Choice Models and Permutation Invariance: Demand Estimation in Differentiated Products Markets

Amandeep Singh
University of Washington
Seattle, WA
amdeep@uw.edu

Ye Liu
University of Washington
Seattle, WA
yeliu@uw.edu

Hema Yoganarasimhan
University of Washington
Seattle, WA
hemay@uw.edu

Abstract

Choice Modeling is at the core of many economics, operations, and marketing problems. In this paper, we propose a fundamental characterization of choice functions that encompasses a wide variety of extant choice models. We demonstrate how non-parametric estimators like neural nets can easily approximate such functionals and overcome the curse of dimensionality that is inherent in the non-parametric estimation of choice functions. We demonstrate through extensive simulations that our proposed functionals can flexibly capture underlying consumer behavior in a completely data-driven fashion and outperform traditional parametric models. As demand settings often exhibit endogenous features, we extend our framework to incorporate estimation under endogenous features. Further, we also describe a formal inference procedure to construct valid confidence intervals on objects of interest like price elasticity. Finally, to assess the practical applicability of our estimator, we utilize a real-world dataset from Berry et al. (1995). Our empirical analysis confirms that the estimator generates realistic and comparable own- and cross-price elasticities that are consistent with the observations reported in the existing literature.

1 Introduction

Demand estimation is a critical component in the field of economics, operations, and marketing, enabling practitioners to model consumer choice behavior and understand how consumers react to changes in a market. This understanding helps policymakers and businesses alike make informed decisions, whether it be about launching new products, adjusting pricing strategies, or analyzing the consequences of mergers. Over the years, various approaches both parametric and non-parametric have been developed to address the complexities inherent in demand estimation. While parametric methods, based on logit or probit assumptions, have remained popular due to their simplicity and interpretability, they often require strong assumptions about the underlying choice process, limiting their ability to capture the true complexity of consumer preferences.

Nonparametric methods, on the other hand, offer a more flexible approach to demand estimation, allowing for more nuanced representations of consumer preferences without restrictive assumptions about the underlying distributions. Despite their potential advantages, non-parametric approaches often suffer from the “curse of dimensionality", as the complexity to estimate choice functions grows exponentially with the number of products. This challenge has limited the widespread adoption of nonparametric methods in practice.

In this paper, we make significant strides in bridging the gap between the flexibility of non-parametric methods and the tractability of parametric models by introducing a fundamental characterization of choice models. This characterization specifically addresses the curse of dimensionality in choice systems and enables the flexible estimation of choice functions via non-parametric estimators.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

A key advantage of contemporary parametric demand estimation approaches is their ability to model counterfactual demand in situations involving new product introductions or mergers. This strength often serves to highlight the limitations of existing non-parametric methods, which struggle with counterfactual estimation. Our proposed characterization, however, successfully estimates counterfactual demand in such scenarios, thereby offering a more robust approach to modeling consumer choice behavior.

Moreover, we recognize that real-world demand systems often contain unobserved demand shocks correlated with observable product features, such as prices. These shocks can lead to endogeneity issues, which can bias the estimated choice functions if not properly addressed. To tackle this challenge, we extend our framework to accommodate endogeneity.

Additionally, we build upon recent advances in automatic debiased machine learning and provide an inference procedure for constructing valid confidence intervals on objects of interest, such as price elasticities.

Finally, to showcase the effectiveness and applicability of our proposed framework, we use the Berry et al. (1995) Automobile dataset to estimate the price elasticities using our non-parametric estimator. The results of our analysis align with existing literature and demonstrate the practical utility of our approach, underscoring its potential for adoption in real-world demand estimation settings.

2 Theory

2.1 Choice Models

In this section, we provide a general characterization of consumer choice functions. In particular, we focus on a scenario where researchers have access only to aggregate market-level demand data, while individual-level choices and characteristics remain unobserved. Suppose consumers in a market t face an offer set $\mathcal{S}_t$ that can comprise any subset of J_t distinct products ($\{1, 2, \dots, J_t\}$) [1]. We use u_{ijt} to represent the index tuple $\{X_{jt}, I_{it}, \varepsilon_{ijt}\}$, where $X_{jt} \in \mathbb{C}^k$ denotes k product features belonging to some countable universe $\mathbb{C}^k$; $I_{it} \in \mathbb{C}^l$ denotes demographics of consumer i in market t, we assume there are l features and belong to some countable universe $\mathbb{C}^l$, and ε_{ijt} denotes random idiosyncratic components pertinent to consumer i for product j in market t that are not unobservable to the researcher but observable to consumers.

Definition 1 (Choice Function)**.** *Given the offer set $\mathcal{S}_t \subset \{1, 2, 3, \dots, J_t\}$, we define a function $\pi : \{u_{ijt} : j \in \mathcal{S}_t\} \to \mathbb{R}^{|\mathcal{S}_t|}$ that maps a set of index tuples $\{u_{ijt}\}_{j \in \mathcal{S}_t}$ to a $|\mathcal{S}_t|$-dimensional probability vector. Each element in the $\pi(\cdot)$ vector represents the probability of consumer i choosing product j in market t.*

Here we present a very general characterization of choice functions that maps the observable and unobservable components of product and individual characteristics to observed choices through some choice function π. Note that, traditionally, u_{ijt} is a scalar that represents utility in choice models. However, in our framework, u_{ijt} doesn't necessarily represent utility. Further, we have not yet imposed any assumption on π, i.e., how consumers make choices.

We now specify a set of assumptions on the model and data-generating process below.

Assumption 1 (Exogeneity). The unobserved error term ε_{ijt} is independent and identically distributed (i.i.d.) across all products. This can be expressed as follows:

$$\mathbb{P}(\varepsilon_{ijt} \mid X_{\cdot t}) = \mathbb{P}(\varepsilon_{ijt})$$

This assumption implies that the error term ε_{ijt} is not correlated with any of the observed variables $X_{\cdot t}$. As such, it precludes the possibility of endogenous prices and/or marketing-mix variables, as is common in observational data (). We start with the basic case with exogenous covariates in this section and later in Section 2.2, we relax this assumption and allow for endogenous covariates.

Assumption 2 (Identity Independence). For any product $j \in \mathcal{S}_t$ and any market t, we assume the choice function π does not depend on the identity of the product (jt). That is:

[1] The offer set S_t also includes an 'outside option,' a hypothetical choice where all product features are zero.

$$\pi_{ijt}(\{u_{ikt}\}_{k\in\mathcal{S}_t}) = \pi_{ijt}(u_{ijt}, \{u_{ikt}\}_{k\in\mathcal{S}_t,k\neq j}) = \pi(u_{ijt}, \{u_{ikt}\}_{k\in\mathcal{S}_t,k\neq j})$$

This assumption implies two things: first, the functional form of the choice probability for different products and markets is the same; second, for any market-level heterogeneity (e.g., in the distribution $F_t(I_{it}, \varepsilon_{ijt})$), we can include them in X_{jt} as features. Intuitively, this assumption suggests that conditional on product and consumer features and the unobserved error term, the choice probabilities are not functions of the identities of the products themselves.

Assumption 3 (Permutation Invariance). The choice function π is invariant under any permutation σ_j applied to the competitors of product j, such that:

$$\pi_{ijt} = \pi(u_{ijt}, \{u_{i\sigma_j(k)t}\}_{k\in\mathcal{S}_t,k\neq j})$$

In this assumption, we state that the choice function for product j is invariant to all permutations of its competitors. This implies that the individual's choice for product j is not affected by the order or identity of the other products in the market, and it only depends on the set of competitors' characteristics.

Since researchers only observe aggregate data, we next define the aggregate demand function. In aggregate demand settings, individual-level choices are not observable and only aggregate demand is observable. It is often the case that the market-specific individual features are not observable and are assumed exogenously drawn from some distribution $\mathcal{F}(m_t)$, where m_t represents the market-level characteristics. For the sake of notional simplicity, we let m_t to be the same across all markets. One can easily incorporate market-specific user demographics in the choice function. Thus the demand of product j in market t denoted by π_{jt} can be expressed as follows –

$$\pi_{jt} = \int\int \pi_{ijt}(\{u_{ikt}\}_{k\in\mathcal{S}_t}) d\mathcal{F}(m_t) d\mathcal{G}(\varepsilon_{ijt}), \tag{1}$$

where $\mathcal{G}(\varepsilon_{ijt})$ denotes the CDF of unobserved errors ε_{ijt}. Since u_{ijt} is determined by $\{X_{jt}, I_{it}, \varepsilon_{ijt}\}$ and $I_{it}, \varepsilon_{ijt}$ are integrated out in a market. Hence, we can express π_{jt} as a function of only the observable product characteristics –

$$\pi_{jt} = g(X_{jt}, \{X_{kt}\}_{k\in S,k\neq j}). \tag{2}$$

Lemma 1. *For any choice function that satisfies Assumption 1 and 3, the aggregate demand function is also permutation invariant.*

This permutation invariance of aggregate demand function exists because, under the exogeneity assumption, the aggregate demand function is simply the sum (or integral) of individual choice functions that are themselves invariant to permutation. Hence, changes to the order of competitors have no impact on the aggregated result. When the assumption of exogeneity is not satisfied, the aggregate demand function does not retain the permutation invariance, notwithstanding the fact that the individual-level choice function exhibit permutation invariance.

Our assumptions 2 (identity independence) and 3 (permutation invariance) are fairly standard in the choice modeling literature, although they might not always be explicitly named. Table A16 summarizes models that satisfy these assumptions.

Theorem 1. For any offer set $\mathcal{S}_t \subset \{1, 2, 3, \ldots, J_t\}$, if a choice function $\pi : \{u_{ijt} : j \in \mathcal{S}_t\} \to \mathbb{R}^{|\mathcal{S}_t|}$ where u_{ijt} represents the index tuple $\{X_{jt}, I_{it}, \varepsilon_{ijt}\}$ satisfies Assumption 1, 2 and 3, then there exists suitable ρ, ϕ_1 and ϕ_2 such that

$$\pi_{jt} = \rho(\phi_1(X_{jt}) + \sum_{k\neq j,k\in\mathcal{S}_t} \phi_2(X_{kt})),$$

Proof: See Appendix B.

This result is the generalization of the results shown in Zaheer et al. (2017) and can be shown following similar arguments. The above result is very powerful and has two important takeaways: (i) input space of the choice function does not grow with the number of products in the assortment. The input space of the choice function (i.e., ϕ_1 and ϕ_2) grows only as a function of the number of features of the products in consideration, and (ii) the result remains valid for all offer sets, denoted by $\mathcal{S}_t$, irrespective of their size. This allows one to easily simulate the demand and entry of new products or changes in market structure, as one does with traditional parametric models. As an example, for the multinomial logit model one possible set of transformations could be $\phi_1(x) = \begin{bmatrix} \exp(x) \\ 0 \end{bmatrix}$ and $\phi_2(x) = \begin{bmatrix} 0 \\ \exp(x) \end{bmatrix}$ that generate two-dimensional vectors, and the function $\rho\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right) = \frac{x_1}{x_1+x_2}$ operates on these vectors. [2]

2.2 Endogenous Covariates

In this section, we relax the exogeneity assumption and handle the potential endogeneity issue that is commonplace in demand settings. Note that, when the price (or other product characteristics or mixed variables, such as promotions, correlate with unobserved variables (ϵ_{ijt}), Assumption 1 (exogeneity) is compromised. As a result, it becomes infeasible to integrate out ϵ_{ijt} in the aggregate demand function, as we did in Equation 1. This means that the aggregate demand function loses its property of permutation invariance with respect to the observable characteristics of competitors. To address this, we will build on the approach developed in Petrin and Train (2010) to allow for endogenous observable features. Without loss of generality, we assume that the price p_{jt} is the endogenous variable and all other characteristics of the product X_{jt} are exogenous variables. i.e.,

$$\mathbb{E}[p_{jt} \cdot \varepsilon_{ijt}] \neq 0 \quad \text{and} \quad \mathbb{E}[X_{jt} \cdot \varepsilon_{ijt}] = 0,$$

Given valid instruments IV_{jt}, we can express p_{jt} as

$$p_{jt} = \gamma\left(X_{jt}, IV_{jt}\right) + \mu_{jt}, \tag{3}$$

At this point, no specific assumptions are made regarding the function γ. However, in the subsequent inference section, we will discuss that the estimator of γ must be estimable at $n^{-1/2}$ in order to construct valid confidence intervals. Next, to address the issue of price endogeneity, we impose a mild restriction on the space of choice functions we consider.

Assumption 4 (Linear Separability). The unobserved product characteristics can be expressed as the sum of an endogenous (CF) and exogenous component

$$\varepsilon_{ijt} = CF\left(\mu_{jt}; \lambda\right) + \tilde{\varepsilon}_{ijt}, \tag{4}$$

where $\mathbb{E}[p_{jt} \cdot \tilde{\varepsilon}_{ijt}] = 0$.

This assumption implies that, after controlling for μ_{jt} using the control function CF, the endogenous variable p_{jt} is uncorrelated with the error term ε_{ijt} in the model, thus it becomes exogenous. Then, we can re-write the index tuple u_{ijt} as

$$u_{ijt} = \{X_{jt}, p_{jt}, CF\left(\mu_{jt}; \lambda\right) + \tilde{\varepsilon}_{ijt}\}, \tag{5}$$

such that $\mathbb{E}\left[\tilde{\varepsilon}_{ijt} | (X_{jt}, p_{jt}, \mu_{jt})\right] = 0$

Theorem 2. For any offer set $\mathcal{S}_t \subset \{1, 2, 3, \ldots, J_t\}$, if a choice function $\pi : \{u_{ijt} : j \in \mathcal{S}_t\} \to \mathbb{R}^{|\mathcal{S}_t|}$ where u_{ijt} represents the index tuple $\{X_{jt}, p_{jt}, I_{it}, \varepsilon_{ijt}\}$ satisfies assumption 2 to 4. Then under the condition of knowing the true function (γ_0) of γ, there exists suitable ρ, ϕ_1 and ϕ_2 such that

$$\pi_{jt} = \rho(\phi_1(X_{jt}, p_{jt}, \mu_{jt}(\gamma_0)) + \sum_{k \neq j, k \in \mathcal{S}} \phi_2(X_{kt}, p_{jt}, \mu_{kt}(\gamma_0))),$$

The result follows straightforwardly from the observation that after controlling for $CF(\mu_{jt}; \lambda)$ the unobservable component $\tilde{\varepsilon}$ is exogenous. This implies the aggregate demand function is invariant under any permutation applied to competitors of product j. The result demonstrates that endogeneity can be addressed by using the residuals from Equation 4 along with product observable characteristics simply as an additional set of features.

[2] $\rho\left(\phi_1(x_{jt}) + \sum_{k \neq j} \phi_2(x_{kt})\right) = \rho\left(\begin{bmatrix} exp(x_{jt}) \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ \sum_{k \neq j} exp(x_{kt}) \end{bmatrix}\right) = \rho\left(\begin{bmatrix} exp(x_{jt}) \\ \sum_{k \neq j} exp(x_{kt}) \end{bmatrix}\right) = \frac{exp(x_{jt})}{exp(x_{jt}) + \sum_{k \neq j} exp(x_{kt})}$

3 Numerical Experiments

In this section, we first conduct several experiments to evaluate our estimator's predictive performance by applying it to different choice models. To assess the predictive performance of our model, we focus on three estimators: market share ($\hat{\pi_{jt}}$), own-elasticity ($\frac{\partial \hat{\pi}_{jt}}{\partial P_{jt}}$) and cross-elasticity ($\frac{\partial \hat{\pi}_{jt}}{\partial P_{k \neq jt}}$). For comparison, we also include the predictive performance of four baseline models : 1) MNL; 2) RCL; 3) A standard neural network-based non-parametric method (NP): We tune the hyperparameters – number of layers, number of nodes in each layer, learning rate, and the number of epochs using 5-fold cross-validation for each data generation. We detail the space of hyperparameters in Appendix F. We also apply the ReLU activation for each layer. 4) A "mean" predictor for all data points (Mean) where we predict the market share to be the same for all products in a market and set it equal to the average market share across all products in the dataset.

One advantage of our model compared to the standard neural network-based non-parametric method is that the parameter of our model does not scale with the number of products. The standard neural network uses the stacked products' feature as input and has $(J \times K + 1) \times h_1$ parameters in the input layer, where h_1 denotes the size of the first hidden layer, while our model only use product features with dimension $K + 1$ as input. Across the simulation runs, we observe that the selected neural network has more parameters than our model.

Recent literature (Allenby et al., 2004) has highlighted a growing trend towards the adoption of non-linear utility functions. In traditional parametric choice models like RCL and MNL, the oversight of non-linear relationships between features and utilities can introduce biases in the estimates. Conversely, non-parametric estimators are adept at capturing these non-linear patterns directly from the data. As part of our analysis, we focus on data generated from a random coefficients logit model with non-linear transformations applied to observable features. Without loss of generalizabilty, we consider the case where there is only one feature x on which we apply non-linear transformation $g(x)$. Specifically, we consider two funtions of $g(x)$ –

a. log(): $g(x) = log(\mid 16x - 8 \mid +1)\text{sign}(x - 0.5)$

b. sin(): $g(x) = sin(x)$

We also estimate baseline models (MNL and RCL) as comparison. Regarding the MAE of predicted market shares, our model surpasses the RCL model by a factor of 8X and 4X across transformations (a) and (b), respectively. Similarly, considering the MAE of predicted own-elasticity in transformations (a) and (b), our model outperforms RCL by factors of 20X and 2.5X, respectively. For the MAE of predicted cross-elasticity, our model is 2X and 1.5X superior to RCL across transformations (a) and (b), respectively. It's worth noting that while our model consistently outperforms the NP method across metrics, the NP method still shows better performance than both RCL and MNL in terms of MAE and RMSE for estimated own-elasticity ($\frac{\partial \hat{\pi}_{jt}}{\partial P_{jt}}$), underscoring the strengths of neural network-driven approaches in navigating non-linearities.

Table 1: Non-linearity - Predicted Market Shares($\hat{\pi_{jt}}$)

#	True Model	Our model		MNL		RCL		NP		Mean		No. Obs.
		MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	
0	RCL-log()	0.0025	0.0063	0.0358	0.0361	0.0213	0.0309	0.0588	0.1235	0.0836	0.1401	200
1	RCL-sin()	0.0029	0.0046	0.0281	0.0340	0.0102	0.0172	0.0315	0.0449	0.0388	0.0527	200

Note: This table presents the results when we add non-linear transformation in data generation. We generate using the Random Coefficient Logit (RCL) model, with 10 products and 100 markets, while only considering a single non-linearly transformed feature, which is the price.

4 Conclusion

Choice models are fundamental in understanding consumer behavior and informing business decisions. Over the years, various methods, both parametric and non-parametric, have been developed to represent consumer behavior. In this paper, we propose a fundamental characterization of choice models that combines the tractability of traditional choice models and the flexibility of non-parametric estimators. This characterization specifically tackles the challenge of high dimensionality in choice

Table 2: Non-linearity - Estimated own-Elasticity ($\frac{\partial \hat{\pi}_{jt}}{\partial P_{jt}}$)

#	True Model	Our Model		MNL		RCL		NP		No. Obs
		MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	
0	RCL-log()	0.0566	0.1523	5.4278	2.8249	1.1961	2.1135	0.6119	0.9085	16000
1	RCL-sin()	0.0609	0.2820	0.6229	1.1350	0.1777	0.4246	0.4057	1.0787	16000

Note: This table presents the results when we add non-linear transformation in data generation. We generate using the Random Coefficient Logit (RCL) model, with 10 products and 100 markets, while only considering a single non-linearly transformed feature, which is the price.

Table 3: Non-linearity - Estimated Cross-Elasticity ($\frac{\partial \hat{\pi}_{jt}}{\partial P_{k\neq jt}}$)

#	True Model	Our Model		MNL		RCL		NP		No. Obs
		MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	
0	RCL-log()	0.0150	0.0543	0.0436	0.1389	0.0372	0.4414	0.2552	0.5444	144000
1	RCL-sin()	0.0226	0.1047	0.0471	0.1794	0.0354	0.1751	0.1448	0.3357	144000

Note: This table presents the results when we add non-linear transformation to features. To limit the influence of other factors that may affect the result, we consider the case when there is only one feature (price).

systems and facilitates flexible estimation of choice functions. Through extensive simulations, we validate the efficacy of our model, demonstrating its superior ability to capture a range of consumer behaviors that traditional choice models fail to capture. We also show how to address the endogeneity issue and estimate counterfactuals in our characterization. Furthermore, leveraging the recent strides in the automatic debiased machine learning literature, we offer an inference procedure that constructs confidence intervals on relevant objects, such as price elasticities. Finally, we apply our method to the automobile dataset from Berry et al. (1995). Our empirical analysis affirms that our model produces results that align well with the extant literature.

Our paper opens many avenues for future research. First, we focus on using neural network-based estimators. However, estimators, such as Gaussian processes and Gradient boosting-based estimators can be adopted to estimate the proposed functionals. Second, we also only consider a very standard multilayer RELU neural network for each component of our model. Another potential future direction could be exploring the connection between transformer networks (Vaswani et al., 2017) and set functions as the attention mechanism used in these architectures have a very similar functional form.

References

J. Abaluck, G. Compiani, and F. Zhang. A method to estimate discrete choice models that is robust to consumer search. Technical report, National Bureau of Economic Research, 2020.

D. Ackerberg, X. Chen, J. Hahn, and Z. Liao. Asymptotic efficiency of semiparametric two-step gmm. *Review of Economic Studies*, 81(3):919–943, 2014.

C. Ai and X. Chen. Estimation of possibly misspecified semiparametric conditional moment restriction models with different conditioning variables. *Journal of Econometrics*, 141(1):5–43, 2007.

G. M. Allenby, T. S. Shively, S. Yang, and M. J. Garratt. A choice model for packaged goods: Dealing with discrete quantities and quantity discounts. *Marketing Science*, 23(1):95–108, 2004.

S. Berry, J. Levinsohn, and A. Pakes. Automobile prices in market equilibrium. *Econometrica: Journal of the Econometric Society*, pages 841–890, 1995.

J. Blanchet, G. Gallego, and V. Goyal. A markov chain approximation to choice modeling. *Operations Research*, 64(4):886–905, 2016.

S. Chatterjee and J. Jafarov. Prediction error of cross-validated lasso. *arXiv preprint arXiv:1502.06291*, 2015.

V. Chernozhukov, W. K. Newey, V. Quintas-Martinez, and V. Syrgkanis. Automatic debiased machine learning via neural nets for generalized linear regression. *arXiv preprint arXiv:2104.14737*, 2021.

V. Chernozhukov, J. C. Escanciano, H. Ichimura, W. K. Newey, and J. M. Robins. Locally robust semiparametric estimation. *Econometrica*, 90(4):1501–1535, 2022a.

V. Chernozhukov, W. K. Newey, and R. Singh. Automatic debiased machine learning of causal and structural effects. *Econometrica*, 90(3):967–1027, 2022b.

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

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

M. J. Funk, D. Westreich, C. Wiesen, T. Stürmer, M. A. Brookhart, and M. Davidian. Doubly robust estimation of causal effects. *American journal of epidemiology*, 173(7):761–767, 2011.

X. Gabaix. Behavioral inattention. In *Handbook of behavioral economics: Applications and foundations 1*, volume 2, pages 261–343. Elsevier, 2019.

A. Gandhi and J.-F. Houde. Measuring substitution patterns in differentiated-products industries. *NBER Working paper*, (w26375), 2019.

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

L. Grigolon and F. Verboven. Nested logit or random coefficients logit? a comparison of alternative discrete choice models of product differentiation. *Review of Economics and Statistics*, 96(5): 916–935, 2014.

J. A. Hausman and D. A. Wise. A conditional probit model for qualitative choice: Discrete decisions recognizing interdependence and heterogeneous preferences. *Econometrica: Journal of the econometric society*, pages 403–426, 1978.

D. A. Hirshberg and S. Wager. Debiased inference of average partial effects in single-index models: Comment on wooldridge and zhu. *Journal of Business & Economic Statistics*, 38(1):19–24, 2020.

E. Honka, A. Hortaçsu, and M. Wildenbeest. Empirical search and consideration sets. In *Handbook of the Economics of Marketing*, volume 1, pages 193–257. Elsevier, 2019.

H. Ichimura and W. K. Newey. The influence function of semiparametric estimators. *Quantitative Economics*, 13(1):29–61, 2022.

W. A. Kamakura and G. J. Russell. A probabilistic choice model for market segmentation and elasticity structure. *Journal of marketing research*, 26(4):379–390, 1989.

D. McFadden and K. Train. Mixed mnl models for discrete response. *Journal of applied Econometrics*, 15(5):447–470, 2000.

D. McFadden et al. Conditional logit analysis of qualitative choice behavior. 1973.

N. Mehta, S. Rajiv, and K. Srinivasan. Price uncertainty and consumer search: A structural model of consideration set formation. *Marketing science*, 22(1):58–84, 2003.

W. K. Newey. The asymptotic variance of semiparametric estimators. *Econometrica: Journal of the Econometric Society*, pages 1349–1382, 1994.

A. Petrin and K. Train. A control function approach to endogeneity in consumer choice models. *Journal of marketing research*, 47(1):3–13, 2010.

J. M. Robins, A. Rotnitzky, and L. P. Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89(427):846–866, 1994.

A. Singh, K. Hosanagar, and A. Gandhi. Machine learning instrument variables for causal inference. In *Proceedings of the 21st ACM Conference on Economics and Computation*, pages 835–836, 2020.

K. E. Train. *Discrete choice methods with simulation*. Cambridge university press, 2009.

K. E. Train, D. L. McFadden, and M. Ben-Akiva. The demand for local telephone service: A fully discrete model of residential calling patterns and service choices. *The RAND Journal of Economics*, pages 109–123, 1987.

E. Van Nierop, B. Bronnenberg, R. Paap, M. Wedel, and P. H. Franses. Retrieving unobserved consideration sets from household panel data. *Journal of Marketing Research*, 47(1):63–74, 2010.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

M. J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

M. Zaheer, S. Kottur, S. Ravanbakhsh, B. Poczos, R. R. Salakhutdinov, and A. J. Smola. Deep sets. *Advances in neural information processing systems*, 30, 2017.

Appendices

A Inference

The aim of this paper is to estimate choice functions flexibly using non-parametric estimators. However, often in social science contexts, one is also interested in conducting inference over some economic objects. Note that because non-parametric regression functions are estimated at a slower rate compared to parametric regressions, it is often infeasible to construct confidence intervals directly on the estimated $\hat{\pi}$. However, it is generally possible to perform inference and construct valid confidence intervals for specific economic objects that are functions of π. In this section, we will provide example of one such important economic object and demonstrate how to construct valid confidence intervals for it. This will be done by leveraging the recent advances in automatic debiased machine learning as shown in the works of Ichimura and Newey (2022); Chernozhukov et al. (2022b,a, 2021), and others. However, unlike existing automatic debiased machine learning setups we also have to account for an additional first-stage estimator $\hat{\gamma}$.

In demand estimation, researchers are often interested in estimating the average effect of a price change on the demand for a product, as it can significantly influence market dynamics, pricing strategies, and regulatory decisions. To proceed with our analysis, let $w_{jt} = (y_{jt}, p_{jt}, x_{jt}, \{p_{kt}, x_{kt}\}_{k\neq j})$ and $z_{jt} = (p_{jt}, x_{jt}, \{p_{kt}, x_{kt}\}_{k\neq j})$ represent the variables associated with product j in market t. Here, $p_{jt} \in \mathbb{C}$ denotes the observed prices, $x_{jt} \in \mathbb{C}^d$ represents other product characteristics and $y_{jt} \in \mathbb{R}$ refers to the observed demand for product j in market t, such as market shares or log shares. Note that either the observed price (p_{jt}) or other characteristics (x_{jt}) could be endogenous. For simplicity and without loss of generality, we focus on p_{jt} as the endogenous variable in the following analysis.

The average effect of a price change[3] can be expressed as the difference between the demand function $\pi_{jt}(\cdot;\gamma)$ evaluated at the original price p_{jt} and at the price incremented by Δp_{jt}, given by the following expression:

$$m(w_{jt}, \pi(\cdot;\gamma)) = \pi(p_{jt} + \Delta p_{jt}, x_{jt}, \{p_{kt}, x_{kt}\}_{k\neq j}); \gamma) - \pi(p_{jt}, x_{jt}, \{p_{kt}, x_{kt}\}_{k\neq j}); \gamma).$$

The parameter of interest, θ_0, is the expected value of this price change effect over the true population distribution[4] of w_{jt}, which can be calculated as:

$$\theta_0 = \mathbb{E}[m(w_{jt}, \pi(\cdot;\gamma))] = \mathbb{E}[\pi(p_{jt} + \Delta p_{jt}, x_{jt}, \{x_{kt}\}_{k\neq j}; \gamma) - \pi(p_{jt}, x_{jt}, \{x_{kt}\}_{k\neq j}; \gamma)].$$

In summary, the average effect of a price change on demand, denoted by θ_0, is calculated by evaluating the difference between the demand function at the original price and at the price incremented by Δp_{jt}, and then computing the expected value of this difference.

In practice, we estimate θ_0 by computing its empirical analog using the estimated demand function $\hat{\pi}$ and first-stage estimator $\hat{\gamma}$, i.e.,

$$\hat{\theta} = \frac{1}{n}\sum_{t=1}^{n} m(w_{jt}, \hat{\pi}(z_{jt}; \hat{\gamma})), \tag{A1}$$

where n is the number of observations. When parametric methods are employed to estimate $\hat{\pi}$ and $\hat{\gamma}$, the estimator for $\hat{\theta}$ is generally $\sqrt{n}$-consistent, assuming that the model is correctly specified. However, $\sqrt{n}$-consistency may not hold when non-parametric estimators are used, particularly if the first-order bias does not vanish at a rate of $\sqrt{n}$. Irrespective of the method used to estimate π, this is often the case, as flexible estimation of π always requires some form of regularization and/or model selection. Debiasing techniques are required to mitigate the effects of regularization and/or model selection when learning flexible demand models. These approaches can help improve the performance of the estimator and facilitate valid inference with $\hat{\theta}$. We therefore adapt recent debiasing

[3] The expression for the average effect of a price change can be adapted to represent average price elasticity by placing the known and fixed value of Δp_{jt} in the denominator.

[4] We assume the data reflects the true population.

techniques developed in recent automatic debiased machine learning literature (see Chernozhukov et al. (2022b)). Specifically, we will focus on problems where there exists a square-integrable random variable $\alpha_0(z)$ such that $\forall\ ||\gamma - \gamma_0||$ small enough –

$$\begin{aligned}\mathbb{E}[m(w_{jt}, \pi(z_{jt};\gamma))] &= \mathbb{E}[\alpha_0(z_{jt})\pi(z_{jt};\gamma)]\\ &\forall \pi \text{ with } \mathbb{E}[\pi_{jt}(z_{jt};\gamma)^2] < \infty\end{aligned}$$

By the Riesz representation theorem, the existence of such $\alpha_0(z_{jt})$ is equivalent to $\mathbb{E}[m(w_{jt}, \pi(z_{jt};\gamma))]$ being a mean square continuous functional of $\pi(z_{jt};\gamma)$. Henceforth, we refer to $\alpha_0(z)$ as Riesz representer (or RR). Newey (1994) shows that the mean square continuity of $\mathbb{E}[m(w_{jt}, \pi_{jt}(z_{jt};\gamma))]$ is equivalent to the semiparametric efficiency bound of θ_0 being finite. Thus, our approach focuses on regular functionals. Similar uses of the Riesz representation theorem can be found in Ai and Chen (2007), Ackerberg et al. (2014), Hirshberg and Wager (2020), and Chernozhukov et al. (2022b) among others. The debiasing term in this case takes the form $\alpha(z_{jt})(y_{jt} - \pi(z_{jt};\gamma))$. To see that, consider the score $m(w_{jt}, \pi(z_{jt};\gamma)) + \alpha(z_{jt})(y_{jt} - \pi(z_{jt};\gamma)) - \theta_0$. It satisfies the following mixed bias property:

$$\begin{aligned}&\mathbb{E}[m(w_{jt}, \pi(z_{jt};\gamma)) + \alpha(z_{jt})(y_{jt} - \pi(z_{jt};\gamma)) - \theta_0]\\ &\qquad = -\mathbb{E}\left[(\alpha(z_{jt}) - \alpha_0(z_{jt}))\left(\pi(z_{jt}) - y_{jt}\right)\right].\end{aligned}$$

This property implies double robustness (Robins et al., 1994; Funk et al., 2011) of the score. That is, if either $\alpha(z_{jt})$ is correctly estimated, which would mean $\alpha(z_{jt}) - \alpha_0(z_{jt}) = 0$, or $\pi(z_{jt})$ is correctly estimated, implying $\pi(z_{jt}) - y_{jt} = 0$, then the term $(\alpha(z_{jt}) - \alpha_0(z_{jt}))(\pi(z_{jt}) - y_{jt})$ will be zero. This results in the score going to zero, thereby making the estimator consistent for θ_0. A debiased machine learning estimator of θ_0 can be constructed from this score and first-stage learners $\widehat{\pi}$ and $\widehat{\alpha}$. Let $\mathbb{E}_n[\cdot]$ denote the empirical expectation over a sample of size n, i.e., $\mathbb{E}_n[x_i] = \frac{1}{n}\sum_{i=1}^{n} x_i$. We consider:

$$\widehat{\theta} = \mathbb{E}_n[m(w_{jt};\widehat{\pi}) + \widehat{\alpha}(z_{jt})(y_{jt} - \widehat{\pi}(z_{jt}))].$$

The mixed bias property implies that the bias of this estimator will vanish at a rate equal to the product of the mean-square convergence rates of $\widehat{\alpha}$ and $\widehat{\pi}$. Therefore, in cases where the demand function π can be estimated very well, the rate requirements on $\widehat{\alpha}$ will be less strict, and vice versa. More notably, whenever the product of the meansquare convergence rates of $\widehat{\alpha}$ and $\widehat{f}$ is larger than $\sqrt{n}$, we have that $\sqrt{n}\left(\widehat{\theta} - \theta_0\right)$ converges in distribution to centered normal law $N\left(0, \mathbb{E}\left[\psi_0(w_{jt})^2\right]\right)$, where

$$\psi_0(w_{jt}) := m\left(w_{jt};\pi_0\right) + \alpha_0(z_{jt})\left(y_{jt} - \pi_0(z_{jt})\right) - \theta_0$$

as proven formally in Theorem 3 of Chernozhukov et al. (2022b). Results in Newey (1994) imply that $\mathbb{E}\left[\psi_0(w_i)^2\right]$ is a semiparametric efficient variance bound for θ_0, and therefore the estimator achieves this bound.

Theorem 3. [Chernozhukov et al. (2021)] One can view the Riesz representer as the minimizer of the loss function:

$$\begin{aligned}\alpha_0 &= \arg\min_{\alpha}\mathbb{E}\left[(\alpha(z_{jt}) - \alpha_0(z_{jt}))^2\right]\\ &= \arg\min_{\alpha}\mathbb{E}\left[\alpha(z_{jt})^2 - 2\alpha_0(z_{jt})\alpha(z_{jt}) + \alpha_0(z_{jt})^2\right]\\ &= \arg\min_{\alpha}\mathbb{E}\left[\alpha(z_{jt})^2 - 2m(w_{jt};\alpha)\right],\end{aligned}$$

In our earlier discussions, we employed the moment function of π, whereas in Theorem 3, we focus on the moment function of α. This shift is justified by the Riesz Representation Theorem, which implies $\mathbb{E}[m(w_{jt};\pi)] = \mathbb{E}[\alpha_0(z_{jt})\pi(z_{jt})]$. Given that π can represent any function, substituting α for π is permissible, thereby validating the transition from the second to the third line in Theorem 3. We use the above theorem to flexibly estimate the RR. The advantage of this approach is that it eliminates the need to derive an analytical form for the RR estimator, allowing it to be addressed as a simple computational optimization problem.

Theorem 4. [Chernozhukov et al. (2021)] Let δ_n be an upper bound on the critical radius (Wainwright (2019)) of the function spaces:

$$\{z \mapsto \zeta\left(\alpha(z)-\alpha_0(z)\right) : \alpha \in \mathcal{A}_n, \zeta \in [0,1]\} \text{ and}$$
$$\{w \mapsto \zeta\left(m(w;\alpha)-m\left(w;\alpha_0\right)\right) : \alpha \in \mathcal{A}_n, \zeta \in [0,1]\}$$

and suppose that for all f in the spaces above: $\|f\|_\infty \leq 1$. Suppose, furthermore, that m satisfies the mean-squared continuity property:

$$\mathbb{E}\left[\left(m(w;\alpha)-m\left(w;\alpha'\right)\right)^2\right] \leq M\left\|\alpha-\alpha'\right\|_2^2$$

for all $\alpha, \alpha' \in \mathcal{A}_n$ and some $M \geq 1$. Then for some universal constant C, we have that w.p. $1-\zeta$:

$$\left\|\widehat{\alpha}-\alpha_0\right\|_2^2 \leq C(\delta_n^2 M + \frac{M \log(1/\zeta)}{n} \\ + \inf_{\alpha_* \in \mathcal{A}_n}\left\|\alpha_*-\alpha_0\right\|_2^2\Big)$$

The critical radius has been widely studied in various function spaces, such as high-dimensional linear functions, neural networks, and superficial regression trees, often showing $\delta_n = O\left(d_n n^{-1/2}\right)$, where d_n represents the effective dimensions of the hypothesis spaces (Chernozhukov et al. (2021)). In our research, we focus on applying Theorem 2.1 from an application standpoint to neural networks.

Assumption 5. i) $\alpha_0(z)$ and $\forall\ ||\gamma-\gamma_0||$ small enough $\mathbb{E}[(y-\pi_0(z_{jt};\gamma))^2|z_{jt}]$ are bounded ii) $\mathbb{E}[m(w_{jt},\pi_0(z_{jt};\gamma_0))^2] < \infty$

These assumptions are standard regularity conditions used in the automatic machine learning literature.

Assumption 6. i) $\forall\ ||\gamma-\gamma_0||$ small enough $||\hat{\pi}(;\gamma)-\pi_0(;\gamma)|| \xrightarrow{p} 0$ and $||\hat{\alpha}-\alpha_0|| \xrightarrow{p} 0$; ii) $\sqrt{n}||\hat{\alpha}-\alpha||||(\hat{\pi}(;\gamma)-\pi_0(;\gamma)|| \xrightarrow{p} 0$; iii) $\hat{\alpha}$ is bounded; (iv) $\sqrt{n}||\hat{\gamma}-\gamma_0|| \xrightarrow{p} 0$

Intuitively these assumptions mean that (i) the estimator of both π and α should be consistent for values of γ in a close enough neighborhood of γ_0. Further, it requires that the product of mean square error of $\hat{\alpha}$ and mean square error of π should vanish at $\sqrt{n}-$ rate. This can be achieved if both these terms converge at least at $n^{-1/4}$ rate. Finally, we also assume that the first stage estimator $\hat{\gamma}$ is estimable at $n^{-1/2}$ rate. This limits the class of functions one can use to estimate γ.

Assumption 7. $m(w,\pi)$ is linear in π and there is $C > 0$ such that

$$|E\left[m(w,\pi)-\theta_0+\alpha_0(z)(y-\pi(z;\gamma))\right]| \leq C\left\|\pi-\pi_0\right\|^2$$

Proposition 1. *If Assumptions 5-7 are satisfied then for* $V = E[\{m(w,\pi_0(z;\gamma_0))-\theta_0$ $+\alpha_0(z)(y-\pi_0(z;\gamma_0))\}^2]$,

$$\sqrt{n}(\hat{\theta}-\theta_0) \xrightarrow{D} N(0,V), \hat{V} \xrightarrow{p} V.$$

We show the proof in Appendix B. This theorem shows that if $\hat{\gamma}$ is estimable at a fast enough rate one can still construct valid confidence intervals for $\hat{\theta}$. This result can be shown following similar arguments as in Chernozhukov et al. (2022a).

A.1 Estimation Outline

Consider the dataset $\{y_t, z_t, IV_t\}_{t=1}^n$ is independently and identically distributed.

- Stage 1 ($\hat{\gamma}$): We estimate γ by regressing the endogenous variable on the exogenous instruments. We then calculate the residual $\hat{\mu}$ with estimated $\hat{\gamma}$.
- Stage 2 ($\hat{\pi}$):
  - Stage 2a (Data partition): We randomly split the data into L folds such that the data $D_l := \{y_t, z_t\}_{t \in I_l}$, where I_l denotes the l^{th} partition.
  - Stage 2b (Estimation): In the second stage for each fold I_l, we estimate both the choice function ($\hat{\pi}$) and the Riesz estimator ($\hat{\alpha}$) on the left out data $D_l^c := \{y_t, x_t\}_{t \notin I_l}$

$$\hat{\pi}_l =_{f \in \mathcal{F}} \frac{1}{\sum_{t \in D_{lc}} J_t} \sum_{t \in D_{lc}} \sum_{j \in J_t} [(y_{jt}-\pi(z_{jt};\gamma))^2] \tag{A2}$$

$$\hat{\alpha}_l =_{\alpha \in \mathcal{A}} \frac{1}{\sum_{t \in D_{l^c}} J_t} \sum_{t \in D_{l^c}} \sum_{j \in J_t} \left[\alpha(z_{jt})^2 - 2m(w_{jt}; \alpha) \right]. \tag{A3}$$

Based on Theorem 2, instead of estimating a function π, we decompose the estimation to 3 components: ρ, ψ_1, and ψ_2. Specifically, for each component of our model (ϕ_1, ϕ_2 and ρ), we use a standard 3-layer neural network, and this is implemented without further hyperparameter tuning. We implement ReLU activation function at each layer as it is standard in feedforward designs due to simplicity and computation efficiency in gradients. Specifically, we pass the focal product's characteristics (x_{jt}) and the residuals (μ_{jt}) estimated from first stage to the ϕ_1. In parallel, we pass each other product's (of the same market) characteristics (x_{kt}) and the residuals (μ_{kt}) to a same ϕ_2 then sum the output up. The output of ϕ_1 and ϕ_2 have the same data structure (e.g., a 64-dimension vector). Next, we pass the summation of the output of ϕ_1 and ϕ_2 to a third neural network ρ. The output of ρ is a scalar which represents the market share of the focal product jt. We use the same structure when estimating α. The only difference is, the loss function of α is not based on the difference between the observed and the predicted. Instead, the loss function is based on the difference between α and the moment function of α as stated in Theorem 3.

- Stage 2c (Cross-fitting): Now we use the cross-fitting technique, same as Chernozhukov et al. (2021) to reduce the bias when estimating $\hat{\theta}$. Specifically, we use the estimators ($\hat{\pi}$ and $\hat{\pi}$) estimated on D_l^c to estimate the $\hat{\theta}_l$ of I_l. By applying cross-fitting, we ensure that the nuisance functions and the parameters are estimated on separate, non-overlapping datasets. This approach diminishes the risk of overfitting and enhances the robustness of our estimation. And finally, to estimate θ, we average it out across all folds. Thus the estimator for θ_0 and its variance can be given as follows –

$$\hat{\theta} = \frac{1}{n} \sum_{\ell=1}^{L} \sum_{t \in D_\ell^c} \left\{ m\left(w_t, \hat{\pi}_\ell\right) + \hat{\alpha}_\ell\left(z_t\right)\left(y_t - \hat{\pi}_\ell(x_t)\right) \right\}$$

$$\hat{V} = \frac{1}{n} \sum_{\ell=1}^{L} \sum_{t \in D_\ell^c} \hat{\psi}_{t\ell}^2, \quad \hat{\psi}_{t\ell} = m\left(w_t, \hat{\pi}_\ell\right) - \hat{\theta} + \hat{\alpha}_\ell\left(z_t\right)\left(y_t - \hat{\pi}_\ell(x_t)\right)$$

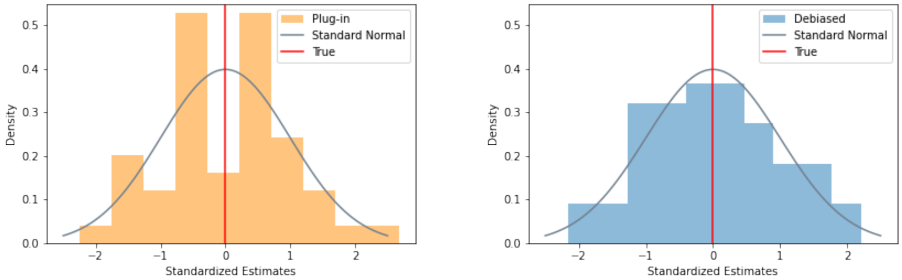
A.2 Inference and Coverage Analysis

We demonstrate the performance of the debiasing and inference procedure. The objective is to demonstrate the validity of the estimated confidence intervals. To this end, we estimate the average effect of a 1% change in own price on demand over all products ($\hat{\theta}$) and compute the corresponding confidence intervals of this effect. The difference between this section and section **??** lies in both the estimators and the methods. In terms of estimators, in section **??**, we predict the market share ($\hat{\pi_{jt}}$), own-elasticity ($\frac{\partial \hat{\pi}_{jt}}{\partial P_{jt}}$) and cross-elasticity ($\frac{\partial \hat{\pi}_{jt}}{\partial P_{k \neq jt}}$) for individual products. In contrast, the object of interest in this section is *average* effect of price on demand across all products ($\hat{\theta}$). As a result, in section **??**, we did not use debiasing techniques which we apply here. It's important to emphasize that in our approach, constructing a confidence interval is viable only for aggregate measures, not for individual observations.

To simulate the data, we consider a random-coefficients logit model of demand with 3 products across 100 markets. We set the true model parameters to be $\beta_{ik} \sim \mathcal{N}(1, 0.5), \alpha_i \sim \mathcal{N}(-1, 0.5)$. The effect of a 1% increase in a product's price is given by

$$\theta_0 = \mathbb{E}[m(w_i, \pi] = \mathbb{E}[\pi(p_{jt} * (1.01), x_{jt}, \{x_{kt}\}_{k \neq j}) - \pi(p_{jt}, x_{jt}, \{x_{kt}\}_{k \neq j})],$$

As discussed earlier, one way to estimate this effect is to compute the sample analog of this using the estimated $\hat{\pi}$, such that $\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} m(w_i, \hat{\pi})$. However, as we pointed out earlier, the distribution of $\hat{\theta}$ might not be asymptotically normal. To demonstrate this, in Figure A1a, we display the histogram of the estimated effect across 50 random samples by using the plug-in method. We standardize the estimates by subtracting the mean and then dividing by the standard deviation and plot them against the standard normal distribution. As one can observe the distribution appears multi-modal and deviates from a standard normal distribution. Next, we use our proposed debiased estimator and plot the standardized estimates across 50 samples of draws in Figure A1b. As one can note the resultant distribution with the debiased estimator is much closer to a standard normal. Finally, we calculate the 95% confidence intervals using our debiased estimator across 50 random draws. In Table A1, we report the bias (mean absolute error from all draws) and the coverage i.e., the percentage of times the true parameter is covered in the estimated confidence intervals. We find that bias across both data-generating processes (RCL and MNL) is notably low, reflecting only a -0.0001 difference from the true effect. The coverage rate of the 95% confidence interval in our corrected model is 90%, indicating good coverage. This shows that our debiased estimator can be used to conduct valid inference in finite samples.



(a) Distribution of Estimated Average Effect of Price Change with Plug-in

(b) Distribution of Estimated Average Effect of Price Change with the Debiased Estimator

Figure A1: Distribution of Estimated Average Effect of Price Change

Note: The figure shows the distribution of the standardized plug-in and debiased estimators of the average effect of 1% change in price on demand. To simulate the data, we consider 3 products across 100 markets with 5 non-price features using RCL for 50 samples. For each sample, we set the true model parameters to be $\beta_{ik} \sim \mathcal{N}(1, 0.5), \alpha_i \sim \mathcal{N}(-1, 0.5)$. Figure A1a displays the distribution of the estimated effect with the Plug-in method and Figure A1b shows the result when employing the debiased estimator.

Table A1: Inference Coverage Analysis

True Model	True Effect	Bias	95% CI Cov.
RCL	-0.0013	-0.0001	90%
MNL	-0.0016	-0.0001	90%

Note: This table presents the bias and coverage rate of 95% confidence interval using our debiased estimator of the average effect of 1% change in price on demand. To simulate the data, we consider 3 products across 100 markets with 5 non-price features using RCL and MNL for 50 samples of draws. For each RCL sample, we set the true model parameters to be $\beta_{ik} \sim \mathcal{N}(1, 0.5), \alpha_i \sim \mathcal{N}(-1, 0.5)$. For each MNL sample, we set the true model parameters to be $\beta_{ik} = 1, \alpha_i = -1$.

B Appendix for the Proof of Main Results

Theorem 1. For any offer set $\mathcal{S}_t \subset \{1, 2, 3, \ldots, J_t\}$, if a choice function $\pi : \{u_{ijt} : j \in \mathcal{S}_t\} \to \mathbb{R}^{|\mathcal{S}_t|}$ where u_{ijt} represents the index tuple $\{X_{jt}, I_{it}, \varepsilon_{ijt}\}$ satisfies Assumption 1, 2 and 3, then there exists suitable ρ, ϕ_1 and ϕ_2 such that

$$\pi_{jt} = \rho(\phi_1(X_{jt}) + \sum_{k \neq j, k \in \mathcal{S}_t} \phi_2(X_{kt})),$$

Proof. The sufficiency follows by observing that the function $\pi_{jt} = \rho(\phi_1(X_{jt}) + \sum_{k \in S \setminus \{j\}} \phi_2(X_{kt}))$ satisfies assumption 2 and 3. To prove necessity, first consider $\mathbb{E} = \{2n \mid n \in \mathbb{N}\}$ and $\mathbb{O} = \{2n + 1 \mid n \in \mathbb{N}\}$ as the set of all even and odd natural numbers respectively. Next, to show that all functions can be decomposed in the above manner, we begin by noting that there must be a mapping from the elements to the set of even number and odd numbers respectively, since the elements belong to a countable universe $\mathbb{C}^k$. Let these mappings be denoted by $c^e : \mathbb{C}^k \to \mathbb{E}$ and $c^o : \mathbb{C}^k \to \mathbb{O}$. Now if we let $\phi_1(x) = 4^{-c^e(x)}$ and $\phi_2(x) = 4^{-c^o(x)}$ then $\phi_1(x) + \sum_{x \in S \setminus \{j\}} \phi_2(x)$ constitutes an unique representation for every product j and competing assortment $\mathcal{S} \setminus \{j\}$. Now a function $\rho : \mathbb{R} \to \mathbb{R}$ can always be constructed such that $\pi_{jt} = \rho\left(\phi_1(x_{jt}) + \sum_{k \in S \setminus \{j\}} \phi_2(x_{kt})\right)$.

Proposition 1. *If Assumptions 5-7 are satisfied then for* $V = E[\{m(w, \pi_0(z;\gamma_0)) - \theta_0$ $+\alpha_0(z)(y - \pi_0(z;\gamma_0))\}^2]$,

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{D} N(0, V), \hat{V} \xrightarrow{p} V.$$

Proof. To show the asymptotic normality we will first verify the Assumptions 1-3 of Chernozhukov et al. (2022a), from now on CEINR, with $g(w, \pi(z;\gamma), \theta) = m(w, \pi(z;\gamma)) - \theta$ and $\phi(w, \pi(z;\gamma), \alpha(z), \theta) = \alpha(z) \cdot (y - \pi(z;\gamma))$. Using Taylor series expansion, Assumption 6 and $||\hat{\pi}(z;\gamma) - \pi_0(z;\gamma)|| \xrightarrow{p} 0$ we have,

$$\begin{aligned}
&\int ||g(w, \hat{\pi}(z_i;\hat{\gamma}), \theta_0) - g(w, \pi_0(z_i;\gamma_0), \theta_0)||^2 \mathcal{P}_0(dw) \\
&= \int ||m(w, \hat{\pi}(z_i;\hat{\gamma})) - m(w, \pi_0(z_i;\gamma_0))||^2 \mathcal{P}_0(dw) \\
&\le C \int ||\hat{\pi}(z_i;\hat{\gamma}) - \pi_0(z_i;\gamma_0)||^2 \mathcal{P}_0(dw) \\
&\le C \int ||\hat{\pi}(z_i;\hat{\gamma}) - \hat{\pi}(z_i;\gamma_0) + \hat{\pi}(z_i;\gamma_0) - \pi_0(z_i;\gamma_0)||^2 \mathcal{P}_0(dw)
\end{aligned}$$

By the triangle inequality

$$\begin{aligned}
&\le C \int ||\hat{\pi}(z_i;\hat{\gamma}) - \hat{\pi}(z_i;\gamma_0)||^2 \mathcal{P}_0(dw) \\
&\quad + C \int ||\hat{\pi}(z_i;\gamma_0) - \pi_0(z_i;\gamma_0)||^2 \mathcal{P}_0(dw) \\
&\quad + C \int ||\hat{\pi}(z_i;\hat{\gamma}) - \hat{\pi}(z_i;\gamma_0)|| ||\hat{\pi}(z_i;\gamma_0) - \pi_0(z_i;\gamma_0)|| \mathcal{P}_0(dw) \xrightarrow{p} 0
\end{aligned} \tag{A-4}$$

The first term converges in probability to 0 by Taylor series expansion.

Also by Assumption 5 i) and ii), and as just showed $||\hat{\pi}(z;\hat{\gamma}) - \pi_0(z;\gamma_0)|| \xrightarrow{p} 0$,

$$\begin{aligned}
\int ||\phi(w, \hat{\pi}(z;\hat{\gamma}), \alpha_0, \theta_0) - \phi(w, \pi_0, \alpha_0, \theta_0)||^2 \mathcal{P}_0(dw) &= \int ||\alpha_0(z)(\pi_0(z;\gamma_0) - \hat{\pi}(z;\hat{\gamma}))||^2 \mathcal{P}_0(dw) \\
&\le C \int ||(\pi_0(z;\gamma_0) - \hat{\pi}(z;\hat{\gamma}))||^2 \mathcal{P}_0(dw) \\
&\le C||\hat{\pi}(z;\hat{\gamma}) - \pi_0(z;\gamma_0)||^2 \xrightarrow{p} 0
\end{aligned} \tag{A-5}$$

Also by Assumption 5 i) and $||\hat{\alpha} - \alpha_0|| \xrightarrow{p} 0$, we have,

$$\begin{aligned}
\int ||\phi(w, \pi_0(z;\gamma_0), \hat{\alpha}, \tilde{\theta}) - \phi(w, \pi_0(z;\gamma_0), \alpha_0, \theta_0)||^2 \mathcal{P}_0(dw) &= \int || \left(\hat{\alpha}(z) - \alpha_0(z)\right) \left(y - \pi_0(z;\gamma_0)\right) ||^2 \mathcal{P}_0(dw) \\
&\le C \int ||\hat{\alpha} - \alpha_0||^2 \mathcal{P}_0(dw) \\
&\le C||\hat{\alpha} - \alpha_0||^2 \xrightarrow{p} 0
\end{aligned} \tag{A-6}$$

This satisfies Assumption 1 parts i), ii), and iii) of CEINR.

Next, consider

$$\begin{aligned}\hat{\Delta}(w) &:= \phi\left(w, \hat{\pi}(z;\hat{\gamma}), \hat{\alpha}, \tilde{\theta}\right) - \phi\left(w, f_0, \hat{\alpha}, \tilde{\theta}\right) - \phi\left(w, \hat{\pi}(z;\hat{\gamma}), \alpha_0, \theta_0\right) + \phi\left(w, f_0, \alpha_0, \theta_0\right) \\ &= -\left[\hat{\alpha}(z) - \alpha_0(z)\right]\left[\hat{\pi}(z;\hat{\gamma}) - \pi_0(z;\gamma)\right].\end{aligned}$$

Then by the Cauchy-Schwartz inequality, and Assumptions 6 i) and ii)

$$\begin{aligned}E\left[\hat{\Delta}(w)\right] &= \int -\left[\hat{\alpha}(z) - \alpha_0(z)\right]\left[(\hat{\pi}(z;\hat{\gamma}) - \pi(z;\gamma))\right]\mathcal{P}_0(dz) \\ &\leq \|\hat{\alpha} - \alpha_0\| \, \|(\hat{\pi}(z;\hat{\gamma}) - \pi(z;\gamma))\| = o_p\left(\frac{1}{\sqrt{n}}\right)\end{aligned} \tag{A-7}$$

Also since $\hat{\alpha}(z)$ and $\alpha(z)$ is bounded, we have

$$\begin{aligned}\int \left\|\hat{\Delta}(w)\right\|^2 \mathcal{P}_0(dw) &= \int \left[\hat{\alpha}(z) - \alpha_0(z)\right]^2 \left[(\hat{\pi}(z;\hat{\gamma}) - \pi_0(z))\right]^2 \mathcal{P}_0(dz) \\ &\leq C \int \left[(\hat{\pi} - \pi_0(z))\right]^2 \mathcal{P}_0(dz) \xrightarrow{p} 0\end{aligned} \tag{A-8}$$

Thus Equation A-7 and Equation A-8 verify Assumption 2 i) of CEINR.

Next Assumption 3 of CEINR is satisfied through Assumption 7. Thus Assumptions 1-3 of CEINR are satisfied. Thus asymptotic normality follows by Lemma 15 of CEINR and the Lindberg-Levy central limit theorem.

Finally, we know $\theta \xrightarrow{p} \theta_0$. And thus we have, $\int \left\|g\left(w, \hat{\pi}(z;\hat{\gamma}), \tilde{\theta}\right) - g\left(w, \hat{\pi}(z;\hat{\gamma}), \theta_0\right)\right\|^2 \mathcal{P}_0(dw) = \xrightarrow{p} 0$

To get the second conclusion, we need to show that $\hat{V}$ is a consistent estimator of V. To show this, we closely follow Chernozhukov et al. (2021). Let $\psi_i = \psi_0(w_i)$ and consider

$$\hat{V} = \frac{1}{n}\sum_{i=1}^{n}\hat{\psi}_i^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\psi}_i - \psi_i\right)^2 + \frac{2}{n}\sum_{i=1}^{n}\left(\hat{\psi}_i - \psi_i\right)\psi_i + \frac{1}{n}\sum_{i=1}^{n}\psi_i^2$$

hence, by re-arranging the terms and Cauchy-Schwarz inequality,

$$\hat{V} - V = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\psi}_i - \psi_i\right)^2 + \frac{2}{n}\sum_{i=1}^{n}\left(\hat{\psi}_i - \psi_i\right)\psi_i \leq \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\psi}_i - \psi_i\right)^2 + 2\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\psi}_i - \psi_i\right)^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\psi_i^2}.$$

Using the triangle inequality, for $i \in I_\ell$,

$$\left(\hat{\psi}_i - \psi_i\right)^2 \leq C\sum_{j=1}^{4} R_{ij} = C\sum_{j=1}^{3} R_{ij} + o_p(1)$$

where

$$\begin{aligned}R_{i1} &= \left[m\left(w_i, \hat{\pi}_\ell(z_i;\hat{\gamma}_\ell)\right) - m\left(w_i, \pi_0(z_i;\gamma_0)\right)\right]^2, \\ R_{i2} &= \hat{\alpha}_\ell^2\left(z_i\right)\left[\hat{\pi}_\ell(z_i;\hat{\gamma}_\ell) - \pi_0(z_i;\gamma_0)\right]^2, \\ R_{i3} &= \left[\hat{\alpha}_\ell\left(z_i\right) - \alpha_0\left(z_i\right)\right]^2\left[y_i - \pi_0\left(z_i;\gamma_0\right)\right]^2, \\ R_{i4} &= \left(\hat{\theta} - \theta_0\right)^2.\end{aligned}$$

We already showed consistency, so $R_{i4} \xrightarrow{p} 0$.

Let $I_{-\ell}$ denote observations not in I_{ℓ}. By Markov's inequality, for some $\delta > 0$,

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\psi}_i - \psi_i\right)^2 > \delta\right) \leq \frac{\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\psi}_i - \psi_i\right)^2\right]}{\delta}$$

Note that the cross-fitting allows us to write

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\psi}_i - \psi_i\right)^2\right] \leq \mathbb{E}\left[\frac{C}{n}\sum_{\ell=1}^{L}\sum_{i\in I_\ell}\sum_{j=1}^{3} R_{ij}\right] + o_p(1) = C\sum_{\ell=1}^{L}\frac{n_\ell}{n}\sum_{j=1}^{3}\mathbb{E}\left[\mathbb{E}\left[R_{ij} \mid I_{-\ell}\right]\right] + o_p(1).$$

We already showed,

$$\mathrm{E}\left[R_{i1} \mid I_{-\ell}\right] = \int \left[m\left(w_i, \hat{\gamma}_\ell\right) - m\left(w_i, \gamma_0\right)\right]^2 F_0(dw) = o_p(1)$$

Next by triangle inequality, we have

$$\begin{aligned}
\mathrm{E}\left[R_{i2} \mid \mathcal{W}_{-l}\right] &= \int \hat{\alpha}_l^2 \left[\hat{\pi}_l(z_i;\hat{\gamma}_l) - \pi_0(z_i;\gamma_0)\right]^2 F_0(dz) \\
&= \int \left[\hat{\alpha}_l + \alpha_0 - \alpha_0\right]^2 \left[\hat{\pi}_l(z_i;\hat{\gamma}_l) - \pi_0(z_i;\gamma_0)\right]^2 F_0(dz) \\
&\leq \int \left[\hat{\alpha}_l - \alpha_0\right]^2 \left[\hat{\pi}_l(z_i;\hat{\gamma}_l) - \pi_0(z_i;\gamma_0)\right]^2 F_0(dz) \\
&\quad + \int \left[\alpha_0\right]^2 \left[\hat{\pi}_l(z_i;\hat{\gamma}_l) - \pi_0(z_i;\gamma_0)\right]^2 F_0(dz) \\
&\leq O_p(1)\int \left[\hat{\pi}_\ell(z_i;\hat{\gamma}_\ell) - \pi_0(z_i;\gamma_0)\right]^2 F_0(dz) \xrightarrow{p} 0
\end{aligned}$$

Finally, we have

$$\begin{aligned}
\mathbb{E}\left[R_{i3} \mid I_{-\ell}\right] &= \mathbb{E}\left[\mathbb{E}\left[\left[\hat{\alpha}_\ell\left(z_i\right) - \alpha_0\left(z_i\right)\right]^2 \left[y_i - \pi_0\left(z_i;\gamma_0\right)\right]^2 \mid z_i, I_{-\ell}\right] \mid I_{-\ell}\right] \\
&= \mathbb{E}\left[\left[\hat{\alpha}_\ell\left(Z_i\right) - \alpha_0\left(Z_i\right)\right]^2 \mathbb{E}\left[\left[y_i - \pi_0\left(z_i;\gamma_0\right)\right]^2 \mid Z_i\right] \mid I_{-\ell}\right] \\
&\leq C\left\|\hat{\alpha}_\ell - \alpha_0\right\|^2 \xrightarrow{p} 0.
\end{aligned}$$

As a result,

$$\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\psi}_i - \psi_i\right)^2 \xrightarrow{p} 0$$

Thus, we have $\hat{V} \xrightarrow{p} V$

Also by Assumption 9 and iterated expectations

$$\begin{aligned}
\mathrm{E}\left[R_{i3} \mid \mathcal{W}_{-\ell}\right] &\leq \int \left\{\hat{\alpha}_\ell(z) - \bar{\alpha}(x)\right\}^2 \mathrm{E}\left[(y - \bar{\pi}(z))^2 \mid Z = z\right] F_Z(dz) \\
&\leq C\int \left\{\hat{\alpha}_\ell(z) - \bar{\alpha}(z)\right\}^2 F_z(dz) = C\left\|\hat{\alpha}_\ell - \bar{\alpha}\right\|^2 = o_p(1).
\end{aligned}$$

□

C Additional Numerical Experiments

We first examine the predictive performance of our model under stylized choice models, such as Multinomial Logit (MNL) and Random-Coefficients Logit (RCL). We also investigate the sensitivity of our model by adjusting key parameters, such as the number of products and markets, within the data generation processes. Second, since the stylized MNL and RCL cannot fully capture how each predictor affects the market shares, for example, the relationship between some predictor and the market share could be non-linear. We expand to other models where models are allowed to be misspecified. Third, we demonstrate the performance of our model when broader consumer behaviors, such as consumer inattention, are considered. Fourth, we demonstrate the capability of our model in estimating counterfactuals.

To assess the predictive performance of our model, we focus on three estimators: market share ($\hat{\pi_{jt}}$), own-elasticity ($\frac{\partial \hat{\pi}_{jt}}{\partial P_{jt}}$) and cross-elasticity ($\frac{\partial \hat{\pi}_{jt}}{\partial P_{k\neq jt}}$). For comparison, we also include the predictive performance of four baseline models : 1) MNL; 2) RCL; 3) A standard neural network-based non-parametric method (NP): We tune the hyperparameters – number of layers, number of nodes in each layer, learning rate, and the number of epochs using 5-fold cross-validation for each data generation. We detail the space of hyperparameters in Appendix F. We also apply the ReLU activation for each layer. 4) A “mean" predictor for all data points (Mean) where we predict the market share to be the same for all products in a market and set it equal to the average market share across all products in the dataset.

One advantage of our model compared to the standard neural network-based non-parametric method is that the parameter of our model does not scale with the number of products. The standard neural network uses the stacked products' feature as input and has $(J \times K + 1) \times h_1$ parameters in the input layer, where h_1 denotes the size of the first hidden layer, while our model only use product features with dimension $K + 1$ as input. Across the simulation runs, we observe that the selected neural network has more parameters than our model.

C.1 Baseline Models - RCL and MNL

Data Generation

We consider the stylized discrete choice models, MNL and RCL models as our baseline data generations. We simulate data using these two models, considering a setting with 10 products ($J = 10$), across 100 markets ($M = 100$), with 10 non-price features ($K = 10$). We calculate the utility $u_{i,m,j}$ that consumer i in market m derives from product j using the formula:

$$u_{i,m,j} = \alpha_i Price_{m,j} + \beta_i X_{m,j} + \varepsilon_{i,m,j}, \tag{A-9}$$

where $\varepsilon_{i,m,j}$ represents an independently and identically distributed (iid) Type-I extreme value across products and consumers. $X_{m,j} \in \mathbb{R}^K$ denotes the non-price features of the product. α_i, β_i are the model coefficients, which are kept constant for all consumers in the MNL, while in the RCL, they are normally distributed across consumers. The probability distribution of features and coefficients are in Appendix E.

We denote the market share of product j in market m generated from MNL by $S^{MNL}_{m,j}$ and the market share generated from RCL by $S^{RCL}_{m,j}$. For each market, we generate the market shares of each product by simulating $N = 10,000$ individual choices and aggregating by each market as below, (the mean utility derived from the outside option is normalized to 0.)

$$S^{MNL}_{m,j} = \frac{1}{N} \sum_{i}^{N} 1(u_{i,m,j} = \max_k(u_{i,m,k}))^5 \tag{A-10}$$

$$S^{RCL}_{m,j} = \frac{1}{N} \sum_{i}^{N} \frac{exp(\alpha_i Price_{m,j} + \beta_i X_{m,j})}{1 + \sum_j exp(\alpha_i Price_{m,j} + \beta_i X_{m,j})} \tag{A-11}$$

[5] Instead of simulating each individual's choice probability, we simulate each indivdual's choice based on the utility maximization. This approach ensures that when we use MNL (true model) for estimation, it does not reproduce the data perfectly.

We then split the generated data into training data (80%) and test data (20%) and only use training data for estimation. We report Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) of the estimators using 20 draws of each simulation.

Results

Table A2, A3, and A4 present MAE and RMSE in the predicted market share ($\hat{\pi_{jt}}$), own-elasticity ($\frac{\widehat{\partial \pi_{jt}}}{\partial P_{jt}}$) and cross-elasticity ($\frac{\widehat{\partial \pi_{jt}}}{\partial P_{k \neq jt}}$) respectively. J represents the number of products, M represents the number of markets (in the full data) and K represents the number of non-price features. The number of observations for market share ($\hat{\pi_{jt}}$) is calculated based on $M \times J \times 20\%$ (the portion of test data) $\times 20$ (the number of draws of simulations). The number of observations for own-elasticity ($\frac{\widehat{\partial \pi_{jt}}}{\partial P_{jt}}$) is calculated based on $M \times J \times 80\%$ (the portion of training data [6]) $\times 20$ (the number of draws of simulations). The number of observations for cross-elasticity ($\frac{\widehat{\partial \pi_{jt}}}{\partial P_{k \neq jt}}$) is calculated based on $M \times J \times (J-1) \times 80\%$ (the portion of training data) $\times 20$ (the number of draws of simulations).

Our model outperforms the benchmark non-parametric method in the predicted market shares in all data generation processes. Note that, this is despite extensive hyperparameter tuning. When the true model is MNL, our model cannot beat RCL or MNL, which is as expected; but the error of our model is close to the true model. When the true model is RCL, our model can beat MNL consistently and the performance of our model is also close to the true model.

A significant advantage of our model, as compared to the benchmark non-parametric estimator, is its ability to circumvent the curse of dimensionality that arises with the increase in the number of products. Specifically, the number of model parameters of our model does not scale with the number of products. On the other hand, each sample in the NP method is a market while one sample in our model is a product. Therefore, the sample size of NP method is essentially M, while the sample size of our model is $M \times J$. Thus, as the number of products escalates, we anticipate an improvement in the performance of our model due to the availability of more training data points or observations. We verify this by varying the number of products to 5, 10, and 20 during data generation. As anticipated, the MAE of the predicted market shares from our model decreases monotonically with an increase in product count. In contrast, the benchmark non-parametric estimator even falls behind that of Mean prediction as the number of products increases, demonstrating the existence of the curse of dimensionality.

We also test the performance of our model with varying market numbers (20, 100, and 200). Although both our model and the NP method show improved performance with more markets, the non-parametric estimator is more adversely affected by a decrease in market numbers due to a more significant reduction in its sample size. This is particularly problematic in scenarios with one market, as the non-parametric estimator becomes infeasible for estimation for only one sample.

In the prediction of own- and cross- elasticity, the patterns observed in the predicted market shares generally remain consistent. Note that, here we use the predicted own-elasticity ($\frac{\widehat{\partial \pi_{jt}}}{\partial P_{jt}}$) and cross-elasticity ($\frac{\widehat{\partial \pi_{jt}}}{\partial P_{k \neq jt}}$) for each observation (product), which is different to the average effect of price on the market share that we discussed in Section A.1. We will elaborate how to conduct inference on the average effect of price ($\hat{\theta}$) in Section A.2.

C.2 Choice Behaviors

Recent literature (e.g., Abaluck et al. (2020); Gabaix (2019); Honka et al. (2019); Compiani (2022); Goeree (2008)) also demonstrate that consumers may not be fully informed of all products when deciding which product to purchase. This violates a general assumption of the choice model: consumers are informed and consider all options when they make purchase decisions. In some parametric models (e.g., Van Nierop et al. (2010)), this issue is managed by constructing the consumer consideration set. However, consideration sets are mostly unobserved in data thus it requires assumptions on how consideration sets are formed, which might not always be appropriate or reflective of actual consumer behavior. Another way to manage this issue is to consider for search

[6] The reason that we use training data for evaluating elasticity is to mirror the process when our method is applied in estimating elasticity. That is when all data is used for estimation.

Table A2: Baseline Results - Predicted Market Share

#	True Model	J	M	K	Our Model		MNL		RCL		NP		Mean		No. Obs.
					MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	
0	MNL	5	100	10	0.0534	0.0834	0.0078	0.0105	0.0082	0.0052	0.1269	0.2364	0.2312	0.2220	2000
1	MNL	10	20	10	0.0585	0.1086	0.0040	0.0131	0.0089	0.0134	0.1129	0.2191	0.1365	0.2948	800
2	MNL	10	100	10	0.0333	0.0591	0.0044	0.0039	0.0026	0.0053	0.1181	0.1717	0.1422	0.1503	4000
3	MNL	10	200	10	0.0307	0.1346	0.0032	0.0102	0.0034	0.0197	0.1096	0.2170	0.1416	0.2106	8000
4	MNL	20	100	10	0.0194	0.0765	0.0015	0.0077	0.0023	0.0068	0.0707	0.2242	0.0768	0.2201	8000
5	RCL	5	100	10	0.0240	0.0314	0.0307	0.0382	0.0033	0.0042	0.0456	0.0583	0.0538	0.0656	2000
6	RCL	10	20	10	0.0206	0.0281	0.0270	0.0343	0.0034	0.0044	0.0540	0.0612	0.0418	0.0525	800
7	RCL	10	100	10	0.0171	0.0231	0.0262	0.0326	0.0025	0.0033	0.0458	0.0583	0.0413	0.0514	4000
8	RCL	10	200	10	0.0141	0.0187	0.0252	0.0318	0.0032	0.0039	0.0431	0.0559	0.0412	0.0513	8000
9	RCL	20	100	10	0.0099	0.0140	0.0262	0.0281	0.0018	0.0024	0.0390	0.0489	0.0276	0.0354	8000

Note: This table presents the baseline results for predicted market share using various models. J, M, and K represent the number of products, non-price features, and markets, respectively. NP denotes a benchmark non-parametric method. Specifically, we use a standard neural network, where we tune the hyperparameters (number of layers, number of nodes in each layer, learning rate, and the number of epochs) using 5-fold cross-validation for each data generation. Mean indicates a prediction method that predicts the market share to be equal to the average market share derived from the training data. The Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) provided for each scenario (i.e., true model, J, K, M) are computed using the test data from 20 iterations of data generation. The column titled "No. Obs." indicates the total number of products in the test data across all draws. Specifically, the number of observations for market share ($\hat{\pi_{jt}}$) is calculated based on $M \times J \times 20\%$ (the portion of test data) ×20 (the number of draws of simulations).

Table A3: Baseline Results - Estimated Own-Elasticity ($\frac{\partial \hat{\pi_{jt}}}{\partial P_{jt}}$)

#	True Model	J	M	K	Our Model		MNL		RCL		NP		No. Obs
					MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	
0	MNL	5	100	10	0.2588	1.1815	0.1414	1.1232	0.1757	1.1322	0.4554	1.2115	8000
1	MNL	10	20	10	0.3523	1.3557	0.1967	1.2970	0.2585	1.3118	1.0057	1.5316	3200
2	MNL	10	100	10	0.3346	1.4327	0.2066	1.3851	0.2150	1.3863	0.9266	1.5857	16000
3	MNL	10	200	10	0.3245	1.4131	0.2007	1.3842	0.2357	1.3876	0.8266	1.5768	32000
4	MNL	20	100	10	0.4146	1.6596	0.3305	1.6203	0.3570	1.6229	1.0151	1.8353	32000
5	RCL	5	100	10	0.1189	0.2310	0.1474	0.3735	0.0125	0.0305	0.1802	0.3145	8000
6	RCL	10	20	10	0.1799	0.3729	0.2039	0.5326	0.0365	0.1196	0.3768	0.3254	3200
7	RCL	10	100	10	0.1498	0.2862	0.2154	0.5531	0.0224	0.0685	0.2987	0.3643	16000
8	RCL	10	200	10	0.1209	0.2416	0.2188	0.5512	0.0233	0.0732	0.2464	0.4241	32000
9	RCL	20	100	10	0.1658	0.3533	1.4591	1.7099	0.0429	0.1319	0.4555	0.4741	32000

Note: This table presents the baseline results for estimated own-elasticity using various models. The number of observations for own-elasticity ($\frac{\partial \hat{\pi_{jt}}}{\partial P_{jt}}$) is calculated based on $M \times J \times 80\%$ (the portion of training data ×20 (the number of draws of simulations).)

Table A4: Baseline Results - Estimated Cross-Elasticity ($\frac{\partial \hat{\pi_{jt}}}{\partial P_{k \neq jt}}$)

#	True Model	J	M	K	Our Model		MNL		RCL		NP		No. Obs
					MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	
0	MNL	5	100	10	0.1349	0.8115	0.0107	0.6780	0.0118	0.6807	0.1968	0.9442	32000
1	MNL	10	20	10	0.0649	0.6742	0.0043	0.5326	0.0059	0.5424	0.1527	0.7123	28800
2	MNL	10	100	10	0.0862	0.6104	0.0043	0.5142	0.0049	0.5143	0.1885	0.7488	144000
3	MNL	10	200	10	0.0901	0.6075	0.0042	0.5140	0.0047	0.5153	0.2110	0.7831	288000
4	MNL	20	100	10	0.0482	0.4665	0.0015	0.4151	0.0016	0.4157	0.1270	0.5303	608000
5	RCL	5	100	10	0.0293	0.0571	0.0492	0.0551	0.0030	0.0070	0.0617	0.0960	32000
6	RCL	10	20	10	0.0261	0.0435	0.0332	0.0455	0.0039	0.0143	0.0972	0.0791	28800
7	RCL	10	100	10	0.0257	0.0493	0.0324	0.0447	0.0028	0.0090	0.0795	0.1124	144000
8	RCL	10	200	10	0.0213	0.0417	0.0353	0.0455	0.0035	0.0097	0.0745	0.1277	288000
9	RCL	20	100	10	0.0220	0.0431	0.0204	0.0359	0.0022	0.0103	0.0715	0.0890	608000

Note: This table presents the baseline results for estimated cross-elasticity using various models. The number of observations for cross-elasticity ($\frac{\partial \hat{\pi_{jt}}}{\partial P_{k \neq jt}}$) is calculated based on $M \times J \times (J-1) \times 80\%$ (the portion of training data) ×20 (the number of draws of simulations).

cost (Mehta et al., 2003). Similarly, it also requires researchers to specify how search cost enters the utility function and decision process. In contrast to these models, our approach refrains from making any parametric assumptions, which allows for a potentially more flexible representation of consumer behavior in the case of consumer attention. To demonstrate how our model can capture the inattentive behavior, we look at a scenario where consumers are inattentive and deviate from the traditional random coefficients logit model.

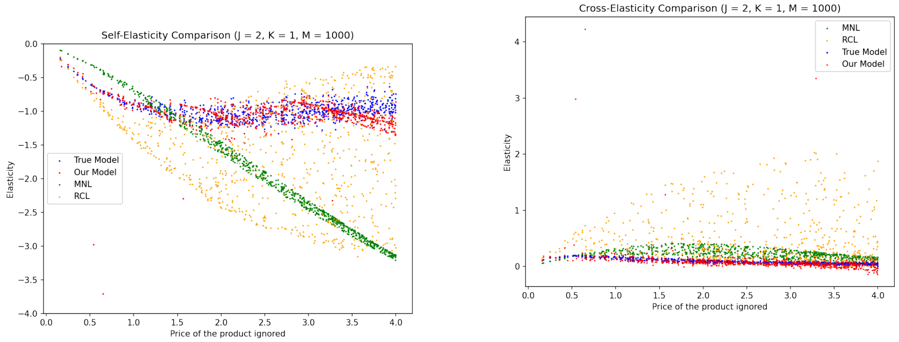
To evaluate how our model performs when there are inattentive consumers, we simulate the market share of each product in each market by assuming there is a portion of consumers who ignore the product with the highest price. We assume the portion is $1 - \frac{1}{1+Price_j}$, so when the price increases, the portion of inattentive consumers increases. In other words, the consideration set of $1 - \frac{1}{1+Price_j}$ consumers excludes the highest-price product. We again only consider there is only one feature, price. The choice probability of the highest price product j_h is

$$\frac{1}{1+Price_{j_h}} \frac{exp(\alpha_i Price_{m,j_h})}{1+\sum_k exp(\alpha_i Price_{m,k})}.$$

The choice probability of other products $j \neq j_h$ is

$$\frac{1}{1+Price_{j_h}} \frac{exp(\alpha_i Price_{j,k})}{1+\sum_k exp(\alpha_i Price_{m,k})} + (1 - \frac{1}{1+Price_{j_h}}) \frac{exp(\alpha_i Price_{m,j})}{1+\sum_{k \neq j_h} exp(\alpha_i Price_{m,k})}.$$

Figure A2 illustrates the performance of different models when there are two products. We consider the number of markets to be 1,000 so that we can observe enough variance in our data. We still only consider there is only one feature, price. Figure A2a captures how the estimated own-elasticity varies with the price of the product ignored. Due to the existence of inattention, when the price is higher, the portion of inattentive consumers is higher. Thus when we change the price, the change in market share is smaller than the case without inattention. Only our model captures this pattern and predicts the own-elasticity to be flat when the price is high. It is also easy to see that only our model is close to the true model. Figure A2b captures how the estimated cross-elasticity varies with the price of the other product. Similarly, due to the ignorance of inattention, both MNL and RCL overestimate the magnitude of the elasticity of the other product. Our model is the only model that captures the elasticity and is closest to the true model.



(a) Own-Elasticity ($\frac{\partial \hat{\pi}_{jt}}{\partial P_{jt}}$) in Consumer Inattention

(b) Cross-Elasticity ($\frac{\partial \hat{\pi}_{jt}}{\partial P_{k \neq jt}}$) in Consumer Inattention

Figure A2: Elasticity Effects in Consumer Inattention

Note: Figure A2 illustrates how different models perform when there is $1 - \frac{1}{1+Price_j}$ consumers who ignore the product with the highest price. We simulate market shares in the case of 2 products, 1000 markets, and 1 feature (price). Due to the existence of inattention, when the price is higher, the portion of inattentive consumers is higher. Thus when we change the price, the change in market share is smaller than the case without inattention.

We consider when there are more products (5, 10) and present our results in Table A5, A6, and A7. Consistently, our model outperforms all other models across these scenarios.

C.3 Counterfactual Analysis

Next, we look at the application of our model to estimate counterfactuals. Our model can handle the counterfactual estimates in multiple scenarios. First, our model can handle any counterfactuals when shocks only result in the change in features of each product. For example, in a choice model where ranking is an important component that influences the choice behavior of each individual, researchers would like to see how the demand would change if the ranking policy is changed. This counterfactual

Table A5: Consumer Inattention - Predicted Market Shares

#	J	Our Model		MNL		RCL		NP		Mean		No. Obs
		MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	
0	2	0.0047	0.0198	0.0590	0.0753	0.0316	0.0439	0.0076	0.0273	0.1840	0.2044	40
1	5	0.0064	0.0139	0.0203	0.0252	0.0178	0.0226	0.0250	0.0345	0.0780	0.1022	100
2	10	0.0033	0.0068	0.0137	0.0166	0.0072	0.0100	0.0258	0.0365	0.0492	0.0656	200

Note: This table presents the MAE and RMSE of predicted market shares when there are inattentive consumers. We simulate the market share of each product in each market by assuming there is a portion of consumers who ignore the product with the highest price. We assume the portion is $1 - \frac{1}{1+Price_j}$. We consider 3 scenarios with 2, 5, and 10 products respectively. We fix the number of markets to 100 and the number of features to 1 (with only price). Other parts are the same as RCL in our baseline.

Table A6: Consumer Inattention - Estimated own-Elasticity

#	J	Our Model		MNL		RCL		NP		No. Obs
		MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	
0	2	0.0609	0.5190	0.6758	1.5698	0.3929	1.4598	0.0573	0.8804	3200
1	5	0.0978	1.8579	0.6917	1.9614	0.3753	1.9782	0.4288	1.9546	8000
2	10	0.0708	2.4273	0.8306	2.1031	0.1842	2.2787	0.5464	2.1450	16000

Note: This table presents the MAE and RMSE of own-elasticity when there are inattentive consumers. We simulate the market share of each product in each market by assuming there is a portion of consumers who ignore the product with the highest price. We assume the portion is $1 - \frac{1}{1+Price_j}$. We consider 3 scenarios with 2, 5, and 10 products respectively. We fix the number of markets to 100 and the number of features to 1 (with only price). Other parts are the same as RCL in our baseline.

Table A7: Consumer Inattention - Estimated Cross-Elasticity

#	J	Our Model		MNL		RCL		NP		No. Obs
		MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	
0	2	0.0419	4.3553	0.1911	8.8197	0.3350	8.5791	0.0378	5.6814	3200
1	5	0.0486	4.7745	0.0827	5.8357	0.0503	5.8404	0.1897	5.5601	32000
2	10	0.0212	3.8765	0.0476	4.0961	0.0146	4.1113	0.1676	4.1198	144000

Note: This table presents the MAE and RMSE of cross-elasticity when there are inattentive consumers. We simulate the market share of each product in each market by assuming there is a portion of consumers who ignore the product with the highest price. We assume the portion is $1 - \frac{1}{1+Price_j}$. We consider 3 scenarios with 2, 5, and 10 products respectively. We fix the number of markets to 100 and the number of features to 1 (with only price). Other parts are the same as RCL in our baseline.

can be estimated by updating the value of the ranking feature in prediction. Second, our model can also handle the counterfactual estimates when the choice set (product set) changes. For example, one common counterfactual of interest is the demand of new product. Since our model only uses product features as input, we can estimate the demand of any new product. In contrast, it is important to acknowledge that estimating counterfactual demand with a standard neural network estimator is infeasible due to its structural constraints on the input space. A change in choice set would result in the change of size of the input vector, making such estimation infeasible. The same case stands for when researchers would like to estimate the demand when one product is removed from the market.

To showcase the capability of our model to estimate counterfactuals, we consider a case where a new product is introduced to the market. For comparison, we will only consider an MNL and RCL estimator since it is infeasible for the NP method to estimate the counterfactual. We use two data generation processes – Multinomial Logit (MNL) and Random-Coefficients Logit (RCL). In each data generation, we consider an 11th product is introduced to each market where there were 10 products and an outside option. The observable characteristics of the new product are simulated from the same distribution as other products. In Table A8, we present the estimated market share of the new product. Our model outperforms the MNL when the underlying data generation process is RCL, and produces results comparable to the true model.

Table A8: New Product Demand Estimation - Predicted Market Share ($\hat{\pi_{jt}}$)

True Model	Our Model		MNL		RCL		No. Obs.
	MAE	RMSE	MAE	RMSE	MAE	RMSE	
MNL	0.0234	0.0644	0.0041	0.0095	0.0023	0.0045	22,000
RCL	0.0186	0.0145	0.0265	0.0331	0.0023	0.0031	22,000

Note: This table presents the MAE and RMSE of predicted market shares of all products in the market when a new product enters. We simulate market shares as in our baseline scenario (10 products and one outside option, 100 markets, 10 features) when a 11th product is introduced.

D Emprical Data Analysis: US Automobile Data (1971 - 1990)

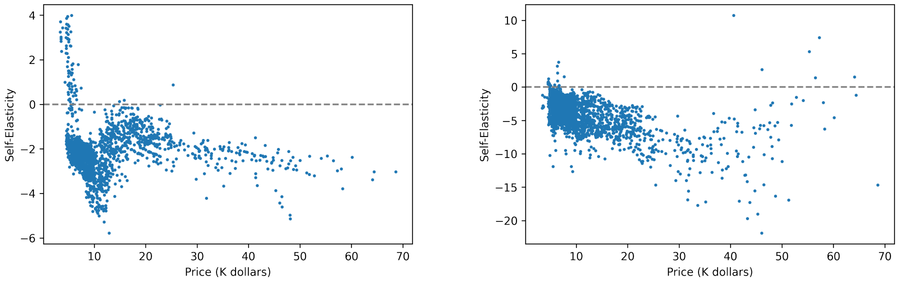
In this section, we apply our model to a real-world dataset. We use the "US Automobile Data (1971 - 1990)" from Berry et al. (1995). The dataset features cars in the US market from 1971 to 1990, with each year regarded as a market. The number of cars varies from 86 to 150 each year. For each car, the dataset provides information such as the car's name, the manufacturing company, factory region, market share, price, and four exogenous car characteristics: horsepower, space, mileage per dollar, and the presence of an air conditioning device.

Even though the dataset is relatively small, it presents two key challenges that make it difficult to use traditional non-parametric estimators: (i) the dataset features markets with more than 100 products and only 20 markets in total, (ii) the product assortment in each year or market varies. In this section, we demonstrate the use of our estimator, which is capable of effectively addressing such challenges posed in real-world datasets.

For each component of our model (ϕ_1, ϕ_2, and ρ), we use a standard 3-layer neural network, and this is implemented without further hyperparameter tuning. We implement ReLU activation function at each layer. This architecture is the same as the one we used in our numerical experiments. For comparison, we replicate the random coefficient logit model (with only the demand side) used by Berry et al. (1995) using the Python package `pyblp` (Conlon and Gortmaker, 2020). In our replication, we allow for heterogeneity in random coefficients across all variables. Our findings show that the estimates obtained from our model are comparable to the random coefficient logit estimation presented in Berry et al. (1995). We estimate our model both without and with consideration of endogeneity. To address endogeneity we utilize three sets of IVs – (i) the sum of characteristics of all car models, excluding the product in focus, produced by the same firm in the same year; (ii) the sum of characteristics of all car models, excluding the product in focus, produced by rival firms in the same year; and (iii) cost shifters, which encompass the wage and exchange rate prevalent in the year and region where the factory is located. The utilization of traditional BLP-style instruments, as discussed by Gandhi and Houde (2019), can be problematic due to their relative weakness, often resulting in considerable bias in the estimation of parameters. These issues are significantly exacerbated in non-parametric models. Thus, to counter potential concerns related to weak instruments, we employ a machine-learning-based IV methodology (MLIV) as proposed by Singh et al. (2020). We detail the estimation procedure and results using BLP style IVs in Appendix G.

In Figure A3, we present the estimated own-elasticity ($\frac{\partial \hat{\pi_{jt}}}{\partial P_{jt}}$) of our model without IV and with IV. The x-axis represents the price of the focal product, while the y-axis shows the product's own elasticity. Each point corresponds to a product in a market, resulting in 2,217 observations. We report the estimated elasticity based on the same price variation used in the BLP paper (a 1,000-dollar change). In Figure A3a, we observe that the majority of low-price products (priced below 6,000 dollars) exhibit positive estimated own-elasticity, demonstrating the existence of the endogeneity.

We report the own-elasticity ($\frac{\partial \hat{\pi_{jt}}}{\partial P_{jt}}$) and cross-elasticity ($\frac{\partial \hat{\pi_{jt}}}{\partial P_{k\neq jt}}$) estimated in our model and random coefficient logit model with a sample of 13 cars in the 1990 market in Table A9 and A10. The sample of 13 cars is the same as the one reported in Berry et al. (1995). Overall, our results are very similar and comparable to Berry et al. (1995). We also plot the distributions of the estimated own-elasticity ($\frac{\partial \hat{\pi_{jt}}}{\partial P_{jt}}$) and cross-elasticity ($\frac{\partial \hat{\pi_{jt}}}{\partial P_{k\neq jt}}$) obtained from our model and the BLP model in Figure A4. The filled areas in the violin plots represent the complete range of the elasticities, while the text labels next to the line indicate the mean values. The estimated mean own- and cross-elasticities appear to

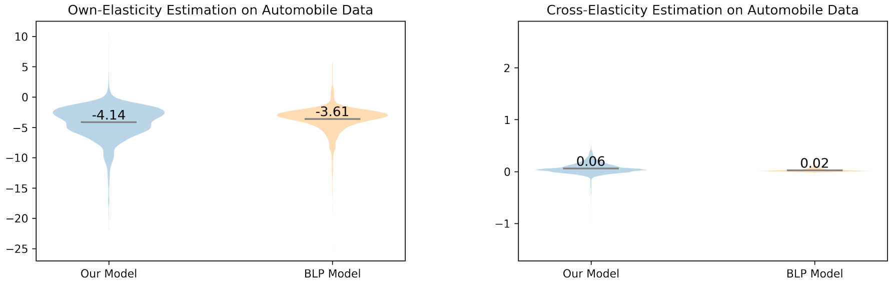


(a) Estimated own-Elasticity ($\frac{\partial \hat{\pi}_{jt}}{\partial P_{jt}}$)without IV (b) Estimated own-Elasticity ($\frac{\partial \hat{\pi}_{jt}}{\partial P_{jt}}$) with IV

Figure A3: Elasticity Estimation Comparison

Figure A3 presents the estimated own-elasticity ($\frac{\partial \hat{\pi}_{jt}}{\partial P_{jt}}$) of our model without IV and with IV. The x-axis represents the price of the focal product, while the y-axis shows the product's own-elasticity. Each point corresponds to a product in a market, resulting in 2,217 observations. We report the estimated elasticity based on the same price variation used in the BLP paper (a 1,000-dollar change). Although we cannot ascertain the true value of own-elasticity, it is widely accepted that own-elasticity should generally be negative for most, if not all, products. In Figure A3a, we observe that the majority of low-price products (priced below 5,000 dollars) exhibit positive estimated own-elasticity, demonstrating the existence of the endogeneity.

be similar between our model and the BLP model, though our model exhibits a larger RMSE in the estimated elasticity values compared to the BLP model.



(a) Own-Elasticity Estimation (Our Model vs. BLP Model) (b) Cross-Elasticity Estimation (Our Model vs. BLP Model)

Figure A4: Elasticity Estimation Comparison

Note: Figure A4 illustrates the distributions of the estimated own- and cross-elasticities obtained from our model and the BLP model. The filled areas in the violin plots represent the complete range of the elasticities, while the text labels indicate the mean values.

We further estimate the average own-elasticity ($\hat{\theta}$) for high-priced, medium-priced, and low-priced cars and construct a confidence interval for each category using our inference procedure. We present our result in Table A11.

E Distribution of Features and Coefficients in Numerical Experiments

	MNL	RCL
$Price_{m,j}$	$U[0,4]$	$U[0,4]$
$X_{m,j}$	$N(0,1)$	$N(0,1)$

Table A12: Distribution of Features

	Acura Legend	BMW 735i	Buick Century	Cadillac Seville	Chevy Cavalier	Ford Escort	Ford Taurus	Honda Accord	Lexus LS400	Lincoln Town Car	Mazda 323	Nissan Maxima	Nissan Sentra
Acura Legend	-5.6060	0.1993	0.2198	0.2221	0.0632	0.2317	0.2199	0.2337	0.2144	0.2187	0.2595	0.2354	0.2595
BMW 735i	0.4095	-6.1528	0.3653	0.4093	0.0352	0.3525	0.3655	0.3807	0.4120	0.4084	0.3547	0.4161	0.3547
Buick Century	0.1234	0.1020	-6.1023	0.1213	0.0376	0.1294	0.1215	0.1205	0.1175	0.1191	0.1400	0.1299	0.1400
Cadillac Seville	0.2895	0.2179	0.2631	-7.2896	0.0647	0.2692	0.2631	0.2566	0.2810	0.2818	0.2892	0.2849	0.2892
Chevy Cavalier	0.0142	-0.0013	0.0202	0.0167	-1.3447	0.0171	0.0202	0.0353	0.0126	0.0167	0.0354	0.0291	0.0355
Ford Escort	0.0410	0.0245	0.0353	0.0413	-0.0148	-1.8519	0.0353	0.0520	0.0392	0.0413	0.0494	0.0513	0.0494
Ford Taurus	0.1166	0.0914	0.1188	0.1160	0.0183	0.1199	-6.1473	0.1258	0.1066	0.1157	0.1451	0.1290	0.1451
Honda Accord	0.0975	0.0647	0.0968	0.1006	-0.0003	0.0945	0.0969	-5.7438	0.0914	0.1006	0.1166	0.1140	0.1165
Lexus LS400	0.3357	0.2606	0.3136	0.3325	0.0822	0.3235	0.3137	0.3126	-6.8791	0.3271	0.3495	0.3348	0.3494
Lincoln Town Car	0.2681	0.2310	0.2548	0.2656	0.0713	0.2663	0.2548	0.2648	0.2586	-5.3996	0.3009	0.2719	0.3009
Mazda 323	0.0361	0.0212	0.0272	0.0326	-0.0127	0.0249	0.0272	0.0363	0.0323	0.0326	-2.6589	0.0404	0.0357
Nissan Maxima	0.1579	0.1367	0.1589	0.1555	0.0425	0.1670	0.1589	0.1689	0.1484	0.1534	0.1884	-7.2216	0.1884
Nissan Sentra	0.0386	0.0239	0.0304	0.0384	-0.0202	0.0294	0.0304	0.0496	0.0375	0.0383	0.0439	0.0470	-1.8754

Table A9: Estimated own- and cross-elasticities of a sample of automobile data using our method

Note: This table presents the estimated own- and cross-elasticity of a sample of 13 cars in the 1990 market using our model. The selected cars are the same as Berry et al. (1995) reports. Each entry with row index i and column index j gives the percentage change in demand divided by the percentage change in price (based on $1,000 change in the price of i).

	Acura Legend	BMW 735i	Buick Century	Cadillac Seville	Chevy Cavalier	Ford Escort	Ford Taurus	Honda Accord	Lexus LS400	Lincoln Town Car	Mazda 323	Nissan Maxima	Nissan Sentra
Acura Legend	-5.4677	0.0489	0.0205	0.1029	0.0221	0.0220	0.0143	0.1477	0.1503	0.0273	0.0013	0.1359	0.0039
BMW 735i	0.1267	-9.8502	0.0156	0.1058	0.0122	0.0121	0.0057	0.1313	0.1546	0.0278	0.0006	0.1375	0.0022
Buick Century	0.0184	0.0054	-5.1978	0.0124	0.1153	0.1043	0.1475	0.1982	0.0165	0.0800	0.0076	0.0379	0.0174
Cadillac Seville	0.1293	0.0513	0.0175	-6.6819	0.0151	0.0150	0.0099	0.1393	0.1576	0.0271	0.0008	0.1406	0.0027
Chevy Cavalier	0.0131	0.0028	0.0766	0.0071	-3.1163	0.1421	0.0849	0.2608	0.0086	0.0404	0.0100	0.0395	0.0241
Ford Escort	0.0137	0.0029	0.0726	0.0074	0.1487	-3.0590	0.0603	0.2781	0.0090	0.0263	0.0106	0.0419	0.0258
Ford Taurus	0.0048	0.0007	0.0554	0.0027	0.0479	0.0326	-4.0258	0.0727	0.0017	0.1779	0.0026	0.0122	0.0057
Honda Accord	0.0387	0.0133	0.0582	0.0291	0.1151	0.1173	0.0568	-4.3399	0.0409	0.0297	0.0081	0.0618	0.0200
Lexus LS400	0.1350	0.0536	0.0166	0.1126	0.0130	0.0130	0.0045	0.1400	-7.4316	0.0243	0.0006	0.1464	0.0024
Lincoln Town Car	0.0087	0.0034	0.0286	0.0069	0.0217	0.0135	0.1693	0.0362	0.0086	-5.6139	0.0011	0.0123	0.0024
Mazda 323	0.0114	0.0020	0.0743	0.0056	0.1476	0.1494	0.0679	0.2723	0.0063	0.0304	-2.8631	0.0390	0.0254
Nissan Maxima	0.1008	0.0393	0.0314	0.0831	0.0493	0.0500	0.0271	0.1749	0.1209	0.0286	0.0033	-4.7872	0.0086
Nissan Sentra	0.0140	0.0031	0.0707	0.0078	0.1471	0.1504	0.0617	0.2763	0.0095	0.0271	0.0105	0.0422	-3.1799

Table A10: Estimated own and cross-elasticities of a sample of automobile data using the BLP model

Note: This table presents the estimated own- and cross-elasticity of a sample of 13 cars in the 1990 market using the BLP model. The selected cars are the same as Berry et al. (1995) reports. Each entry with row index i and column index j gives the percentage change in demand divided by the percentage change in price (based on $1,000 change in the price of i).

	BLP Model Mean Estimate	Our Model Mean Estimate (95% Confidence Interval)	No. Obs.
High	-5.6705	-4.1922 (-6.2204, -2.1641)	20
Medium	-3.7354	-3.7215 (-4.9260, -2.5169)	20
Low	-2.9174	-2.0697 (-3.1980, -0.9415)	20

Table A11: Estimates of Average Own-Elasticity

Note: This table presents the estimated average own-elasticity for cars across various price categories. We categorize cars with a price over $20k as “high-priced", cars priced between $8k and $20k as “medium-priced", and all other cars as “low-priced". For each category, we randomly select one car of the category from each market as one observation. For the BLP model, we calculate the average own-elasticity of the sampled cars as the mean estimate. For our model, we estimate the average own-elasticity and construct the confidence interval following our inference procedure.

	MNL	RCL
α_i	-1	$N(-1, 1)$
β_{ik}	1	$N(\mu_{\beta_k}, 1)$

Notes : $\mu_{\beta_k} \sim N(0, 1/2K)$

Table A13: Distribution of Coefficients

F Hyperparameter Space for Tuning Non-parametric Estimator Benchmark

Hyperparameter	Space
Number of hidden layers	[3, 4, 5]
Number of nodes in each layer	[64, 128, 256]
Learning rate	[1e-2, 1e-3, 1e-4]
Number of epochs	[1, 2, 4]

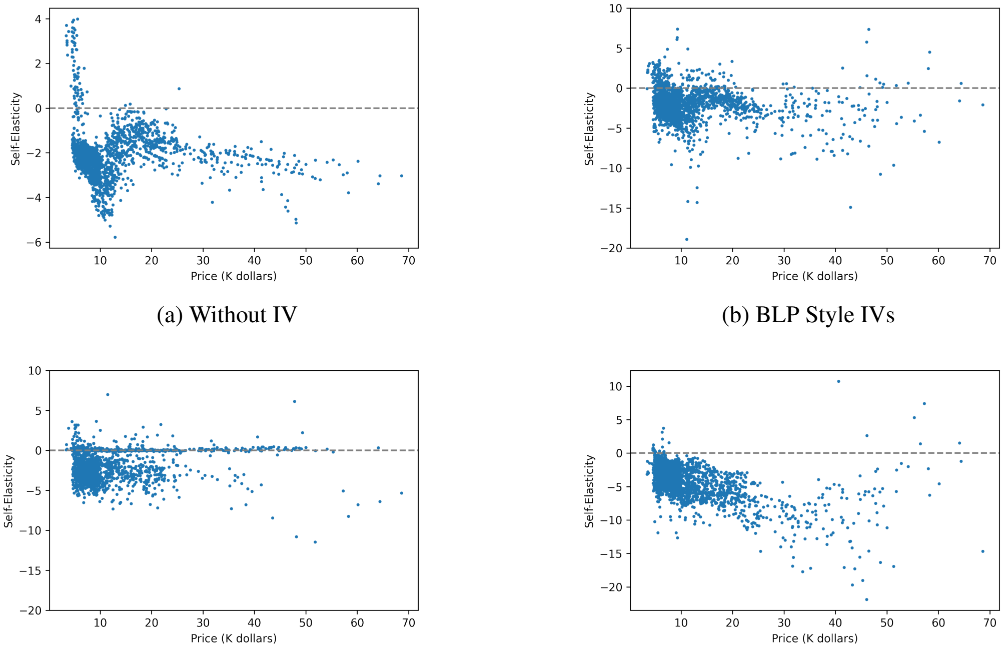
Table A14: Hyperparameter Space for Tuning Non-parametric Estimator Benchmark

G Details in Adopting the “MLIV” Method

Following Singh et al. (2020), we perform the steps below to construct the machine-learning-based IV (MLIV) and use them to estimate $\hat{\gamma}$ to control for endogeneity in prices.

- **Step 1: Data Partition** We randomly split the data set, S, into three separate partitions of markets, each denoted as S_k. Each market is exclusively assigned to only one partition. For each partition, we define its complement set, S_k^c, as the subset of data in S that is not included in S_k.
- **Step 2: Cross-fitting** For each partition S_k, we first estimate a linear regression model on the complement data set, S_k^c, using the Lasso method with hyperparameters tuned by 3-fold cross-validation. As discussed in section A, we need the estimator of γ to converge at $n^{-1/2}$ rate, a similar result that bounds the in-sample prediction error of the lasso estimator has been established in Chatterjee and Jafarov (2015). Then, we use this trained model to predict the outcomes (prices) of the S_k. We denote the fitted value as $\hat{f}_k$, which is essentially the MLIV.
- **Step 3: First-stage Regression** We run a first-stage linear regression on the entire dataset using the MLIV as the only predictor. Then, we use the residuals estimated from this first-stage regression as one additional feature as detailed in Section 2.2.

As a supplement to our main result, we also run our model using non-machine learning-based IVs. Similar to Figure 4 in the main text, we present the estimated own-elasticity of our model without IV, with BLP-style IVs, with differentiation IVs, and with MLIV in Figure A5. In Figure A5b, even when IVs are applied, the persistence of many positive own-elasticities suggests the weakness of the BLP style IVs. Furthermore, we apply the differentiation IVs (Gandhi and Houde, 2019), which use exogenous measures of differentiation and provide a more robust instrument compared to the conventional BLP IVs. As one can see from Figure A5c, the use of differentiation IV provides a more realistic estimation of own-elasticities, strengthening the issue of weak instruments of the BLP style IVs. We also include the distributions of the estimated own- and cross-elasticities obtained from our model using different sets of IVs in Figure A6.



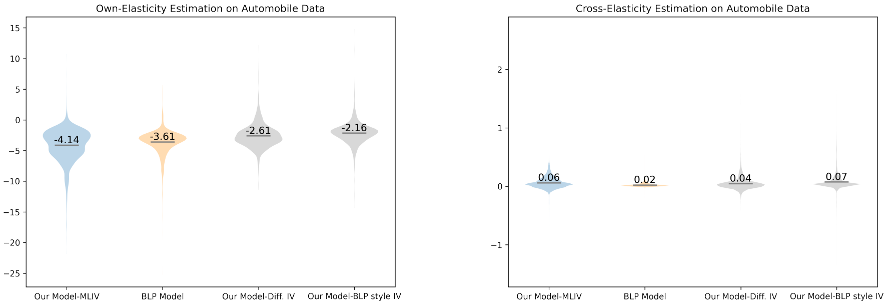
(a) Without IV

(b) BLP Style IVs

(c) Differentiation IVs

(d) MLIV

Figure A5: Elasticity Estimation Comparison
Figure A5 presents the estimated own-elasticity of our model without IV, with BLP Style IVs, with differentiation IVs and with MLIV. The x-axis represents the price of the focal product, while the y-axis shows the product's own-elasticity. Each point corresponds to a product in a market, resulting in 2,217 observations. We report the estimated elasticity based on the same price variation used in the BLP paper (a 1,000-dollar change).



(a) Own-Elasticity Estimation (Our Model vs. BLP Model)

(b) Cross-Elasticity Estimation (Our Model vs. BLP Model)

Figure A6: Elasticity Estimation Comparison
Note: Figure A6 illustrates the distributions of the estimated own- and cross-elasticities obtained from our model (using different sets of IVs) and the BLP model. The filled areas in the violin plots represent the complete range of the elasticities, while the text labels indicate the mean values.

In addition, we also perform a weak instrument test on both BLP Style IVs and the MLIV and report the F-statitics and p-value in Table A15. Both BLP Style IVs and MLIV pass the weak instrument tests.

	F-statistic	P-value
BLP Style IVs	241.5	$<$ 1e-8
MLIV	280.9	$<$1e-8

Table A15: Weak Instrument Test

H Choice Models Satisfying Permutation Invariance

Table A16: Choice Models Satisfying Permutation Invariance

Choice Model	Literature
Multinomial Logit Model	McFadden et al. (1973)
Nested Logit Model	Train et al. (1987)
Mixed Logit Model	McFadden and Train (2000)
Generalized Extreme Value (GEV) Model	Train (2009)
Probit Model	Hausman and Wise (1978)
Latent Class Logit Model	Kamakura and Russell (1989)
Random Coefficients Nested Logit	Grigolon and Verboven (2014)
Markov Chain Choice Model	Blanchet et al. (2016)
Customer Inattention Based Models	Goeree (2008)
Customer Search Models	Mehta et al. (2003)