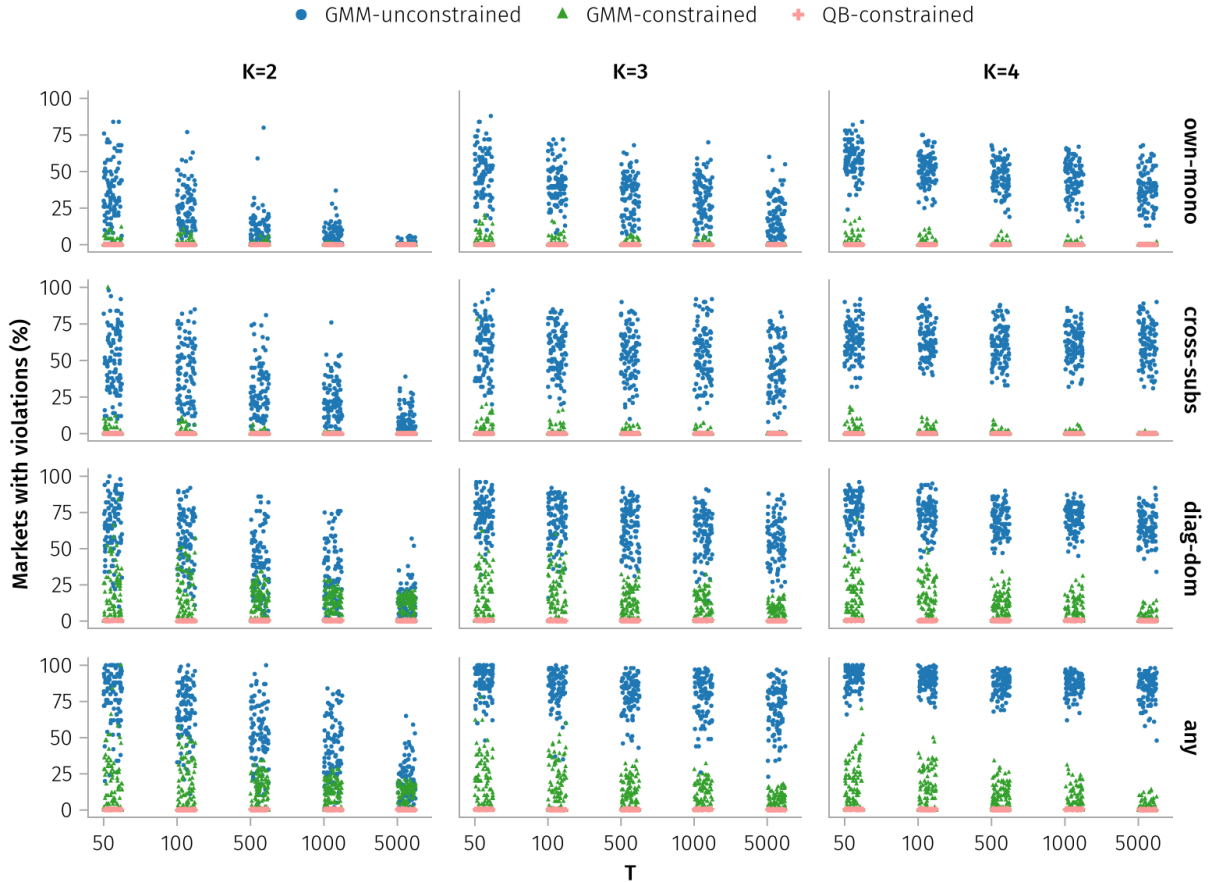


Figure 3: The distribution of constraint violations across estimator specifications. Each point represents the outcome from one of 100 simulation runs for a given sample size  $T$  from a logit DGP with  $J = 2$  products. Points are jittered about the x-axis to aid in visualization. The columns correspond to different orders  $K$  of each univariate Bernstein polynomial in the tensor product sieve approximation and the rows correspond to different constraints.

enforcing such constraints in estimation.

Similar to the results on estimation accuracy, we find that our quasi-Bayes approach outperforms the constrained GMM estimator in terms of violations. Incorporating linear restrictions in the GMM problem leads to better alignment with own and cross-good monotonicity constraints, which are satisfied in more than 90% of markets. Diagonal dominance is satisfied less often. In contrast, our approach to enforcing constraints through the prior ensures that they hold across all markets.

Next, we move to a slightly higher-dimensional estimation problem and examine the performance across the two different DGPs outlined above. For each DGP, we simulate 25 data sets with  $J = 4$  products and  $T = 500$  markets. We take the same three candidate estimators to the data with  $K \in \{2, 3\}$ . The results are reported in Table 2. We report constraint violations across markets as well as three different elasticity estimation error metrics: mean squared error (MSE), mean absolute error (MAE), and median absolute deviation (MAD). When reporting the MSE, we

Table 2: Simulation Results Across DGPs

	$K = 2$			$K = 3$		
	GMM-U	GMM-C	QB-C	GMM-U	GMM-C	QB-C
<b>(A) Logit</b>						
Elasticity Estimation Error						
- MSE	8.63	1.02	<b>0.13</b>	10.28	1.20	0.15
- MAE	73.87	0.92	<b>0.14</b>	25.62	0.85	0.16
- MAD	0.23	0.06	<b>0.04</b>	0.32	0.07	0.04
Markets with Violations (%)						
- own-good monotonicity	30	2	0	56	3	0
- cross-good substitutes	92	38	0	97	45	0
- diagonal dominance	82	32	0	92	31	0
- any	96	50	0	99	53	0
<b>(B) Complements</b>						
Elasticity Estimation Error						
- MSE	11.90	10.74	<b>2.48</b>	21.2	25.8	3.47
- MAE	10.90	2.97	<b>0.58</b>	3.88	7.88	0.76
- MAD	0.29	0.28	<b>0.08</b>	0.39	0.41	0.16
Markets with Violations (%)						
- own-good monotonicity	38	13	0	63	24	0
- within-group substitutes	67	64	0	80	73	0
- across-group complements	95	90	1	98	95	6
- any	99	96	1	100	19	6

Notes: (1)  $J = 4$  and  $T = 500$ . (2)  $K$  represents the order of each univariate Bernstein polynomial in the tensor product sieve approximation. (3) Estimation error metrics are averaged across 25 simulations. (4) The within and across group constraints used in the complements DGP do not have linear representations so they are omitted from the constrained GMM specification. (5) We cap the number of SMC iterations to save on total run time which contributes to the small residual share of violations for the QB-C specification in the complements DGP.

trim the top 1% of elasticity estimates due to extreme outliers of the unconstrained specification, in particular. The metrics reported in Table 2 are averaged across the 25 data replications.

We again find large performance gains from incorporating constraints in NPD estimation. When  $K = 2$ , the unconstrained GMM estimator violates the simplest own-good monotonicity constraint in more than 30% of markets in the logit DGP and 38% of markets in the complements DGP. Constraint violations tend to increase as the dimension of the sieve space grows. Incorporating linear restrictions into the GMM problem reduces violations and improves the accuracy of the estimated elasticities. For example, in the logit DGP, violations of own-good monotonicity drop from 30% to 2% when  $K = 2$ , and these reductions are accompanied by an eight-fold reduction in the MSE and four-fold reduction to the MAD. Despite the gains, there is still sizable headroom given that a large share of markets continue to violate some of the key shape constraints exhibited by each DGP, despite the inclusion of linear restrictions in estimation. As we push our estimates in the direction of the constraints holding everywhere, the estimation accuracy continues to improve. For the logit DGP ( $K = 2$ ), the MSE drops from 1.02 to 0.13 and the MAD drops from 0.06 to

0.04. The gains are also large for the complements DGP, where the MSE is reduced by four-fold and the MAD is reduced by more than three-fold.

In summary, we have shown through two simulation exercises that constraints improve both statistical and economic properties of estimated NPD systems. We also find that our quasi-Bayes estimator offers superior finite sample performance relative to a constrained GMM estimator, which can in part be attributed to differences in the treatment of constraints. That is, our approach ensures that the desired shape constraints hold on the target demand functions, which cannot be guaranteed using linear restrictions alone. We show that our approach is also effective at enforcing constraints which do not admit a linear restriction, such as constraints on cross-group substitution that commonly arise in multicategory demand settings. In the next section we demonstrate how many of these differences play out in retail scanner data.

## 6 Application to Grocery Retail Data

In this section, we provide further empirical evidence for the value of economic constraints in nonparametric demand estimation. We use grocery retail data and estimate demand for several categories of consumer goods. Specifically, we construct 12 separate data sets—each coming from a different product category with its own unique characteristics, such as the number of products, the number of subcategories, and the nature of substitution across products/subcategories. We first show that across all 12 product categories, unconstrained NPD estimators consistently fail to recover economically valid demand systems. We then zoom in and compare unconstrained and constrained NPD estimates on four representative categories. We find that constraints provide much needed economic discipline and that quasi-Bayes methods more effectively enforce such constraints. In doing so, QBNPD provides estimates of substitution patterns that are more reasonable in both sign and magnitude.

### 6.1 Data

We use point-of-sale transaction data from a regional supermarket chain in the United States with nearly 500 stores spanning five states.<sup>7</sup> Our raw data consist of every transaction made in each store between January 1, 2015 and December 31, 2016. We observe the quantity, price, and promotional activity for each UPC scanned at the register. We also observe wholesale prices which we use as excluded price instruments in estimation.

We construct data sets for several product categories. Examples of product categories include Frozen Ice Cream, Mayonnaise, Snacks, and Jams, Jellies, & Peanut Butter. Within each category, we select all non-fringe brands from the most popular 1-2 subcategories. A list of all product categories and subcategories is provided in [Table 3](#). We aggregate UPCs across pack sizes and types to define products at the brand-subcategory level, and define markets at the ZIP3-week level.

---

<sup>7</sup>The data are provided by DecaData (<https://decadata.io>).

Table 3: Description of Product Categories

Category	Subcategories (number of products)
1 Baking Goods	Cake Mix (5), Frosting (3)
2 Beer	Premium (4), Economy (3)
3 Butter, Margarine, & Spreads	Margarine and Spreads (4), Butter (2)
4 Cookies	Regular (4)
5 Fish Canned	Chunk Light Tuna (3), Solid White Tuna (2)
6 Franks	Meat (2), Beef (3)
7 Frozen Ice Cream	Premium (4)
8 Frozen Pizza	Value (1), Core (4)
9 Jams, Jellies, & Peanut Butter	Peanut Butter (4), Jams, Jellies, Preserves (2)
10 Ketchup	Regular (2)
11 Mayonnaise	Regular (3), Light (3)
12 Snacks	Potato Chips (4), Tortilla Chips (4)

In total, each data set contains  $T = 5,565$  markets and includes anywhere from  $J = 2$  to 8 products. Additional data processing details are provided in [Appendix C](#).

## 6.2 Econometric NPD Specifications

We specify the market share for product  $j = 1, \dots, J$  in each market  $t = 1, \dots, T$  as

$$s_{jt} = \sigma_j(\delta_1(\mathbf{x}_{1t}, p_{1t}, \xi_{1t}), \dots, \delta_J(\mathbf{x}_{Jt}, p_{Jt}, \xi_{Jt})) \quad (28)$$

where  $\mathbf{x}_{jt}$  is a vector of product-market characteristics,  $p_{jt}$  is price, and  $\xi_{jmt}$  is an unobserved demand shock. We model the product indices  $\delta_j(\cdot)$  as linear functions of product characteristics:

$$\delta_j(\mathbf{x}_{jt}, p_{jt}, \xi_{jt}) = \alpha p_{jmt} + \beta x_{jt}^{\text{promo}} + \tilde{\mathbf{x}}_{jt}^{\text{FE}'} \gamma + \xi_{jt} \quad (29)$$

where  $x_{jt}^{\text{promo}}$  is a feature promotion variable and  $\tilde{\mathbf{x}}_{jt}^{\text{FE}}$  is a vector of the remaining product-market characteristics including year, quarter, state, holiday, and product dummies. We normalize  $\alpha = 1$  and drop one level from each variable contained in  $\tilde{\mathbf{x}}_{jt}^{\text{FE}}$ .

Note that relative to (2), we have made the additional assumption that prices enter via the index. While this specification is more restrictive, we believe this is a reasonable starting point for two reasons. First, the model remains flexible in that we have not placed any restrictions on the ways that the indices interact in the production of demand. Second, as we will show below, the estimated demand functions are sufficiently noisy and continue to violate constraints *despite* the added structure. The problem will become worse as the assumptions are relaxed, which only further motivates our regularizing quasi-Bayes approach.

As described in Section 2, we can invert  $\sigma_1, \dots, \sigma_J$  to get the following system of equations:

$$p_{jt} = \sigma_j^{-1}(s_{1t}, \dots, s_{Jt}) - \beta x_{jt}^{\text{promo}} - \tilde{\mathbf{x}}_{jt}^{\text{FE}'} \gamma - \xi_{jt}. \quad (30)$$

Each inverse demand equation has two sets of endogenous variables: shares and prices. We use

feature promotions as included instruments and wholesale prices as excluded instruments. [Table D1](#) reports  $R^2$  and F-test statistics from the first-stage price regressions and shows that the instruments are relevant.

### 6.3 Estimation Details

We approximate  $\sigma^{-1}$  using the same tensor product sieve as in [\(27\)](#) where each univariate Bernstein basis function has order  $K_j = K$ . As in any nonparametric analysis, there is a question of how to choose tuning parameters like  $K$  which govern the dimension of the approximation. In practice, researchers often rely on cross-validation procedures and choose the value of the tuning parameter which minimizes the mean-squared error (MSE). However, cross-validation produces a biased estimate of the MSE in models with endogeneity and is therefore an invalid procedure for selecting  $K$  in our analysis ([Chen et al., 2024](#)). Given that our goal is to show the empirical value of economic constraints, we instead take several specifications to the data ( $K \in \{2, 3\}$ ) and simply compare the resulting elasticity estimates in terms of their sign and magnitude. As we will show, unconstrained GMM estimators already tend to exhibit high variance when  $K = 2$ , so our approach to regularization should become more valuable as the dimension of the sieve space grows.

We consider the same three candidate estimators as in our simulation experiments: (1) unconstrained GMM, (2) constrained GMM, and (3) constrained quasi-Bayes. In both unconstrained and constrained specifications, we assume that demand is exchangeable within but not across product subcategories. For categories exclusively comprised of substitutes, we impose the following three constraints: own-good monotonicity (negative own-price effects), cross-good substitution (positive cross-price effects), and diagonal dominance. For categories with a complementary subcategories, we impose own-good monotonicity, cross-good substitution within subcategory, and cross-good complementarity across subcategories. In our quasi-Bayes approach, we incorporate linear restrictions via a reparameterization and then specify priors of the form in [\(14\)](#), where the constrained set  $\mathcal{C}_{\mathcal{G}}(\Theta)$  is defined over a grid of “in-sample” shares. We also specify Gaussian baseline priors  $\bar{\pi}(\theta)$  with a mean of zero and a variance of 100.

The analysis is carried out using our Julia package `NPDemand.jl`. For the quasi-Bayes specification, we first estimate a model using only linear restrictions that can be encoded using reparameterizations. We sample from this model’s posterior distribution using the No-U-Turn Sampler (NUTS) ([Hoffman and Gelman, 2014](#)) with 1000 adaption steps and a target accept ratio of 0.65. We run the sampler for 2,500 iterations and discard the first 20% of draws as burn-in. We then implement the constrained SMC method outlined in [Section 4.2](#). For this, we follow the adaptive approach popularized by [Jasra et al. \(2011\)](#) which solves for an optimal sequence of penalty values  $\lambda$  subject to a minimum ESS threshold, which we set to 500. We stop the sampler whenever the share of markets violating any constraints falls below 0.1%. For the final particle rejuvenation step in [Algorithm 1](#), we use 20 iterations of a random-walk MH algorithm with Gaussian proposals centered at the current particles and scaled by the sample covariance matrix of the current particles.

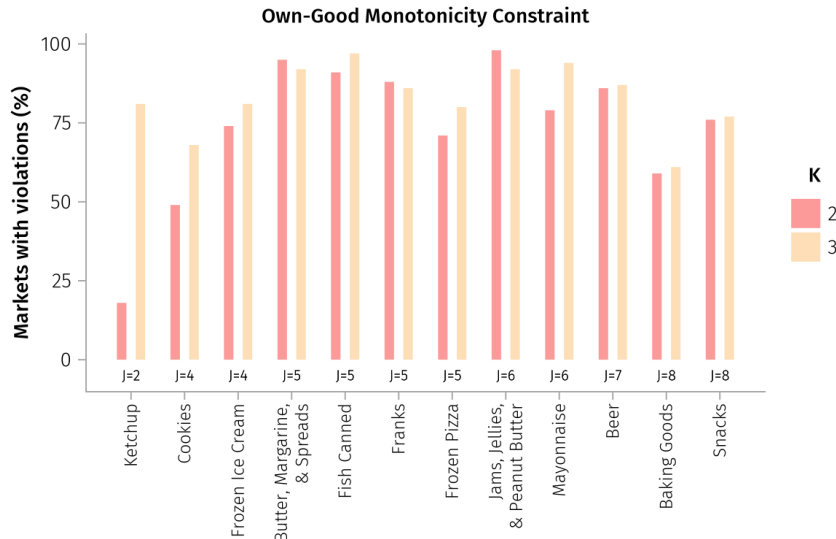


Figure 4: Share of markets violating own-good monotonicity when NPD is estimated via GMM without any economic shape constraints. Product categories are ordered by  $J$ .  $K$  represents the degree of each univariate Bernstein polynomial in the tensor product sieve approximation.

## 6.4 Results

### Constraint violations

We start our analysis by exploring the performance of an unconstrained GMM estimator across a large set of product categories. In each category, we compute the share of markets violating own-good monotonicity which should be the easiest of the shape constraints to satisfy. The results are displayed in Figure 4. We arrange the 12 product categories in ascending order by  $J$ . Ketchup is the smallest product category with  $J = 2$  products and Baking Goods and Snacks are the largest categories with  $J = 8$  products.

There are two takeaways from Figure 4. The first is that the share of violations is appreciably large across all categories. For instance, whenever  $J > 2$ , the unconstrained estimator produces upward-sloping demand curves in more than half of the markets in our data. The second takeaway is that the share of violations tends to increase as the dimension of the estimation problem grows. Since  $T$  is fixed in our analysis, we are asking more of the data as we increase either  $J$  or  $K$ . This in turn makes it harder for an unconstrained estimator to learn economic structure. Together, the results highlight the practical challenges of NPD estimation in finite samples and call for judicious constraints to ensure that the estimated demand system adheres to economic principles.

We now move to a smaller set of representative categories and report more detailed results for the three candidate estimators. For ease of exposition, we only discuss results from the  $K = 2$  specification. Estimates from the  $K = 3$  specification are qualitatively similar and the insights do not change. We report results for the following four categories: Ketchup ( $J = 2$ ), Frozen Ice Cream ( $J = 4$ ), Fish Canned ( $J = 5$ ), and Jams, Jellies, & Peanut Butter ( $J = 6$ ). We choose these

Table 4: Share of Markets Violating Constraints

	GMM-unconstrained	GMM-constrained	QB-constrained
<i>Ketchup</i>			
- own-good monotonicity	18	0	0
- cross-good substitutes	32	0	0
- diagonal dominance	87	25	0
- any	92	25	0
<i>Frozen Ice Cream</i>			
- own-good monotonicity	74	94	0
- cross-good substitutes	100	99	0
- diagonal dominance	95	97	0
- any	100	99	0
<i>Fish Canned</i>			
- own-good monotonicity	91	38	0
- cross-good substitutes	100	98	0
- diagonal dominance	100	69	0
- any	100	98	0
<i>Jams, Jellies, &amp; Peanut Butter</i>			
- own-good monotonicity	98	97	0
- within-group substitutes	92	93	0
- across-group complements	100	100	0
- any	100	100	0

Notes: Constraints are checked across all estimated market-level demand Jacobian matrices. When checking sign restrictions, we say that a constraint is violated if any of the element-wise sign restrictions are violated.

categories because they span several important dimensions, including the number of products, the number of subcategories, and the nature of cross-subcategory substitution. For example, Ketchup and Frozen Ice Cream each have only one subcategory while Fish Canned and Jams, Jellies, & Peanut Butter each have two. In Fish Canned, the two subcategories (“Chunk Light” vs. “Solid White”) are substitutes while in Jams, Jellies, & Peanut Butter, the two subcategories (“Jams, Jellies, Preserves” and “Peanut Butter”) are complements. Thus, across the four categories we can highlight the flexibility of NPD and the ability to accommodate flexible cross-product and cross-subcategory substitution, as well as the value of our quasi-Bayes approach to disciplining high-dimensional NPD estimators.

Table 4 reports the share of markets violating the desired economic constraints for each of the four categories and each of the three NPD specifications. We report violations of each constraint separately, but also include a row that reports the share of markets violating any of the desired constraints. We find that an unconstrained GMM specification struggles to recover economic structure. In all four categories, the share of markets violating any of the desired constraints exceeds 90%. Incorporating (necessary) linear restrictions into the GMM problem tends to help, but leaves an appreciable share of markets with demand functions violating economic theory. The residual violations are especially large for the bigger demand systems. For instance, linear restrictions are

sufficient for enforcing own-good monotonicity in the  $J = 2$  system for Ketchup, but produce (at least one) upward-sloping demand curve in 94% of markets in the Frozen Ice Cream category, 38% of markets in the Fish Canned category, and 97% of markets in the Jams, Jellies, & Peanut Butter category. Our quasi-Bayes approach closes this gap and produces estimated demand functions which satisfy the desired constraints across all markets.

## Price elasticities

Next, we move on to investigate the impact of constraints on the sign and magnitude of estimated elasticities. [Figure 5](#) plots the marginal distributions of each specification’s estimated elasticities across all products and markets. The rows correspond to product categories and the columns correspond to the three types of price elasticities: own elasticities, cross (within-subcategory) elasticities, and cross (across subcategory) elasticities. Note that Fish Canned and Jams, Jellies, & Peanut Butter are the only two categories with multiple subcategories, and thus the only two categories where the notion of across-subcategory substitution is relevant. In total, there are 94,605 own-price elasticities, 200,340 cross-price (within-subcategory) elasticities, and 155,820 cross-price (across-subcategory) elasticities.

We find that that constraints deliver useful regularization.<sup>8</sup> When left unconstrained, GMM estimators are extremely high variance. Across all categories, 21% of product-market own-price elasticities are larger than 50 in magnitude and 45% are incorrectly signed. Incorporating linear restrictions tends to shift the distribution of estimates in the right direction, corroborating the results of [Table 4](#). However, the distribution of estimates remains widely dispersed and the estimator cannot guarantee that the monotonicity constraints hold across all markets. For instance, 28% of all own-price elasticities remain incorrectly signed. In comparison to GMM, the quasi-Bayes specification effectively shrinks the tails and shifts the location to ensure that all elasticities are correctly signed.

[Figure 6](#) provides a different view on the dispersion of elasticity estimates by plotting own-price elasticities against prices (top row) and shares (bottom row). We find that the largest elasticities (in magnitude) tend to come from markets with very small market shares, which helps rationalize the wide tails of the distributions reported in [Figure 5](#). More generally, we find that the magnitude of own elasticities tends to decrease in shares and increase in prices, as one might expect. However, both GMM estimators are noticeably high-variance. Unconstrained estimates are scattered across the  $(-10^4, 10^4)$  interval, becoming larger in magnitude in both positive and negative directions as shares get smaller. Incorporating linear constraints does correct the sign of some estimates, but the distribution of estimates still spans the  $(-10^4, 10^4)$  interval. In contrast, our quasi-Bayes estimator produces a much tighter range of estimates that are correctly signed across all markets.

Taken together, [Figures 5](#) and [6](#) provide visual evidence that GMM estimators are prohibitively

---

<sup>8</sup>[Figure D1](#) plots the joint distribution of own-price elasticity estimates for each possible pair of estimators. We also separately report the constrained quasi-Bayes estimates before SMC (where only linear restrictions are imposed) and after SMC to showcase the incremental value of economic regularization induced by the hard constraint in [\(14\)](#).

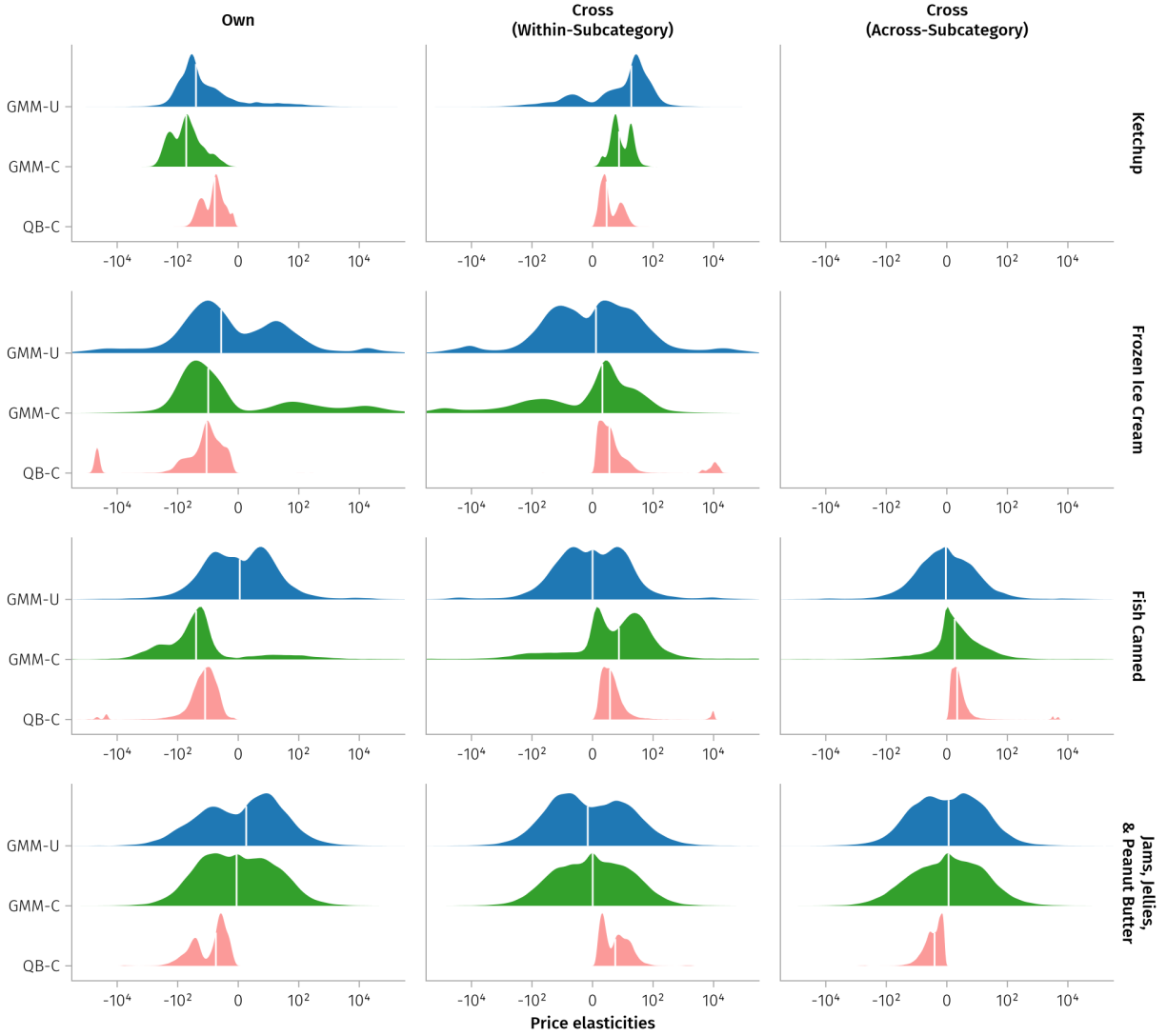


Figure 5: Marginal distributions of all product-market price elasticities across specifications. The columns correspond to the different types of price elasticities and the rows are different product categories. The Ketchup and Frozen Ice Cream categories only contain one subcategory and therefore only have within-subcategory cross-price elasticities. Subcategories are substitutes in the Fish Canned category and complements in the Jams, Jellies, & Peanut Butter category. White vertical lines represent the median elasticity across all products and markets.

high-variance, producing product-market elasticity estimates that are often incorrectly signed and orders of magnitude larger than one may expect. In [Appendix D](#), we provide evidence that the same insights hold when elasticities are aggregated over markets and reported at the product level. [Table D2](#) reports the median own-price estimated elasticity for each of the 17 products in our data. We find that the unconstrained GMM estimator produces incorrectly signed estimates for about half of the products. Enforcing linear restrictions in the GMM problem leads to negative own-price effects for 15 of the 17 total products, and corrects the sign for seven out of the nine

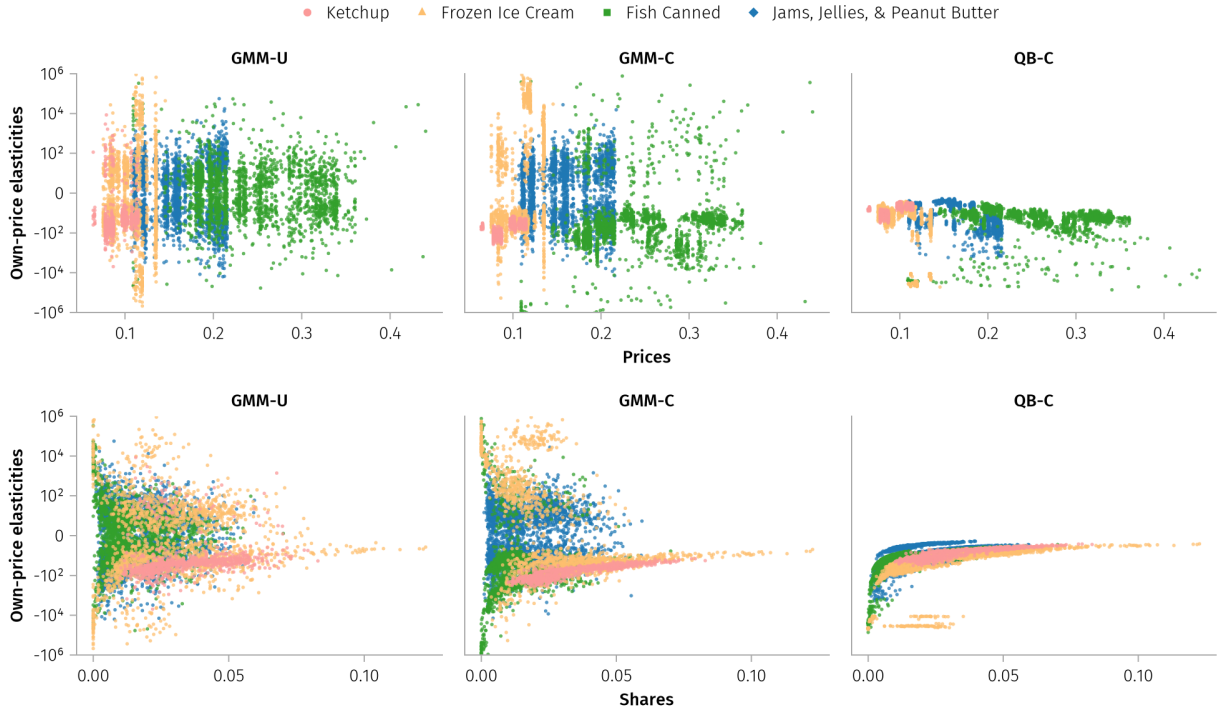


Figure 6: Own-price elasticities plotted against prices and shares. We average elasticity estimates up to the product-ZIP3-year-quarter level to aid in visualization.

incorrectly signed estimates from the unconstrained specification. However, many of the linearly constrained estimates are still large in magnitude (e.g., two point estimates are larger than 100 in absolute value) and also exhibit large sampling uncertainty. Quasi-Bayes estimates provide additional improvements, ensuring that all point estimates are negative.<sup>9</sup>

## Diversion ratios

For our last set of empirical results, we present estimates of diversion ratios which offer a complementary perspective on substitution patterns relative to the estimated price elasticities reported above. Diversion ratios are defined as:

$$D_{jk} = \frac{\partial \sigma_k / \partial p_j}{|\partial \sigma_j / \partial p_j|} \quad (31)$$

<sup>9</sup>It is also worth noting that some of the product-level NPD elasticities reported in Table D2 are larger in magnitude relative to estimates reported in the literature from similar CPG markets but based on more restrictive logit-like specifications of demand. This is perhaps not surprising given that the logit functional form also regularizes, and so paring back functional form could indeed lead to a wider range of admissible estimates on any given data set. For example, Compiani (2022) also finds that NPD produces larger elasticities in magnitude than a standard mixed logit model. While we do not find any systematic differences in the magnitude of elasticity estimates across GMM vs. quasi-Bayes specifications, some quasi-Bayes estimates are larger than GMM estimates. This is a consequence of regularization and the way we operationalize constraints—i.e., we require the constraints to hold across all markets in our data. If the researcher only requires constraints to be satisfied over a smaller set of points, such as median elasticity values, then the amount of regularization could change. We return to discuss this point in Section 7.

Table 5: Diversion Ratio Estimates Across Products and Markets

	1st	10th	50th	90th	99th
<i>Ketchup</i>					
- GMM-unconstrained	-4.78	0.00	0.63	2.21	16.98
- GMM-constrained	0.00	0.01	0.22	1.16	8.94
- QB-constrained	0.31	0.32	0.36	0.43	0.61
<i>Frozen Ice Cream</i>					
- GMM-unconstrained	-29.39	-2.17	0.14	3.13	33.58
- GMM-constrained	-15.77	-0.88	0.13	1.98	13.41
- QB-constrained	0.20	0.22	0.23	0.25	0.27
<i>Fish Canned</i>					
- GMM-unconstrained	-36.93	-2.77	0.10	3.29	35.73
- GMM-constrained	-8.43	-0.24	0.03	0.94	10.22
- QB-constrained	0.07	0.07	0.15	0.22	0.25
<i>Jams, Jellies, &amp; Peanut Butter</i>					
- GMM-unconstrained	-36.97	-2.99	0.04	3.00	36.38
- GMM-constrained	-58.27	-4.22	0.03	4.55	61.62
- QB-constrained	-0.57	-0.51	-0.15	0.35	1.29

Notes: The table reports percentiles of the distribution of diversion ratios  $D_{jkt}$  for all product pairs  $(j, k)$  and markets  $t$ .

and provide a measure of the direction of substitution when the price of good  $j$  increases. Diversion ratios are also focal arguments of a multi-product firm’s first-order conditions, and as such, are of central importance to the analysis of horizontal mergers and unilateral price effects (Conlon and Mortimer, 2021).

Table 5 reports various percentiles of the distribution of product-market diversion ratio estimates. Not surprisingly, we find that a lack of economic structure in the estimated demand functions translates to unreasonable estimates of diversion ratios. For example, estimates of upward-sloping demand from GMM specifications lead to negative diversion ratios, implying that price increases can steal customers away from competitors. Diversion ratios should also be bound between zero and one, which is not guaranteed to hold in a nonparametric system without diagonal-dominance-like restrictions tying together the magnitude of the own-price effects with the sum of cross-price effects. The quasi-Bayes estimator produces improved diversion ratio estimates that fall in the unit interval. The only exception is the Jams, Jellies, & Peanut Butter category which contains negative estimates due to the presence of complements and a small share of estimates larger than one.

## 7 Extensions

Although our QBNPD estimation method delivers several improvements over existing NPD estimation approaches, it still has its own limitations. We see multiple directions for future research.

**Constraints.** In all of our empirical work, we enforce constraints on a pre-specified grid of “in-sample” market shares. There are two potential drawbacks to this approach. First, by requiring constraints to hold everywhere on this large grid, we are inducing significant bias. Although (quasi-)Bayesian models tend to favorably navigate the bias-variance tradeoff, the magnitude of bias may be undesirable if all that is required for the analysis is that constraints should hold at some small representative set of shares. In this case, researchers can easily change the specification of  $\mathcal{G}$  and continue to use our framework to ensure that constraints hold on this smaller grid. Second, because this grid is defined prior to estimation, it is possible that our approach here is both stronger than strictly necessary (because it enforces constraints on many irrelevant points) and may not enforce constraints on important parts of the domain. A fruitful extension of this work would be to define model- and data-informed constraint grids. For example, enforcing constraints only at or near counterfactuals of interest, or at chosen model equilibria, may provide sufficient regularization for some settings and would require significantly weaker assumptions than the constraints we currently impose.

**Asymptotics.** Our Monte Carlo simulation results raise several open questions about the theoretical advantages offered by our constrained QBNPD estimator and what can be said about its behavior in large samples. For example, the results of [Figure 2](#) suggest that more judicious constraints can deliver faster rates of convergence. Future work could consider characterizing such gains via non-asymptotic bounds on the frequentist risk ([Chetverikov and Wilhelm, 2017](#)) or a novel BvM ([Bochkina and Green, 2014](#); [Gallant et al., 2022](#)).

**Curse of dimensionality.** Our quasi-Bayes approach to regularizing demand can in principle accommodate any nonparametric approximation of (inverse) demand. In our analysis, we use linear sieves which are well-studied and have been shown to exhibit many attractive theoretical properties in NPIV settings. However, linear sieves are also subject to a curse of dimensionality that renders them impractical for a lot of applied work. We see two fruitful paths forward. The first is to impose additional restrictions on the type and interaction of basis functions used. For example, Bernstein sieves grow exponentially in  $J$ , but it is possible that much smaller sieves may well approximate simple inverse demand functions. One option is to then drop some interaction terms in a common sieve, as in [Christensen \(2017\)](#). Although this is beyond the scope of this paper, we show preliminary evidence in [Table D3](#) that our complements DGP can be approximated well by fewer parameters than the number induced by a tensor product of Bernstein polynomials. We have also implemented options in `NPDemand.jl` to allow for low-dimensional polynomials so that researchers can explore the practical performance of this approach. A second path forward is to move beyond the class of linear sieves and turn to a growing class of ML-based NPIV estimators. For example, [Chiang et al. \(2026\)](#) exploit the scalability of variational method of moments ([Bennett and Kallus, 2023](#)) to estimate nonparametric supply functions for up to  $J = 30$  goods. Future work could integrate such estimators into our quasi-Bayes regularization framework.

**Cross-market heterogeneity.** We also see value in exploring methods for incorporating additional market-level heterogeneity into the specification of demand. In our analysis, we assume that prices and all other observed market characteristics affect demand only through product indices. While we believe this is a reasonable assumption given the goals of the paper, it does restrict the degree of market heterogeneity captured by our empirical specifications. The “simplest” solution is to include prices and/or market-level data as extra covariates in the inverse demand functions—with the caveat that the dimension of the estimand will increase, raising concerns to the practicality of linear sieves as discussed above. In some cases, researchers may also have access to individual-level data, which could be incorporated similarly by including market-level moments of individual-level information in the inverse demand function. In the longer term, identifying routes that are closer to the “micro moments” used in mixed logit estimation (Conlon and Gortmaker, 2025) would likely provide significantly more identifying power.

## 8 Conclusion

In this paper, we propose a quasi-Bayesian approach to estimating multi-product NPD systems subject to economic constraints. Our approach transforms a classical nonparametric estimator of (inverse) demand functions into a quasi-likelihood and then uses priors to regularize towards theory. We employ a novel Sequential Monte Carlo sampling algorithm which targets a sequence of posteriors with softened constraints converging to the desired hard constraint. We use simulated data to show that our approach delivers superior finite sample performance relative to two benchmark GMM estimators. We also apply our methods to grocery retail data and estimate demand for several CPG categories which together span 17 products and more than 450,000 product-market elasticities. Our results show that our quasi-Bayes delivers useful economic regularization. The analysis throughout the paper showcases our accompanying Julia package `NPDemand.jl`, which allows researchers to estimate NPD using both GMM and quasi-Bayes estimators—all with minimal user overhead. We hope that our methods and tools, accompanied by evidence of their practical value, provide a useful step forward and encourage other researchers to test nonparametric approaches in their analysis of consumer demand.

## References

- Abril-Pla, O., Andreani, V., Carroll, C., Dong, L., Fonnesebeck, C. J., Kochurov, M., Kumar, R., Lao, J., Luhmann, C. C., Martin, O. A., et al. (2023). PyMC: A modern, and comprehensive probabilistic programming framework in python. *PeerJ Computer Science*, 9:e1516.
- Bennett, A. and Kallus, N. (2023). The variational method of moments. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(3):810–841.
- 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, Volume 4*, pages 1–62. Elseiver.
- Beskos, A., Jasra, A., Kantas, N., and Thiery, A. (2016). On the convergence of adaptive sequential Monte Carlo methods. *The Annals of Applied Probability*, 26(2):1111–1146.
- Betancourt, M. (2017). A conceptual introduction to Hamiltonian Monte Carlo.
- Birchall, C., Mohapatra, D., and Verboven, F. (2024). Estimating substitution patterns and demand curvature in discrete-choice models of product differentiation. *Review of Economics and Statistics*.
- Bissiri, P. G., Holmes, C. C., and Walker, S. G. (2016). A general framework for updating belief distributions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):1103–1130.
- Blundell, R., Chen, X., and Kristensen, D. (2007). Semi-nonparametric IV estimation of shape-invariant Engel curves. *Econometrica*, 75(6):1613–1669.
- 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.
- Blundell, R., Horowitz, J. L., and Parey, M. (2012). Measuring the price responsiveness of gasoline demand: Economic shape restrictions and nonparametric demand estimation. *Quantitative Economics*, 3(1):29–51.
- Bochkina, N. A. and Green, P. J. (2014). The Bernstein–von Mises theorem and nonregular models. *The Annals of Statistics*, 42(5):1850–1878.

- Brand, J. (2021). Differences in differentiation: Rising variety and markups in retail food stores.
- Carrasco, M., Florens, J.-P., and Renault, E. (2007). Linear inverse problems in structural econometrics estimation based on spectral decomposition and regularization. volume 6 of *Handbook of Econometrics*, pages 5633–5751. Elsevier.
- Chen, X., Christensen, T., and Kankanala, S. (2024). Adaptive estimation and uniform confidence bands for nonparametric structural functions and elasticities. *Review of Economic Studies*, 92(1):162–196.
- Chen, X. and Christensen, T. M. (2018). Optimal sup-norm rates and uniform inference on nonlinear functionals of nonparametric IV regression. *Quantitative Economics*, 9(1):39–84.
- Chen, X., Christensen, T. M., and Tamer, E. (2018). Monte Carlo confidence sets for identified sets. *Econometrica*, 86(6):1965–2018.
- Chen, X. and Pouzo, D. (2012). Estimation of nonparametric conditional moment models with possibly nonsmooth generalized residuals. *Econometrica*, 80(1):277–321.
- Chernozhukov, V. and Hong, H. (2003). An MCMC approach to classical estimation. *Journal of Econometrics*, 115(2):293–346.
- Chernozhukov, V., Newey, W. K., and Santos, A. (2023). Constrained conditional moment restriction models. *Econometrica*, 91(2):709–736.
- Chetverikov, D. and Wilhelm, D. (2017). Nonparametric instrumental variable estimation under monotonicity. *Econometrica*, 85(4):1303–1320.
- Chiang, H. D., Collison, J., Magnolfi, L., and Sullivan, C. (2026). Market counterfactuals with nonparametric supply: An ML/AI approach.
- Chopin, N. (2002). A sequential particle filter method for static models. *Biometrika*, 89(3):539–552.
- Chopin, N. (2004). Central limit theorem for sequential Monte Carlo methods and its application to Bayesian inference. *The Annals of Statistics*, 32(6):2385–2411.
- Christensen, T. M. (2017). Nonparametric stochastic discount factor decomposition. *Econometrica*, 85(5):1501–1536.
- Compiani, G. (2022). Market counterfactuals and the specification of multiproduct demand: A nonparametric approach. *Quantitative Economics*, 13(2):545–591.
- 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 Gortmaker, J. (2025). Incorporating micro data into differentiated products demand estimation with PyBLP. *Journal of Econometrics*.

- Conlon, C. and Mortimer, J. H. (2021). Empirical properties of diversion ratios. *The RAND Journal of Economics*, 52(4):693–726.
- Crisan, D. and Doucet, A. (2002). A survey of convergence results on particle filtering methods for practitioners. *IEEE Transactions on Signal Processing*, 50(3):736–746.
- Darolles, S., Fan, Y., Florens, J. P., and Renault, E. (2011). Nonparametric instrumental regression. *Econometrica*, 79(5):1541–1565.
- Deaton, A. and Muellbauer, J. (1980). An almost ideal demand system. *American Economic Review*, 70(3):312–326.
- Del Moral, P., Doucet, A., and Jasra, A. (2006). Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(3):411–436.
- Douc, R. and Moulines, E. (2008). Limit theorems for weighted samples with applications to sequential Monte Carlo methods. *The Annals of Statistics*, 36(5):2344–2376.
- Doucet, A., de Freitas, N., and Gordon, N. (2001). *Sequential Monte Carlo Methods in Practice*. Springer.
- Fosgerau, M., Monardo, J., and De Palma, A. (2024). The inverse product differentiation logit model. *American Economic Journal: Microeconomics*, 16(4):329–370.
- Freyberger, J. and Horowitz, J. L. (2015). Identification and shape restrictions in nonparametric instrumental variables estimation. *Journal of Econometrics*, 189(1):41–53.
- Gallant, A. R., Hong, H., Leung, M. P., and Li, J. (2022). Constrained estimation using penalization and MCMC. *Journal of Econometrics*, 228(1):85–106.
- Ge, H., Xu, K., and Ghahramani, Z. (2018). Turing: A language for flexible probabilistic inference. In *International Conference on Artificial Intelligence and Statistics, AISTATS 2018, 9-11 April 2018, Playa Blanca, Lanzarote, Canary Islands, Spain*, pages 1682–1690.
- Gelman, A. (2004). Parameterization and Bayesian modeling. *Journal of the American Statistical Association*, 99(466):537–545.
- Ghosal, S. and van der Vaart, A. (2017). *Fundamentals of Nonparametric Bayesian Inference*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.
- Golchi, S. and Campbell, D. A. (2016). Sequentially constrained Monte Carlo. *Computational Statistics & Data Analysis*, 97:98–113.
- Haag, B. R., Hoderlein, S., and Pendakur, K. (2009). Testing and imposing Slutsky symmetry in nonparametric demand systems. *Journal of Econometrics*, 153(1):33–50.

- Hausman, J. A. and Newey, W. K. (1995). Nonparametric estimation of exact consumers surplus and deadweight loss. *Econometrica*, 63(6):1445–1476.
- Herbst, E. and Schorfheide, F. (2014). Sequential Monte Carlo sampling for DSGE models. *Journal of Applied Econometrics*, 29(7):1073–1098.
- Herbst, E. and Schorfheide, F. (2019). Tempered particle filtering. *Journal of Econometrics*, 210(1):26–44.
- Hoderlein, S. and Lewbel, A. (2012). Regression dimension reduction with economic constraints: The example of demand systems with many goods. *Econometric Theory*, 28(5):1087–1120.
- Hoffman, M. D. and Gelman, A. (2014). The No-U-Turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15:1593–1623.
- Jasra, A., Stephens, D. A., Doucet, A., and Tsagaris, T. (2011). Inference for Lévy-driven stochastic volatility models via adaptive sequential Monte Carlo. *Scandinavian Journal of Statistics*, 38(1):1–22.
- Kankanala, S. (2025). Generalized Bayes in conditional moment restriction models.
- Kato, K. (2013). Quasi-Bayesian analysis of nonparametric instrumental variables models. *The Annals of Statistics*, 41(5):2359–2390.
- Liao, Y. and Jiang, W. (2011). Posterior consistency of nonparametric conditional moment restricted models. *The Annals of Statistics*, 39(6):3003–3031.
- Lubin, M., Dowson, O., Dias Garcia, J., Huchette, J., Legat, B., and Vielma, J. P. (2023). JuMP 1.0: Recent improvements to a modeling language for mathematical optimization. *Mathematical Programming Computation*, 15:581–589.
- Miravete, E. J., Seim, K., and Thurk, J. (2024). Elasticity and curvature of discrete choice demand models.
- Newey, W. K. and Powell, J. L. (2003). Instrumental variable estimation of nonparametric models. *Econometrica*, 71(5):1565–1578.
- Pachali, M. J., Kurz, P., and Otter, T. (2020). How to generalize from a hierarchical model? *Quantitative Marketing and Economics*, 18(4):343–380.
- Robert, C. P. and Casella, G. (2004). *Monte Carlo Statistical Methods*. Springer.
- Stan Development Team (2024). RStan: The R interface to Stan. R package version 2.32.6.

# APPENDIX

## A Proof of [Theorem 1](#)

In this appendix, we show that the baseline quasi-posterior  $\bar{\pi}(\theta|\mathcal{D}_T) \propto L_T(\theta)\bar{\pi}(\theta)$  fits within the general framework of [Kato \(2013\)](#) and thus possesses the nice asymptotic properties stated in [Theorem 1](#).

### A.1 Setup

For ease of exposition, we focus on the following model:

$$x_{jt} = h_{0j}(\mathbf{y}_t) + \xi_{jt}, \quad \mathbb{E}(\xi_{jt}|\mathbf{w}_t) = 0, \quad j = 1, \dots, J \quad (32)$$

where  $x_{jt} \in \mathbb{R}$  is the outcome variable,  $\mathbf{y}_t \in \mathcal{Y} \subseteq \mathbb{R}^D$  represents the endogenous variable(s),  $\mathbf{w}_t \in \mathcal{W} \subseteq \mathbb{R}^D$  represents the exogenous variable(s), and  $h_{0j} \in \mathcal{H}$  is one of  $j = 1, \dots, J$  true the structural functions of interest. Let  $\mathcal{D}_t = \{x_{jt}, \mathbf{y}_t, \mathbf{w}_t\}$  denote the set of observed variables which are IID over  $t = 1, \dots, T$ . Going forward, we drop the  $t$  subscript unless necessary.

To equate (32) to the conditional moment restriction model in (9), assume without loss of generality there is only one observed non-price characteristic  $x_{jt}$  with normalized  $\beta_j = 1$  and let  $\mathbf{y}_t \equiv \tilde{\mathbf{s}}_t$ ,  $\mathbf{w}_t \equiv (\mathbf{X}_t, \mathbf{z}_t)$ ,  $D = \tilde{J}$ , and  $h_j(\cdot) \equiv \sigma_j^{-1}(\cdot)$ . The structure of (32) also matches the NPD model for which [Compiani \(2022\)](#) derives asymptotic results.

### A.2 Discussion of Assumptions

In what follows let  $C$  denote some positive and sufficiently large constant and let  $c$  denote a positive and sufficiently small constant. We start with the following assumptions about the observed data and moments.

**Assumption A8** (Data).

- (i)  $(\mathbf{y}, \mathbf{w})$  has a joint density  $f(\mathbf{y}, \mathbf{w})$  on  $\mathcal{Y} \times \mathcal{W}$  which satisfies  $f_{Y,W}(\mathbf{y}, \mathbf{w}) \leq C$ .
- (ii)  $\mathcal{Y} = \mathcal{W} = [0, 1]^D$ .

**Assumption A9.** For all  $j = 1, \dots, J$ ,  $k = 1, \dots, J$ , and  $j \neq k$  we have

- (i)  $\sup_{\mathbf{w} \in \mathcal{W}} \mathbb{E}(\xi_j^2|\mathbf{w}) \leq C$ ;
- (ii)  $\sup_{\mathbf{w} \in \mathcal{W}} \mathbb{E}(\xi_j \xi_k|\mathbf{w}) \leq C$ ;
- (iii)  $\sup_{\mathbf{w} \in \mathcal{W}} \mathbb{E}(\xi_j^2 \mathbf{1}(|\xi_j^2| > \lambda)|\mathbf{w}) \rightarrow 0$  as  $\lambda \rightarrow \infty$ ;
- (iv)  $\mathbb{E}(|\xi_j|^{2+\delta}) \leq C$  for some  $\delta > 0$ ;
- (v)  $\mathbb{E}(|\xi_j \xi_k|^{1+\delta'}) \leq C$  for some  $\delta' > 0$ .

Assumptions A8 and A9 are standard in the NPIV literature. Conditions on the random variables and associated densities in A8 are basic ingredients that help establish regularities in the tools of functional analysis used throughout. Assumption A9(i) helps ensure finite population variances and A9(iii) is a uniform integrability condition which is assumed by Compiani (2022) and eventually required by Kato (2013) to establish sharper contraction rates. Assumptions 2(ii), (iv), and (iv) impose additional regularity on cross-equation covariances in our system of  $J$  structural functions.

We now turn to the identification of each structural function  $h_{0j}$ . Let  $\mathbb{T} : L_2([0, 1]^D) \rightarrow L_2([0, 1]^D)$  denote the linear operator defined by  $\mathbb{T}(h_j)(\mathbf{w}) := \mathbb{E}(h_j(\mathbf{y})|\mathbf{w})f_W(\mathbf{w})$  and  $r_j(\mathbf{w}) := \mathbb{E}(x_j|\mathbf{w})f_W(\mathbf{w})$ . We can then rewrite the NPIV model in (32) as the following operator equation.

$$\mathbb{T}h_{0j} = r_j \tag{33}$$

**Assumption A10** (Operator). *The operator  $\mathbb{T}$  is injective.*

While Assumption A10 ensures that  $h_{0j}$  is identified, recovering  $h_{0j}$  from the data requires an inversion  $h_{0j} = \mathbb{T}^{-1}r_j$ . Because  $\mathbb{T}$  is a compact operator, its singular values tend towards zero and small changes in  $r_j$  can imply arbitrarily large changes in  $h_{0j}$  making it hard to infer the true function from even large amounts of data. It is in this sense that the problem is ill-posed.

Our first layer of regularization comes via assumptions of smoothness and a subsequent series representation (Blundell et al., 2007; Chen and Pouzo, 2012; Chen and Christensen, 2018).

**Assumption A11** (Smoothness). *Each of the true structural functions  $h_{01}, \dots, h_{0J}$  belongs to a Hölder class  $\mathcal{H}^s$  with smoothness parameter  $s > D/2$ .*

Following Assumption A11, we can write

$$h_{0j}(\mathbf{y}) = \sum_{k_1=0}^{K_{0j}} \cdots \sum_{k_D=0}^{K_{0j}} \theta_{0jk_1 \cdots k_D} \prod_{\ell=1}^D \phi_{jk_\ell}(y_\ell) \tag{34}$$

for a set of basis functions  $\phi_j^{M_j}(\cdot) = (\phi_{j1}(\cdot), \dots, \phi_{jM_j}(\cdot))'$  and some sufficiently large values  $K_{01}, \dots, K_{0J}$ . We approximate each series representation above by a finite-dimensional sum:

$$h_j(\mathbf{y}) = \sum_{k_1=0}^{K_j} \cdots \sum_{k_D=0}^{K_j} \theta_{jk_1 \cdots k_D} \prod_{\ell=1}^D \phi_{jk_\ell}(y_\ell) \tag{35}$$

Hence, each structural function  $h_j$  is approximated by a tensor product of univariate basis functions of order  $K_j$ . Let  $M_j = (K_j+1)^D$  denote the corresponding sieve dimension for  $h_j$  and  $M = \sum_{j=1}^J M_j$  denote the sieve dimension for the entire system of equations  $h = (h_1, \dots, h_J)$ . While we do not include  $T$  subscripts on  $M$  or  $M_j$ , we allow both to grow with  $T$ .

Going forward, it will be useful to define

$$G_j = \mathbb{E}(\phi_j^{M_j}(\mathbf{w})\phi_j^{M_j}(\mathbf{w})') \quad (36)$$

$$\Psi_j = \mathbb{E}(\phi_j^{M_j}(\mathbf{w})\phi_j^{M_j}(\mathbf{y})') \quad (37)$$

as well as the basis sup-norm  $\zeta_j^{M_j} = \sup_{\mathbf{w} \in \mathcal{W}} \|G_j^{-\frac{1}{2}}\phi_j^{M_j}(\mathbf{w})\|$  with  $\zeta_M = \max(\zeta_1^{M_1}, \dots, \zeta_J^{M_J})$ .

We impose the following regularity conditions on the basis functions.

**Assumption A12** (Basis functions).

- (i)  $\phi_j^{M_j}(\cdot) = \phi_k^{M_k}(\cdot)$  for all  $j \neq k$ ,  $j = 1, \dots, J$ ,  $k = 1, \dots, J$
- (ii) We use the same basis  $\phi(\cdot)$  to approximate the structural functions  $h_j \in \mathcal{H}$  as we do to approximate the space of instruments.
- (iii)  $s_{\min}(G_j) \geq C^{-D}$  for all  $M_j \geq M_{0j}$  and for each  $j = 1, \dots, J$ ;
- (iv) There exists  $s_j > 0$  such that  $s_{\min}(\Psi_j) \geq C^{-1}M_j^{-s_j}$  for all  $M_j \geq M_{0j}$  and for each  $j = 1, \dots, J$ ;
- (v)  $\|\mathbb{E}(\phi_j^{M_j}(\mathbf{w})(h_{0j} - h_j)(\mathbf{y}))\| \leq C\tau_j^{M_j}\|h_{0j} - h_j\|$  for all  $M_j \geq M_{0j}$  and for each  $j = 1, \dots, J$ .

Assumptions A12(i) and (ii) are not necessary but are consistent with our operationalization and are thus imposed for simplicity. Assumptions A12(iii) and (iv) bound the smallest singular values of  $G_j$  and  $\Psi_j$  away from zero to ensure invertibility. Note that we can accordingly define a sieve measure of ill-posedness as  $\tau_j^{M_j} = s_{\min}(\Psi_j)$  which is the reciprocal of the measure of ill-posedness typically used in the literature (Blundell et al., 2007). To define a single measure across the system of functions, let  $\tau_M = \max(\tau_1^{M_1}, \dots, \tau_J^{M_J})$ . Lastly, Assumption A12(v) represents a stability condition on the bias  $h_{0j} - h_j$ .

Next, we define the quasi-likelihood. Following Section 3, we start with the conditional moment restriction model induced by (32):

$$m_j(h_{0j}) = \mathbb{E}(x_j - h_{0j}(\mathbf{y})|\mathbf{w}) = 0, \quad j = 1, \dots, J \quad (38)$$

which is approximated by  $\hat{m}_j(\theta_j^{M_j})$  where  $\theta_j^{M_j}$  is the vector of coefficients on the sieve approximation  $h_j(\mathbf{y}; \theta_j^{M_j})$  in (35). Let  $\theta^M = ((\theta_1^{M_1})', \dots, (\theta_J^{M_J})')'$  denote the  $M$ -dimensional vector of parameters across all  $J$  demand equations with corresponding ‘‘true’’ values  $\theta_0^{M_0} = ((\theta_{01}^{M_{01}})', \dots, (\theta_{0J}^{M_{0J}})')$ .

Define the quasi-likelihood as  $L_T(\theta^M) = e^{-\frac{T}{2}Q_T(\theta^M)}$  where:

$$Q_T(\theta^M) = \sum_{j=1}^J \hat{m}_j(\theta_j^{M_j})' \hat{\Omega}_j \hat{m}_j(\theta_j^{M_j}) \quad (39)$$

and  $\hat{m}_j(\theta_j^{M_j}) = \frac{1}{T} \sum_{t=1}^T (x_{jt} - h(\mathbf{y}_t; \theta_j^{M_j})) \cdot \phi_j^{M_j}(\mathbf{w}_t)$  and  $\hat{\Omega}_j$  is a  $M_j \times M_j$  weighting matrix satisfying the following properties.

**Assumption A13** (Weighting matrix). *There exists a set of symmetric weighting matrices  $\Omega_1, \dots, \Omega_J$  such that:*

(i)  $\|\hat{\Omega}_j - \Omega_j\| = o(\gamma_T)$  with  $\gamma_T M_j \rightarrow 0$  for each  $j = 1, \dots, J$ ;

(ii)  $\mathbb{P}(c \leq \lambda_{\min}(\Omega_j) \leq \lambda_{\max}(\Omega_j) \leq C) \rightarrow 1$  for each  $j = 1, \dots, J$ .

Each target function is estimated via  $\hat{h}_j(\mathbf{y}; \hat{\theta}_j^{M_j}) = \phi_j^{M_j}(\mathbf{y})' \hat{\theta}_j^{M_j}$  where

$$\hat{\theta}_j^{M_j} = \frac{1}{T} (\hat{\Phi}_j' \hat{\Omega}_j \hat{\Phi}_j)^{-1} \hat{\Phi}_j' \hat{\Omega}_j \phi_j^{M_j}(\mathbf{w})' \mathbf{x}_j \quad (40)$$

and  $\hat{\Phi}_j = \frac{1}{T} \sum_{t=1}^T \phi_j^{M_j}(\mathbf{w}_t) \phi_j^{M_j}(\mathbf{y}_t)'$  is the finite-sample analog of  $\Phi_j$ . Stacking the  $J$  estimators together yields

$$\hat{\theta}^M = \frac{1}{T} (\hat{\Phi}' \hat{\Omega} \hat{\Phi})^{-1} \hat{\Phi}' \hat{\Omega} \phi^M(\mathbf{w})' \mathbf{x} \quad (41)$$

where  $\hat{\Phi} = \text{diag}(\hat{\Phi}_1, \dots, \hat{\Phi}_J)$ ,  $\hat{\Omega} = \text{diag}(\hat{\Omega}_1, \dots, \hat{\Omega}_J)$ , and  $\phi^M(\cdot) = (\phi_1^{M_1}(\cdot), \dots, \phi_J^{M_J}(\cdot))'$ . The asymptotic variance takes the form:

$$\mathbb{V}(\hat{\theta}^M) = \frac{1}{T} (\Phi' \Omega \Phi)^{-1} \Phi' \Omega \Sigma \Omega \Phi (\Phi' \Omega \Phi)^{-1} \quad (42)$$

where  $\Sigma$  is the  $M \times M$  covariance matrix of the stacked sample moments and is comprised of the  $M_j \times M_k$  submatrices

$$\Sigma_{jk} = \mathbb{E}(\xi_j \xi_k \phi_j^{M_j}(\mathbf{w}) \phi_k^{M_k}(\mathbf{w})'). \quad (43)$$

Finally, we turn to the specification of suitable priors on parameters  $\theta^M$  which add another layer of regularization to the model. Let  $\bar{\Pi}(\cdot)$  denote the prior distribution which admits density  $\bar{\pi}(\cdot)$ . We require the following two prior regularity conditions to hold.

**Assumption A14** (Prior small ball). *There exists a constant  $C_\theta > 0$  such that for all  $T$  sufficiently large,  $\bar{\Pi}(\|\theta^M - \theta_0^{M_0}\|_{\ell^2} \leq \epsilon_T) \geq e^{-C_\theta T \epsilon_T^2}$  as  $\epsilon_T \rightarrow 0$ .*

**Assumption A15** (Prior flatness). *Let  $\gamma(\epsilon_T) = M^{-s/D} + \tau_M^{-1} \epsilon_T$ . There exists a sequence  $B_T \rightarrow \infty$  sufficiently slowly such that for all  $T$  sufficiently large,  $\bar{\pi}(\theta^M)$  is positive for all  $\|\theta^M - \theta_0^{M_0}\| \leq B_T \gamma(\epsilon_T)$  and*

$$\sup_{\|\theta^M\| \leq B_T \gamma(\epsilon_T), \|\tilde{\theta}^M\| \leq B_T \gamma(\epsilon_T)} \left| \frac{\bar{\pi}(\theta_0^{M_0} + \theta^M)}{\bar{\pi}(\theta_0^{M_0} + \tilde{\theta}^M)} - 1 \right| \rightarrow 0 \quad \text{as } \epsilon_T \rightarrow 0.$$

Assumption A14 ensures that the prior assigns sufficient mass to small neighborhoods around the true parameter  $\theta_0$ . Assumption A15 requires the prior to become sufficiently flat in a ball centered around  $\theta_0$ . Both conditions, in some form, are required to establish nonparametric BvM results (Ghosal and van der Vaart, 2017). In our operationalization, for example, we assume the baseline prior is Gaussian:  $\theta^M \sim N(0, a^{-2} I_M)$ . Hence, we can write  $\bar{\pi}(\theta^M) = \prod_{j=1}^J \prod_{k=1}^{M_j} \varphi(\theta_{jk})/a$ , where  $\varphi(\cdot)$  is the standard normal density function. Kato (2013) shows that this type of Gaussian “product prior” satisfies Assumptions A14 and A15.

### A.3 Results

**Theorem 1.** Let  $\bar{\Pi}(\cdot|\mathcal{D}_T)$  denote the baseline posterior distribution admitting the posterior density  $\bar{\pi}(\theta^M|\mathcal{D}_T) \propto L_T(\theta^M)\bar{\pi}(\theta^M)$ , and suppose that all regularity conditions on the target structural functions (1, 2), data (3, A8), moments (A9), conditional expectation operator (A10), sieve approximation (A11, A12), weighting matrix (A13), and prior (A14, A15) hold. Then

- (i)  $\bar{\Pi}(\cdot|\mathcal{D}_T)$  is consistent and approximately Gaussian in large samples;
- (ii) The credible intervals derived from  $\bar{\Pi}(\cdot|\mathcal{D}_T)$  have asymptotically exact frequentist coverage under optimal weighting of the moments in  $L_T(\theta^M)$ .

*Proof.* Of the assumptions stated above, we introduce A9(ii), A9(iv), A9(v) and A13 to extend the results of Kato (2013) to a setting that includes a non-diagonal weighting matrix and a system of  $J$  multivariate structural functions. We also modify some of the basis function assumptions in A12 to be consistent with this setup (e.g., see also Chen and Christensen, 2018; Compiani, 2022).

- (i) Consistency and asymptotic normality follow from Theorems 1-2 in Kato (2013). Specifically, Theorem 1 establishes both consistency and asymptotic normality, and Theorem 2 establishes sharper contraction rates under an additional uniform integrability condition stated in Assumption A9(iii). We state the main results here for completeness.

**Consistency.** If we let  $M \rightarrow \infty$  in such a way that  $\zeta_M^2 \log M/T = o(\tau_M^2)$ , then there exists a constant  $C > 0$  such that

$$\bar{\Pi} \left( \|\theta^M - \theta_0^{M_0}\| > C(M^{-s/D} + \tau_M^{-1}\sqrt{M/T}) \mid \mathcal{D}_T \right) \rightarrow 0. \quad (44)$$

**BvM.** If we let  $M \rightarrow \infty$  in such a way that  $\zeta_M^6 \log M/T = o(\tau_M^2)$ , then

$$\left\| \bar{\Pi}(\cdot|\mathcal{D}_T) - N(\hat{\theta}^M, T^{-1}(\Phi'\Omega\Phi)^{-1})(\cdot) \right\|_{\text{TV}} \rightarrow 0. \quad (45)$$

- (ii) Establishing an asymptotic equivalence between quasi-Bayes credible intervals and frequentist confidence intervals amounts to showing that the limiting variance of the quasi-posterior matches the asymptotic variance of an efficient frequentist estimator  $\hat{\theta}^M$ . As can be seen in (45), the limiting variance of the quasi-posterior does not match the asymptotic variance in (42). However, if we take  $\Omega = \Sigma^{-1}$  where the elements  $\Sigma_{jk}$  are defined in (43), then

$$\begin{aligned} \mathbb{V}(\hat{\theta}^M) &= T^{-1}(\Phi'\Omega\Phi)^{-1}\Phi'\Omega\Sigma\Omega\Phi(\Phi'\Omega\Phi)^{-1} \\ &= T^{-1}(\Phi'\Sigma^{-1}\Phi)^{-1}\Phi'\Sigma^{-1}\Sigma\Sigma^{-1}\Phi(\Phi'\Omega\Phi)^{-1} \\ &= T^{-1}(\Phi'\Sigma^{-1}\Phi)^{-1} \end{aligned} \quad (46)$$

which does match the limiting variance of the quasi-posterior in (45). □

## B NPDemand.jl Package Usage and Examples

In order to begin using `NPDemand.jl`, one has to structure data as in Table B1. Here, we show an example data frame for a problem in which the index contains prices and a second characteristic  $x$ . For each product (indexed starting at 0), the data frame should have a numbered column for `shares`, `prices`, `x`, `share_iv`, and `price_iv`. The first three are straightforward, and the last two denote the instruments for market shares (in  $\sigma^{-1}$ ) and prices (in the index). The instruments for market shares must be full rank, because we require  $J$  dimensions of variation to identify  $\sigma^{-1}$ . In contrast, the instruments for price require much less variation as they jointly identify only a single coefficient (the implicit coefficient on price in the index relative to other covariates).

Table B1: Data Formatting

shares0	prices0	x0	price_iv0	shares1	prices1	x1	...
0.151	1.584	0.070	0.615	0.259	1.737	0.490	...
0.174	0.000	0.528	0.176	0.232	2.006	0.894	...
0.183	0.707	0.441	0.284	0.192	0.923	0.783	...
0.291	1.068	0.236	0.201	0.190	1.211	0.814	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

The next step is to define an NPD problem, which is analogous to a BLP problem in PyBLP (Conlon and Gortmaker, 2020). Defining the problem amounts to inputting the data frame as well as specifying (i) the set of variables entering demand through the index, (ii) sieve approximation details, including the degree of the polynomial approximation, and (iii) constraints, if any. The `define_problem()` command constructs all design and constraint matrices required to evaluate the candidate GMM objective function (or quasi-likelihood function) at a candidate parameter value. Estimation directly follows via `estimate!()`, which will store all estimation output inside of the NPD problem. By default, `estimate!()` estimates the NPD problem via GMM and uses `JuMP.jl` (Lubin et al., 2023) as the back-end solver. When `quasi_bayes=true` as an option inside `estimate!()`, then the NPD problem is estimated using our proposed quasi-Bayes approach. This step specifically uses MCMC to sample from a quasi-posterior defined by the quasi-likelihood (which is just the scaled GMM objective function) and default Gaussian priors over sieve coefficients. If constraints are specified in the NPD problem, then any corresponding linear restrictions will be imposed using a reparameterization. Note, however, that this step alone does not sample from the quasi-posterior in (15) because it does not include the hard constraints  $\mathbf{1}(\theta \in \mathcal{C}_G(\Theta))$ . To sample from the fully constrained quasi-posterior in (15), `estimate!()` should be followed by `smc!()`. The second step will use the SMC algorithm described in Algorithm 1 to transform draws from the “unconstrained” (or linearly-constrained) quasi-posterior targeted in the first step into draws from the fully constrained target posterior. An example workflow is shown in Figure B1.

Figure B1: Example Code to Define and Estimate NPD Problems

---

```
using NPDemand, Turing

# Define an NPD problem
problem = define_problem(df;
    exchange = [[1 2], [3 4]],
    index_vars = ["prices", "x"],
    approximation_details = Dict(
        :sieve_type => "bernstein",
        :order => 2,
        :tensor => true
    ),
    constraints = [
        :exchangeability, # products are exchangeable wrt the exchange vector
        :monotone, # demand function is monotonic in own-price
        :diagonal_dominance_all # diagonal dominance of Jacobian of demand
    ]
)

# Estimate via GMM
estimate!(problem)

# Estimate via Quasi-Bayes
# (a) Run HMC
estimate!(problem,
    quasi_bayes = true,
    n_samples = 2_000,
    burn_in = 0.25,
    sampler = NUTS()
)

# (b) Run SMC
smc!(problem,
    burn_in = 0.25,
    mh_steps = 20,
    ess_threshold = 200,
    smc_method = :adaptive,
    max_violations = 0.01
)

# Calculate price elasticities in every market (w/ 95% confidence intervals)
price_elasticities!(problem, CI = 0.95)

# Calculate the median own-price elasticity across all product-markets
summarize_elasticities(problem, "own", "median")

# Calculate the median price elasticity matrix across all markets
summarize_elasticities(problem, "matrix", "median")
```

---

## C Additional Data Details

### Aggregation

Our raw data consists of every unique transaction made in each store, and so for each  $\text{UPC} \times \text{date}$  we have repeated observations of transacted prices. The following steps outline our approach to product-market aggregation when constructing the dataset for a particular product category.

1.  $\text{UPC} \times \text{date} \times \text{household} \rightarrow \text{UPC} \times \text{date} \times \text{store}$

We compute store-level prices for each  $\text{UPC} \times \text{date}$  by taking the median price across transactions (there is rarely any within-date, cross-household variation in prices).

2.  $\text{UPC} \times \text{date} \times \text{store} \rightarrow \text{UPC} \times \text{week} \times \text{ZIP3}$

Next, we aggregate to the  $\text{week} \times \text{ZIP3}$  level by taking total quantity sold and the median UPC prices across all  $\text{date} \times \text{stores}$  within a  $\text{week} \times \text{ZIP3}$ .

3.  $\text{UPC} \times \text{week} \times \text{ZIP3} \rightarrow \text{brand} \times \text{subcategory} \times \text{week} \times \text{ZIP3}$

Finally, we aggregate from UPCs to brands by summing the total equivalent unit sales for all UPCs within a  $\text{brand} \times \text{subcategory}$ . We measure prices as a quantity-weighted averages of UPC-level prices (per equivalent unit), where the weights are computed separately for each year.

### Market sizes

To compute market sizes for each category, we leverage the fact that our raw data is at the household-trip level. Following [Brand \(2021\)](#), we start by calculating a ratio of the number households who ever purchased from the category in a given year to the number of households who purchased from the category in a week. This ratio is informative about the relative size of outside good demand—i.e., for every one household who we observed make a purchase in a given week, there were  $x$  additional households who were in the market but did not make a purchase. We compute this ratio at the weekly level for each category-year, and then take an average for the year. We then use this average ratio as a scaling factor for the maximum weekly unit sales for the category, where the maximum is taken across all cross-sectional units (e.g., ZIP3s).

### Product selection

The last step is to select a focal set of products within each category to use in our analysis. Because of the high-dimensional nature of our estimator, we only consider products from up to two subcategories. For each category, we first compute subcategory-level market shares. If the top subcategory has greater than 65% share, then we only consider products from that single subcategory. Otherwise, we consider products from the top two subcategories. Then, among the chosen 1-2 subcategories, we compute product-level market shares and select all products with greater than 5% share.

## D Additional Empirical Results

Table D1: First Stage Price Regressions

Category	Estimated Coefficient		$R^2$	$F$ -stat
	Feature Promo	Wholesale Price		
Baking Goods	-0.02		0.66	9045.88
	-0.02	0.72	0.68	6585.17
Beer	-0.01		0.86	7213.35
	-0.00	0.49	0.91	14364.86
Butter, Margarine, Spreads	-0.07		0.95	23919.90
	-0.07	0.24	0.96	13428.04
Cookies	-0.06		0.68	9876.66
	-0.06	0.01	0.68	4939.68
Fish Canned	-0.05		0.62	1339.88
	-0.05	-0.02	0.62	704.64
Franks	-0.06		0.76	5848.42
	-0.06	-0.19	0.76	2974.93
Frozen Ice Cream	-0.02		0.79	14130.04
	-0.02	0.23	0.79	7395.83
Frozen Pizza	-0.04		0.73	7727.72
	-0.04	0.70	0.78	7813.22
Jams, Jellies, Peanut Butter	-0.03		0.84	11434.28
	-0.03	0.06	0.84	5749.33
Ketchup	-0.01		0.68	1632.27
	-0.01	0.11	0.69	832.90
Mayonnaise	-0.03		0.52	10003.26
	-0.03	0.13	0.52	5143.90
Snacks	-0.04		0.84	8134.78
	-0.04	0.32	0.84	4769.14

Notes: All specifications include the same fixed effects as those included in our NPD specifications (product, year, quarter, state, and holiday).

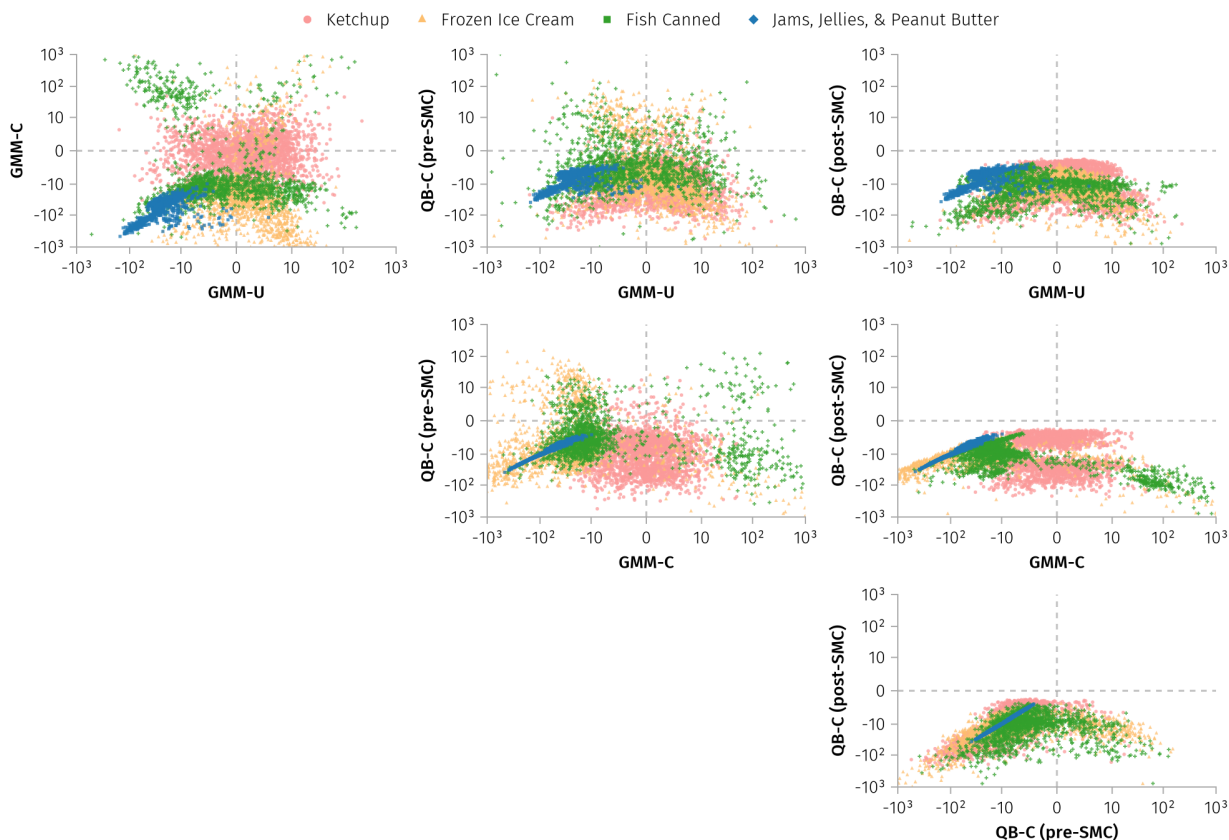


Figure D1: Joint distributions of own-price elasticities plotted across pairs of estimators. We average elasticity estimates up to the product-ZIP3-year-quarter level to aid in visualization. We split the constrained quasi-Bayes elasticities into values before SMC (where only linear restrictions are imposed) and after SMC to demonstrate the incremental value of economic regularization induced by the  $\mathbf{1}(\theta \in \mathcal{C}_{\mathcal{G}}(\Theta))$  component of our priors.

Table D2: Product-Level Median Own-Price Elasticity Estimates

	GMM-U		GMM-C		QB-C	
	Est	95% Interval	Est	95% Interval	Est	95% Interval
<i>Ketchup</i>						
1. Heinz	-20.73	(-26.81, 10.41)	-37.41	(-50.17, -30.31)	-3.37	(-10.02, -1.37)
2. Hunts	-35.90	(-40.80, 39.13)	-130.40	(-176.28, -103.55)	-10.52	(-32.03, -4.23)
<i>Frozen Ice Cream</i>						
1. Private Label	-1.96	(-4.00, -0.57)	-12.58	(-14.93, -11.11)	-5.42	(-15.95, -2.46)
2. Blue Bell	-7.14	(-9.33, -1.89)	-6.51	(-8.38, -5.55)	-8.67	(-27.93, -3.78)
3. Breyers	-1.07	(-4.72, 0.08)	-12.52	(-15.23, -9.95)	-7.54	(-24.15, -3.42)
4. Mayfield	-4.30	(-7.84, -0.24)	-0.41	(-2.91, 1.30)	-31.95	(-100.54, -14.01)
<i>Fish Canned</i>						
1. Starkist (CL)	0.25	(-1.09, 1.37)	-34.98	(-80.15, -24.49)	-4.65	(-9.71, -2.72)
2. Chicken of the Sea (CL)	1.29	(-0.26, 3.14)	-59.23	(-123.77, -3.03)	-11.80	(-25.69, -6.70)
3. Bumble Bee (CL)	4.99	(-0.24, 8.53)	-248.80	(-448.50, 116.29)	-35.95	(-78.91, -20.01)
4. Starkist (SW)	-0.95	(-2.35, 0.26)	-16.10	(-18.21, -14.34)	-11.68	(-26.92, -6.70)
5. Chicken of the Sea (SW)	-0.81	(-1.89, 0.15)	-14.90	(-16.34, -13.51)	-9.62	(-22.04, -5.51)
<i>Jams, Jellies, &amp; Peanut Butter</i>						
1. Jif (PB)	0.16	(-1.67, 0.76)	-0.38	(-2.13, 0.14)	-3.11	(-5.21, -1.78)
2. Peter Pan (PB)	1.56	(-0.73, 6.87)	-0.92	(-2.90, 1.17)	-19.48	(-33.34, -11.53)
3. Skippy (PB)	0.11	(-4.42, 9.82)	-0.15	(-1.94, 0.77)	-54.04	(-96.30, -31.02)
4. Private Label (PB)	0.40	(-1.91, 5.14)	0.01	(-0.46, 0.67)	-25.66	(-45.64, -15.00)
5. Smuckers (JJP)	0.90	(-0.07, 2.33)	0.22	(-1.01, 1.33)	-1.37	(-3.16, -0.55)
6. Welch's (JJP)	1.99	(-0.39, 4.40)	-0.20	(-0.93, 0.67)	-2.95	(-6.08, -1.08)

Notes: (1) GMM confidence intervals are calculated using quantiles of a bootstrapped sampling distribution with 100 bootstrap samples. (2) Quasi-Bayes credible intervals are calculated using quantiles of the empirical distribution of quasi-posterior draws. (3) Product subcategory abbreviations are listed in parentheses for categories with two subcategories (CL = “Chuck Light,” SW = “Solid White,” PB = “Peanut Butter,” JJP = “Jams, Jellies, Preserves”).

Table D3: Performance of Restricted Bernstein Polynomials

	Tensor		Non-Tensor		
	$K = 2$	$K = 3$	$K = 2$	$K = 3$	$K = 4$
Median Absolute Deviation	0.29	0.39	0.37	0.34	0.27
Violations (%)					
- own-good monotonicity	49	73	85	93	95
- any	99	100	100	100	100
Total Sieve Dimension	8	12	2	3	4
Number of Parameters	108	320	22	48	92

Notes: (1) We evaluate the performance of the unconstrained GMM estimator. (2) Performance metrics are averaged across 25 data sets generated from the complements DGP with  $J = 4$  goods and  $T = 500$  markets. (3) The tensor specifications match the implementation of the sieve estimator used throughout the paper. In this specification, each  $\sigma_j^{-1}$  is approximated by a tensor product of  $K$ -order Bernstein polynomials. (4) Non-tensor specifications naturally restrict interactions between univariate polynomials. (5) The violations represent shares of markets violating own-good monotonicity or any of the following three constraints: own-good monotonicity, within-group cross-good substitution, and across-group cross-good complements.