Web Appendix

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

By Amandeep Singh, Ye Liu, and Hema Yoganarasimhan

A Identity Independence and Permutation Invariance of Traditional Choice Models

In this Web Appendix, we analyze various choice models with respect to the identity independence and permutation invariance properties. Table 1 categorizes these models based on whether they satisfy these properties.

We first examine models that satisfy both properties. Identity independence is achieved when all the information about a product utilized in the model is included in its features. To demonstrate permutation invariance, we permutate any two competitor products associated with a focal product and show that such permutations do not alter the demand for the focal product.

We then present models that violate the permutation invariance assumption, explaining why the demand for a focal product changes when competitor products are permutated, even when their features remain identical.

Since all models require parametric assumptions about the utility function, for the sake of simplicity in notation, we assume that the utility of a product is a parametric function of its observed features X_{jt} and price p_{jt}, determined by a set of parameters β unless otherwise specified. Mathematically, we express the utility as:

$$v_{jt} = f(X_{jt}, p_{jt}; \beta)$$

A.1 MNL

Assuming there are J products, the market share of each product under the MNL model is

$$\pi_{jt} = g(X_{jt}, p_{jt}, \{X_{kt}, p_{kt}\}_{k \in S, k \neq j}) = \frac{exp(f(X_{jt}, p_{jt}; \beta))}{exp(f(X_{jt}, p_{jt}; \beta) + \sum_{k \in \mathcal{S}, k \neq j} exp(f(X_{kt}, p_{kt}; \beta))}. \quad \text{(A1)}$$

If we permute two products k_1 and k_2 in the set $\mathcal{S} \backslash \{j\}$, the only change is in the order of terms in the summation in the denominator of the market share. Since addition is commutative, the sum $\sum_{k \in \mathcal{S}, k \neq j} exp(f(X_{kt}, p_{kt}; \beta))$ remains the same regardless of the order of k_1 and k_2. Therefore, the value of π_{jt} remains unchanged, demonstrating that the MNL model satisfies permutation invariance.

A.2 Mixed Logit

Assuming there are J products, the market share of each product under the RCL model is

$$\pi_{jt} = g(X_{jt}, \{X_{kt}\}_{k \in S, k \neq j}) = \int \frac{exp(f(X_{jt}, p_{jt}; \beta_i))}{exp(f(X_{jt}, p_{jt}; \beta_i) + \sum_{k \in \mathcal{S}, k \neq j} exp(f(X_{kt}, p_{kt}; \beta_i))} dF(\beta_i) \quad \text{(A2)}$$

, where $F(\beta_i)$ is the CDF of random coefficient β_i. Similar to MNL, if we permute the position of any other two products $k_1, k_2 \in \mathcal{S} \backslash \{j\}$, it only affects the sequence of the summation in the denominator, thus the demand of product j remains the same. Therefore, it satisfies the permutation invariance.

A.3 Nested Logit

In a nested logit model, products are grouped into nests to account for the correlation in unobserved factors among products within the same nest. Consider a scenario where there are J products categorized into L nests. The market share of a product j in a nest l is calculated taking into account this nesting structure. A key aspect of showing permutation invariance in the Nested Logit model is to include the nest affiliation of each product as a feature. This is represented by an L-dimensional 0-1 vector $N_{jt} \in \mathbb{R}^{1\times L}$, which denotes the affiliation of product jt to one of the L nests. Each element of the vector corresponds to a specific nest, indicating the product's membership in that nest. Specifically,

$$\pi_{jt} = g(X_{jt}, p_{jt}, N_{jt}, \{X_{kt}, p_{kt}, N_{kt}\}_{k\in S, k\neq j}). \tag{A3}$$

Then, the Nested Logit model can be represented as:

$$\pi_{jt} = P(jt|l)P(l) \tag{A4}$$

$$P(j|l) = \frac{\exp\left(\frac{v_{jt}}{N_{jt}\sigma}\right)}{\sum_{k\in\mathcal{N}_l}\exp\left(\frac{v_{kt}}{N_{jt}\sigma}\right)} \tag{A5}$$

$$P(l) = \frac{\exp\left(N_l\sigma\log\left(\sum_{k\in\mathcal{N}_l}\exp\left(\frac{v_{kt}}{N_{kt}\sigma}\right)\right)\right)}{\sum_{m=1}^{L}\exp\left(N_m\sigma\log\left(\sum_{k\in\mathcal{N}_m}\exp\left(\frac{v_{kt}}{N_m\sigma}\right)\right)\right)} \tag{A6}$$

Where:

- $\mathcal{N}_l$ is the set of products in nest l.
- $\sigma \in R^L$ is an L-dimentional vector where the lth element represents the scale parameter for nest l which captures the correlation in unobserved utilities among products in the same nest. Therefore, the scale parameter of product jt, which belongs to nest l, can be expressed as $N_{jt}\sigma$.
- N_m denotes a binary vector of dimension L, in which the mth element is one and all other elements are zero. Therefore, $N_m\sigma$ is the scale parameter of the nest m.

It is intuitive that when we swap the position of any two products, the probability of choosing a nest ($P(l)$) would not change because both the denominator and the numerator only depend on the nest affiliations. Moreover, $P(j|l)$ also satisfies permutation invariance because of the additive nature of the denominator in its formula, where the sum over all products in nest l remains constant despite any permutation of product positions within the nest. The permutation invariance property could be easily extended to that random coefficient nested logit following the same logic as random coefficient logit.

A.4 Latent Class Logit Model

In a latent class logit model, products can be divided into multiple latent classes and each latent class has a distinct set of parameters. Assume there are c latent classes, the utility of a product in class c is denoted by $v_{jt,c} = f(X_{jt}, p_{jt}; \beta_c)$. The probability of one product belonging to a class c is determined by the linear combination of Z_{jt} with parameters τ_c.

$$\begin{aligned}\pi_{jt} &= g(X_{jt}, p_{jt}, Z_{jt}, \{X_{kt}, p_{kt}, Z_{kt}\}_{k\in S, k\neq j}) \\ &= \sum_c \frac{\exp(f(X_{jt}, p_{jt}; \beta_c))}{\exp(f(X_{jt}, p_{jt}; \beta_c)) + \sum_{k\in S, k\neq j}\exp(f(X_{kt}, p_{kt}; \beta_c)} \frac{\exp(\tau_c Z_{jt})}{\sum_c \exp(\tau_c Z_{jt})}\end{aligned} \tag{A7}$$

Similarly, the $\sum_{k\in S,k\neq j}\exp(f(X_{kt},p_{kt};\beta_c)$ is commutative thus the demand remains the same when we alter the order of any two products except jt. Therefore, latent class models satisfy the permutation invariance.

A.5 Consumer Inattention and Search Model

In this part, we specifically demonstrate the permutation invariance property of the consumer inattention model that we used in the numerical experiments. Our analysis could easily be extended to other models based on consumer inattention in the literature, such as Goeree (2008), Turlo et al. (2023), Compiani (2022) and Joo (2023), as well as search models, such as Mehta et al. (2003).

In our simulation, we generate the market shares of products by considering a segment of consumers who ignore the highest-priced product. Here we generalize the specification by assuming the portion of the inattentive consumers is a function of observed features of the highest-priced product $1-s(X_{ht})$, where h denotes the highest-priced product. A consumer forms their consideration set with the products they pay attention to. For products in their consideration set, they use the logit function as the decision rule. We allow for each consumer to have a set of random coefficients. Then the choice probability of the highest-priced product becomes

$$\pi_{ht}=g(X_{ht},\{X_{kt}\}_{k\in\mathcal{S},k\neq h})=\frac{s(X_{ht})exp(\beta_i X_{ht})}{f(X_{jt},p_{jt};\beta_i)+\sum_{k\in\mathcal{S},k\neq h}f(X_{kt},p_{kt};\beta_i)} \tag{A8}$$

When we permutate the order of any two competitor products, the choice probability of the highest-price product remains the same. The choice probability of other products can be expressed as

$$\begin{aligned}\pi_{jt}&=g(X_{jt},\{X_{kt}\}_{k\in\mathcal{S},k\neq j})\\&=\frac{s(X_{ht})exp(f(X_{kt},p_{jt};\beta_i))}{exp(f(X_{kt},p_{jt};\beta_i))+\sum_{k\in\mathcal{S},k\neq j}f(X_{kt},p_{kt};\beta_i)}+\frac{(1-s(X_{ht}))exp(f(X_{kt},p_{jt};\beta_i))}{exp(f(X_{kt},p_{jt};\beta_i))+\sum_{k\in\mathcal{S},k\neq j,h}f(X_{kt},p_{kt};\beta_i)}\end{aligned} \tag{A9}$$

Similar to the highest-priced product, the choice probability for other products remains unaffected when the order of competitor products is altered. This is because the feature values of the most expensive product X_h do not change even though the index h may change. Therefore, consumer inattention models are also permutation invariant.

A.6 Other Models

We have illustrated that the Multinomial Logit model, Nested Logit model, Mixed Logit model, and Latent Class Logit model all exhibit the property of permutation invariance. It becomes apparent that other models, which employ a similar random utility framework but differentiate in the probability distribution of the error term, likewise possess this permutation invariance property.

For example, this permutation invariance property extends Generalized Extreme Value (GEV) models (Train, 2009), where the error term follows the generalized extreme value distribution, and the Probit model (Hausman and Wise, 1978; Chintagunta, 2001), where the error term follows a normal distribution. Although there is no straightforward closed-formed representation of choice probability for these models, the choice probability can be expressed as

$$\pi_{ijt}=\Pr(u_{ijt}\geq u_{ikt},\forall k\neq j,k\in\mathcal{S}_t), \tag{A10}$$

where u_{ijt} represents the consumer i's utility of product j in market t. Since u_{ikt} for any $k\in\mathcal{S}_t$ is parameterized by its own feature (X_{kt}), permuting any two $k\neq j,k\in\mathcal{S}_t$, does not affect the choice probability of π_{ijt} as well as the aggregate demand π_{jt}. Indeed, this shows that any models building on the random utility framework with the decision rule as stated in Equation A10 are permutation invariant.

The permutation invariance property also extends to other recently developed choice models, such as Markov Chain Choice model(Blanchet et al., 2016). The choice probability of a product jt is modeled as a Markov Chain, which consists of two parts: arrival probability q_{jt} and transition probability $p_{jk,t}$, defined as

$$q_{jt} = f(j, \mathcal{S}_t), \tag{A11}$$

$$p_{jk,t} = \begin{cases} 1, & \text{if } j = 0 \text{ and } k = 0, \\ \frac{f(j,\mathcal{S}_t \setminus \{j\}) - f(k,\mathcal{S}_t)}{f(j,\mathcal{S}_t)}, & \text{if } j, k \in \mathcal{S}_t, k \neq j, \\ 0, & \text{otherwise.} \end{cases} \tag{A12}$$

Given that both q_{jt} and $p_{jk,t}$ only rely on jt, $\mathcal{S}_t$ and $\mathcal{S}_t \setminus \{j\}$, permuting the order of any two products that are not the focal product jt in the market t does not change both arrival probability and transition probability. Therefore, the choice probability does not change.

A.7 Almost Ideal Demand System (AIDS)

Unlike traditional discrete choice models, the Almost Ideal Demand System (AIDS) specifies demand directly through expenditure shares as a function of prices and total expenditure, rather than through random utility maximization. The AIDS model, proposed by Deaton and Muellbauer (1980), defines the budget share w_{jt} of product j as:

$$w_{jt} = \nu_j + \sum_{k=1}^{J} \iota_{jk} \log(p_{kt}) + \beta_j \log\left(\frac{x_t}{P_t}\right), \tag{A13}$$

where p_{kt} is the price of product k, x_t is total expenditure, and P_t is the price index. The AIDS model explicitly incorporates cross-price elasticities via parameters ι_{jk}, which measure the responsiveness of the expenditure share of product j to changes in the price of product k. Importantly, these cross-price elasticities are specific to product pairs and are not symmetric in general.

When two products' positions are permuted, their identities directly affect the structure and interpretation of these cross-price elasticities. Unlike models relying on additive utilities, where permutation merely rearranges summation order, in the AIDS model, permuting two products k_1 and k_2 leads to a fundamental alteration of the parameters ι_{jk} linked to product identities. Therefore, the AIDS model does not satisfy permutation invariance, as changing the order of product identities alters the structure of cross-price relationships.

A.8 Models with Flexible Error Structures

Any demand system whose error structure (e.g., the covariance of unobserved components) depends on the labels of products rather than on a symmetric function of their characteristics is not permutation invariant. In essence, if the "off-diagonal" relationships in the covariance matrix differ simply because of which product is called "i" and which is called "j," the model predictions will change when we swap labels.

An example is Random Utility Models (RUMs) with a flexible covariance structure, such as the Multinomial Probit Model. In these models, individual n's utility for alternative i is given by:

$$U_{ni} = V_{ni} + \varepsilon_{ni},$$

where V_{ni} is the systematic utility component and ε_{ni} is an error term capturing unobserved factors. If the vector of errors

$$\boldsymbol{\varepsilon}_n = (\varepsilon_{n1}, \varepsilon_{n2}, \ldots, \varepsilon_{nJ})^\top$$

follows a general multivariate normal distribution

$$\boldsymbol{\varepsilon}_n \sim \mathcal{N}(\mathbf{0}, \Sigma),$$

then the structure of Σ determines whether permutation invariance holds. If Σ is not permutation invariant (i.e., if the covariance terms are indexed by product labels rather than a symmetric function of characteristics), then relabeling alternatives will alter choice probabilities. For example, if some products are modeled as having stronger correlations with specific others based on their index positions rather than their characteristics, the invariance property fails. However, also note that if correlation structure across products can be modelled as product characteristics the resultant choice model will again satisfy permutation invariance.

B Proof of Main Results

B.1 Proof of Theorem 1

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 $\{X_{jt}, p_{jt}, I_{it}, \varepsilon_{ijt}\}$ satisfies Assumption 1, 2 and 3, then there exists suitable ρ, ϕ_1 and ϕ_2 such that

$$\pi_{jt} = \rho(\phi_1(X_{jt}, p_{jt}) + \sum_{k \neq j, k \in \mathcal{S}_t} \phi_2(X_{kt}, p_{kt})), \tag{A14}$$

Proof. The sufficiency follows by observing that the function $\pi_{jt} = \rho(\phi_1(X_{jt}, p_{jt}) + \sum_{k \in S \setminus \{j\}} \phi_2(X_{kt}, p_{kt}))$ 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 choice functions satisfying assumptions 2 (identity independence) and 3 (permutation invariance) 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 $\phi_1(X_{jt}) = 4^{-c^e(X_{jt}, p_{jt})}$ and $\phi_2(X_{kt}, p_{kt}) = 4^{-c^o(X_{kt}, p_{kt})}$ then $\phi_1(X_{jt}, p_{jt}) + \sum_{k \in \mathcal{S}_t \setminus \{j\}} \phi_2(X_{kt}, p_{kt})$ constitutes an unique representation for every product j and competing assortment $\mathcal{S}_t \setminus \{j\}$. Now a function $\rho : \mathbb{R} \to \mathbb{R}$ can always be constructed such that $\pi_{jt} = \rho\left(\phi_1(X_{jt}, p_{jt}) + \sum_{k \in S \setminus \{j\}} \phi_2(X_{kt}, p_{kt})\right)$.

B.2 Convergence Rate and Estimation of Riesz Representers

Theorem 4. [Chernozhukov et al. (2021)] One can view the Riesz representer as the minimizer of the loss function:

$$\begin{aligned} \alpha_0 &= \arg\min_{\alpha} \mathbb{E}\left[(\alpha(z) - \alpha_0(z))^2\right] \\ &= \arg\min_{\alpha} \mathbb{E}\left[\alpha(z)^2 - 2\alpha_0(z)\alpha(z) + \alpha_0(z)^2\right] \\ &= \arg\min_{\alpha} \mathbb{E}\left[\alpha(z)^2 - 2m(w, \alpha)\right]. \end{aligned} \tag{A15}$$

Now, we explain the intuition behind this theorem. Consider the objective function we use in the optimization problem.

$$\mathbb{E}\Big[\big(\alpha(z) - \alpha_0(z)\big)^2\Big]. \tag{A16}$$

Expanding and splitting the expectation yields:

$$\mathbb{E}[\alpha(z)^2] - 2\,\mathbb{E}\big[\alpha_0(z)\,\alpha(z)\big] + \mathbb{E}[\alpha_0(z)^2]. \tag{A17}$$

Since $\mathbb{E}[\alpha_0(z)^2]$ is constant with respect to α, minimizing the objective is equivalent to minimizing:

$$\mathbb{E}[\alpha(z)^2] - 2\,\mathbb{E}\big[\alpha_0(z)\,\alpha(z)\big]. \tag{A18}$$

Now, lets consider definition of $m(\cdot, \pi)$. The Riesz Representation Theorem guarantees that there exists a function α_0 defined on z such that for any function π,

$$\mathbb{E}\big[m(w, \pi)\big] = \mathbb{E}\big[\alpha_0(z)\, \pi(z)\big]. \tag{A19}$$

Thus, we also have:

$$\mathbb{E}\big[m(w, \alpha)\big] = \mathbb{E}\big[\alpha_0(z)\, \alpha(z)\big]. \tag{A20}$$

By substituting $\mathbb{E}[m(w, \alpha)]$ for $\mathbb{E}[\alpha_0(z)\, \alpha(z)]$, our minimization problem becomes:

$$\arg\min_{\alpha} \left\{ \mathbb{E}\Big[\alpha(z)^2 - 2\, m(w, \alpha)\Big] \right\}. \tag{A21}$$

Theorem 5. [Chernozhukov et al. (2021)] Let δ_n be an upper bound on the critical radius (Wainwright (2019)) of the function spaces:

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

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

$$\mathbb{E}\left[\big(m(w, \alpha) - m\left(w, \alpha'\right)\big)^2\right] \leq M \left\|\alpha - \alpha'\right\|_2^2 \tag{A22}$$

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

$$\begin{aligned} \|\widehat{\alpha} - \alpha_0\|_2^2 \leq C\Big(\delta_n^2 M + \frac{M \log(1/\zeta)}{n} \\ + \inf_{\alpha_* \in \mathcal{A}_n} \|\alpha_* - \alpha_0\|_2^2\Big) \end{aligned} \tag{A23}$$

The critical radius is defined as the smallest radius around a true parameter or function within which the complexity of the function space (measured by metrics such as covering numbers or entropy integrals) is sufficiently controlled. In simpler terms, it provides an upper bound that characterizes how large deviations between an estimated function and the true function can become, given the complexity of the function class considered. Intuitively, the critical radius measures how complex the estimation problem is relative to the sample size; it indicates how precisely we can estimate the unknown function given the available data and complexity constraints of our model class. In practical terms, the smaller the critical radius, the more tightly we can guarantee convergence and reliable estimation of the function. The result shows that the estimation error in α (RR) is driven by three terms: (i) the complexity of the function class, $\delta_n^2 M$, which reflects the *statistical complexity* of the function class used to estimate the Riesz representer—more complex models (e.g., deeper neural networks) can approximate richer functions but also require more data to learn accurately; (ii) the variance due to finite samples, captured by the term $\frac{M \log(1/\zeta)}{n}$, which shrinks as the sample size grows and ensures uniform convergence with high probability $1 - \zeta$; and (iii) the approximation error, $\inf_{\alpha_* \in \mathcal{A}_n} \|\alpha_* - \alpha_0\|_2^2$, which measures how well the best function in the chosen class $\mathcal{A}_n$ can represent the true α_0. In our research, we focus on applying Theorem 4 from an application standpoint to neural networks.

B.3 Proof of Proposition 1

Proposition 1. *If Assumptions 5–8 are satisfied, then for*

$$V = E\left[\{m(w, \pi_0(z;\gamma_0)) - \theta_0 + \alpha_0(z;\gamma_0)\,(y - \pi_0(z;\gamma_0))\}^2\right],$$

we have

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

where

$$\hat{\theta} = \frac{1}{n}\sum_{\ell=1}^{L}\sum_{t\in\mathcal{I}_\ell}\left\{m\left(w_{t^*}, \hat{\pi}_\ell(z_{t^*};\hat{\gamma}_\ell)\right) + \hat{\alpha}_\ell(z_{t^*};\hat{\gamma}_\ell)\left(y_{t^*} - \hat{\pi}_\ell(z_{t^*};\hat{\gamma}_\ell)\right)\right\},$$

$$\hat{V} = \frac{1}{n}\sum_{\ell=1}^{L}\sum_{t\in\mathcal{I}_\ell}\hat{\psi}^2_{t^*\ell}, \quad \hat{\psi}_{t^*\ell} = m\left(w_{t^*}, \hat{\pi}_\ell(z_{t^*};\hat{\gamma}_\ell)\right) - \hat{\theta} + \hat{\alpha}_\ell(z_{t^*};\hat{\gamma}_\ell)\left(y_{t^*} - \hat{\pi}_\ell(z_{t^*};\hat{\gamma}_\ell)\right).$$

Here, t^* is randomly drawn from each market t.

Proof. To show the asymptotic normality we will first verify the Assumptions 1-3 of Chernozhukov et al. (2022), from now on CEINR, with $g(w, \pi(z;\gamma), \theta) = m(w, \pi(z;\gamma)) - \theta$ and $\phi(w, \pi(z;\gamma), \alpha(z,\gamma), \theta) = \alpha(z,\gamma)\cdot(y - \pi(z;\gamma))$. Using Taylor series expansion, Assumption 7 and $||\hat{\pi}_\ell(z;\gamma) - \pi_0(z;\gamma)|| \xrightarrow{p} 0$ we have,

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

By the triangle inequality

$$\begin{aligned}
&\leq C\int ||\hat{\pi}_\ell(z_i;\hat{\gamma}_\ell) - \hat{\pi}_\ell(z_i;\gamma_0)||^2 \mathcal{P}_0(dw) \\
&\quad + C\int ||\hat{\pi}_\ell(z_i;\gamma_0) - \pi_0(z_i;\gamma_0)||^2 \mathcal{P}_0(dw) \\
&\quad + C\int ||\hat{\pi}_\ell(z_i;\hat{\gamma}_\ell) - \hat{\pi}_\ell(z_i;\gamma_0)||||\hat{\pi}_\ell(z_i;\gamma_0) - \pi_0(z_i;\gamma_0)||\mathcal{P}_0(dw) \xrightarrow{p} 0
\end{aligned} \tag{A24}$$

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

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

$$\begin{aligned}\int ||\phi(w,\hat{\pi}_\ell(z;\hat{\gamma}_\ell),\alpha_0(z;\gamma_0),\theta_0) - \phi(w,\pi_0,\alpha_0,\theta_0)||^2\mathcal{P}_0(dw) &= \int ||\alpha_0(z)(\pi_0(z;\gamma_0) - \hat{\pi}_\ell(z;\hat{\gamma}_\ell))||^2\mathcal{P}_0(dw) \\ &\leq C\int ||(\pi_0(z;\gamma_0) - \hat{\pi}_\ell(z;\hat{\gamma}_\ell))||^2\mathcal{P}_0(dw) \\ &\leq C||\hat{\pi}_\ell(z;\hat{\gamma}_\ell) - \pi_0(z;\gamma_0)||^2 \xrightarrow{p} 0\end{aligned} \tag{A25}$$

Next, we need to show that

$$\|\hat{\alpha}_\ell(z;\hat{\gamma}_\ell) - \alpha_0(z;\gamma_0)\| \xrightarrow{p} 0,$$

observe that

$$\|\hat{\alpha}_\ell(z;\hat{\gamma}_\ell) - \alpha_0(z;\gamma_0)\| \;\leq\; \|\hat{\alpha}_\ell(z;\hat{\gamma}_\ell) - \alpha_0(z;\hat{\gamma}_\ell)\| \;+\; \|\alpha_0(z;\hat{\gamma}_\ell) - \alpha_0(z;\gamma_0)\|.$$

By by Assumption 5 i) for each γ near γ_0, $\|\hat{\alpha}_\ell(z;\gamma) - \alpha_0(z;\gamma)\| \xrightarrow{p} 0$, and by Assumption 7 $\hat{\gamma}_\ell \xrightarrow{p} \gamma_0$. Thus we have

$$\|\hat{\alpha}_\ell(z;\hat{\gamma}_\ell) - \alpha_0(z;\hat{\gamma}_\ell)\| \xrightarrow{p} 0.$$

Next, by continuity of $\alpha_0(z;\gamma)$ in γ at γ_0 and the fact that $\hat{\gamma}_\ell \xrightarrow{p} \gamma_0$, we also get

$$\|\alpha_0(z;\hat{\gamma}_\ell) - \alpha_0(z;\gamma_0)\| \xrightarrow{p} 0.$$

Hence each term on the right converges to 0 in probability, and so the sum does as well:

$$\|\hat{\alpha}_\ell(z;\hat{\gamma}_\ell) - \alpha_0(z;\gamma_0)\| \xrightarrow{p} 0.$$

Thus, next we have,

$$\begin{aligned}&\int \left\|\phi(w,\pi_0(z;\gamma_0),\hat{\alpha}_\ell(z;\hat{\gamma}_\ell),\tilde{\theta}) - \phi(w,\pi_0(z;\gamma_0),\alpha_0,\theta_0)\right\|^2 \mathcal{P}_0(dw) \\ &= \int \|(\hat{\alpha}_\ell(z;\hat{\gamma}_\ell) - \alpha_0(z;\gamma_0))\,(y - \pi_0(z;\gamma_0))\|^2\,\mathcal{P}_0(dw) \\ &\leq C\int \|\hat{\alpha}_\ell(z;\hat{\gamma}_\ell) - \alpha_0(z;\gamma_0)\|^2\,\mathcal{P}_0(dw) \\ &\leq C\,\|\hat{\alpha}_\ell(z;\hat{\gamma}_\ell) - \alpha_0(z;\gamma_0)\|^2 \xrightarrow{p} 0\end{aligned} \tag{A26}$$

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

$$\begin{aligned}\hat{\Delta}_\ell(w) := \phi\left(w,\hat{\pi}_\ell(z;\hat{\gamma}_\ell),\hat{\alpha}_\ell(z;\hat{\gamma}_\ell),\tilde{\theta}\right) - \phi\left(w,\pi_0(z;\gamma_0),\hat{\alpha}_\ell(z;\hat{\gamma}_\ell),\tilde{\theta}\right) \\ - \phi\left(w,\hat{\pi}_\ell(z;\hat{\gamma}_\ell),\alpha_0,\theta_0\right) + \phi\left(w,\pi_0(z;\gamma_0),\alpha_0(z;\gamma_0),\theta_0\right) \\ = -\left[\hat{\alpha}_\ell(z;\hat{\gamma}_\ell) - \alpha_0(z;\gamma_0)\right]\left[\hat{\pi}_\ell(z;\hat{\gamma}_\ell) - \pi_0(z;\gamma_0)\right]\end{aligned} \tag{A27}$$

Then by the Cauchy–Schwarz inequality and Assumptions 7 i) and ii),

$$E\left[\hat{\Delta}_\ell(w)\right] = \int -\left[\hat{\alpha}_\ell(z;\hat{\gamma}_\ell) - \alpha_0(z;\gamma_0)\right]\left[(\hat{\pi}_\ell(z;\hat{\gamma}_\ell) - \pi(z;\gamma_0))\right]\mathcal{P}_0(dz)$$
$$\leq \|\hat{\alpha}_\ell(z;\hat{\gamma}_\ell) - \alpha_0(z;\gamma_0)\|\,\|(\hat{\pi}_\ell(z;\hat{\gamma}_\ell) - \pi(z;\gamma_0))\| \tag{A28}$$

We want to control the product

$$\left\|\hat{\alpha}_\ell(z,\hat{\gamma}_\ell) - \alpha_0(z,\gamma_0)\right\|\,\left\|\hat{\pi}_\ell(z,\hat{\gamma}_\ell) - \pi_0(z,\gamma_0)\right\|.$$

For α. Consider,

$$\hat{\alpha}_\ell(z,\hat{\gamma}_\ell) - \alpha_0(z,\gamma_0) \;=\; \underbrace{\left[\hat{\alpha}_\ell(z,\hat{\gamma}_\ell) - \alpha_0(z,\hat{\gamma}_\ell)\right]}_{A_1} \;+\; \underbrace{\left[\alpha_0(z,\hat{\gamma}_\ell) - \alpha_0(z,\gamma_0)\right]}_{A_2},$$

where

$$A_1 \;=\; \hat{\alpha}_\ell(z,\hat{\gamma}_\ell) - \alpha_0(z,\hat{\gamma}_\ell), \quad A_2 \;=\; \alpha_0(z,\hat{\gamma}_\ell) - \alpha_0(z,\gamma_0).$$

For π. Similarly,

$$\hat{\pi}_\ell(z,\hat{\gamma}_\ell) - \pi_0(z,\gamma_0) \;=\; \underbrace{\left[\hat{\pi}_\ell(z,\hat{\gamma}_\ell) - \pi_0(z,\hat{\gamma}_\ell)\right]}_{P_1} \;+\; \underbrace{\left[\pi_0(z,\hat{\gamma}_\ell) - \pi_0(z,\gamma_0)\right]}_{P_2},$$

where

$$P_1 \;=\; \hat{\pi}_\ell(z,\hat{\gamma}_\ell) - \pi_0(z,\hat{\gamma}_\ell), \quad P_2 \;=\; \pi_0(z,\hat{\gamma}_\ell) - \pi_0(z,\gamma_0).$$

Next consider

$$\left\|\hat{\alpha}_\ell(z,\hat{\gamma}_\ell) - \alpha_0(z,\gamma_0)\right\|\,\left\|\hat{\pi}_\ell(z,\hat{\gamma}_\ell) - \pi_0(z,\gamma_0)\right\| \;=\; \left\|A_1 + A_2\right\|\,\left\|P_1 + P_2\right\|.$$

By the triangle inequality for norms,

$$\|A_1 + A_2\| \;\leq\; \|A_1\| + \|A_2\|, \quad \|P_1 + P_2\| \;\leq\; \|P_1\| + \|P_2\|.$$

Hence,

$$\left\|A_1{+}A_2\right\|\,\left\|P_1{+}P_2\right\| \;\leq\; \left(\|A_1\|{+}\|A_2\|\right)\left(\|P_1\|{+}\|P_2\|\right) \;=\; \|A_1\|\|P_1\| + \|A_1\|\|P_2\| + \|A_2\|\|P_1\| + \|A_2\|\|P_2\|.$$

Substituting back the definitions of A_1, A_2, P_1, P_2 yields the sum of four terms:

$$\begin{aligned}
&\left\|\hat{\alpha}_\ell(z,\hat{\gamma}_\ell) - \alpha_0(z,\gamma_0)\right\|\,\left\|\hat{\pi}_\ell(z,\hat{\gamma}_\ell) - \pi_0(z,\gamma_0)\right\| \\
&\quad\leq \underbrace{\left\|\hat{\alpha}_\ell(z,\hat{\gamma}_\ell) - \alpha_0(z,\hat{\gamma}_\ell)\right\|\,\left\|\hat{\pi}_\ell(z,\hat{\gamma}_\ell) - \pi_0(z,\hat{\gamma}_\ell)\right\|}_{\text{Term (1)}} + \underbrace{\left\|\hat{\alpha}_\ell(z,\hat{\gamma}_\ell) - \alpha_0(z,\hat{\gamma}_\ell)\right\|\,\left\|\pi_0(z,\hat{\gamma}_\ell) - \pi_0(z,\gamma_0)\right\|}_{\text{Term (2)}} \\
&\qquad + \underbrace{\left\|\alpha_0(z,\hat{\gamma}_\ell) - \alpha_0(z,\gamma_0)\right\|\,\left\|\hat{\pi}_\ell(z,\hat{\gamma}_\ell) - \pi_0(z,\hat{\gamma}_\ell)\right\|}_{\text{Term (3)}} + \underbrace{\left\|\alpha_0(z,\hat{\gamma}_\ell) - \alpha_0(z,\gamma_0)\right\|\,\left\|\pi_0(z,\hat{\gamma}_\ell) - \pi_0(z,\gamma_0)\right\|}_{\text{Term (4)}}.
\end{aligned} \tag{A29}$$

From assumption 6, α_0 and π_0 are each Lipschitz in γ. That is, there exist constants L_{α_0} and L_{π_0} such

that, for all $\gamma, \hat{\gamma}_\ell$,

$$\left\|\alpha_0(z,\hat{\gamma}_\ell) - \alpha_0(z,\gamma)\right\| \;\le\; L_{\alpha_0}\,\|\hat{\gamma}_\ell - \gamma\|, \qquad \left\|\pi_0(z,\hat{\gamma}_\ell) - \pi_0(z,\gamma)\right\| \;\le\; L_{\pi_0}\,\|\hat{\gamma}_\ell - \gamma\|.$$

Then the terms in (A29) simplify as follows:

- **Term (2):**

$$\left\|\hat{\alpha}_\ell(z,\hat{\gamma}_\ell) - \alpha_0(z,\hat{\gamma}_\ell)\right\| \left\|\pi_0(z,\hat{\gamma}_\ell) - \pi_0(z,\gamma_0)\right\| \;\le\; \left\|\hat{\alpha}_\ell(z,\hat{\gamma}_\ell) - \alpha_0(z,\hat{\gamma}_\ell)\right\| \left[L_{\pi_0}\,\|\hat{\gamma}_\ell - \gamma_0\|\right].$$

- **Term (3):**

$$\left\|\alpha_0(z,\hat{\gamma}_\ell) - \alpha_0(z,\gamma_0)\right\| \left\|\hat{\pi}_\ell(z,\hat{\gamma}_\ell) - \pi_0(z,\hat{\gamma}_\ell)\right\| \;\le\; \left[L_{\alpha_0}\,\|\hat{\gamma}_\ell - \gamma_0\|\right] \left\|\hat{\pi}_\ell(z,\hat{\gamma}_\ell) - \pi_0(z,\hat{\gamma}_\ell)\right\|.$$

- **Term (4):**

$$\left\|\alpha_0(z,\hat{\gamma}_\ell) - \alpha_0(z,\gamma)\right\| \left\|\pi_0(z,\hat{\gamma}_\ell) - \pi_0(z,\gamma_0)\right\| \;\le\; \left[L_{\alpha_0}\,\|\hat{\gamma}_\ell - \gamma_0\|\right] \left[L_{\pi_0}\,\|\hat{\gamma}_\ell - \gamma_0\|\right] \;=\; L_{\alpha_0} L_{\pi_0}\,\|\hat{\gamma}_\ell - \gamma_0\|^2.$$

Putting it all together and from assumption 7, we get the final bound:

$$\begin{aligned}
&\left\|\hat{\alpha}_\ell(z,\hat{\gamma}_\ell) - \alpha_0(z,\gamma_0)\right\| \left\|\hat{\pi}_\ell(z,\hat{\gamma}_\ell) - \pi_0(z,\gamma_0)\right\| \\
&\quad\le \underbrace{\left\|\hat{\alpha}_\ell(z,\hat{\gamma}_\ell) - \alpha_0(z,\hat{\gamma}_\ell)\right\| \left\|\hat{\pi}_\ell(z,\hat{\gamma}_\ell) - \pi_0(z,\hat{\gamma}_\ell)\right\|}_{\text{Term (1): estimator error product}} \\
&\qquad + \underbrace{\left\|\hat{\alpha}_\ell(z,\hat{\gamma}_\ell) - \alpha_0(z,\hat{\gamma}_\ell)\right\| \left[L_{\pi_0}\,\|\hat{\gamma}_\ell - \gamma_0\|\right]}_{\text{Term (2)}} + \underbrace{\left[L_{\alpha_0}\,\|\hat{\gamma}_\ell - \gamma_0\|\right] \left\|\hat{\pi}_\ell(z,\hat{\gamma}_\ell) - \pi_0(z,\hat{\gamma}_\ell)\right\|}_{\text{Term (3)}} \\
&\qquad + \underbrace{L_{\alpha_0} L_{\pi_0}\,\|\hat{\gamma}_\ell - \gamma_0\|^2}_{\text{Term (4)}} = o_p\left(\frac{1}{\sqrt{n}}\right).
\end{aligned} \tag{A30}$$

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

$$\begin{aligned}
\int \left\|\hat{\Delta}(w)\right\|^2 \mathcal{P}_0(dw) &= \int \left[\hat{\alpha}_\ell(z) - \alpha_0(z)\right]^2 \left[(\hat{\pi}_\ell(z;\hat{\gamma}_\ell) - \pi_0(z))\right]^2 \mathcal{P}_0(dz) \\
&\le C \int \left[(\hat{\pi}_\ell - \pi_0(z))\right]^2 \mathcal{P}_0(dz) \xrightarrow{p} 0
\end{aligned} \tag{A31}$$

Thus Equation A30 and Equation A31 verify Assumption 2 i) of CEINR.

Next Assumption 3 of CEINR is satisfied through Assumption 8. 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, $\int \left\|g\left(w, \hat{\pi}_\ell(z;\hat{\gamma}_\ell), \tilde{\theta}\right) - g\left(w, \hat{\pi}_\ell(z;\hat{\gamma}_\ell), \theta_0\right)\right\|^2 \mathcal{P}_0(dw) \xrightarrow{p} 0$

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

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

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

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

□

C Imposing probability constraints without sacrificing permutation invariance.

While the proposed framework does not rely on restrictive probability and discrete-choice constraints, researchers may sometimes wish to impose them for interpretability or to align with traditional formulations. In this appendix, we discuss such constraints and illustrate how they can be incorporated within our architecture while retaining permutation invariance.

- **Choice probabilities are positive (and bounded):** If one wishes to ensure positivity or $[0, 1]$-bounded outputs for each product, this can be enforced at the final layer per product using standard links (e.g., *softplus*/$\exp$ for nonnegativity or *sigmoid* for $(0, 1)$). These impose the same structural constraints used in traditional choice models.
- **Probabilities summing to one:** In standard choice models (e.g., MNL), probabilities sum to one because of the restrictive assumption that each consumer selects exactly one product from the assortment. This is a limitation rather than an advantage—many real-world markets exhibit consumers purchasing more than one product. Our framework does not impose this restriction by default, enabling a more data-driven estimation of consumer preferences.

- **Relaxing discrete choice: continuous choices and quantities (aggregate demand):** Our framework even relaxes the discrete-choice assumption and allows continuous choices at the product level (e.g., multi-unit purchases), so outputs need not be probabilities. It can model *demand quantities directly* rather than only discrete choice probabilities. In this formulation, the "choice function" in Definition 1 maps index tuples to *nonnegative demand quantity vectors* (enforced with a nonnegative link such as softplus or $\exp$), accommodating both single- and multi-purchase settings as well as continuous purchase intensities, without imposing unnecessary probability constraints.[1]

If a researcher nevertheless wishes to impose the usual probability constraints and ensure probability summing to one, this can be done within our architecture while retaining permutation invariance. Concretely, compute a per-product score via the permutation-invariant structure in Theorem 1:

$$s_{jt} = h\left(\phi_1(X_{jt}, p_{jt}) + \sum_{k \neq j,\, k \in \mathcal{S}_t} \phi_2(X_{kt}, p_{kt})\right),$$

and then apply a market-level normalization over the set $\{s_{jt} : j \in \mathcal{S}_t\}$. A *softmax* yields $P_{jt} \in (0,1)$ with $\sum_{j \in \mathcal{S}_t} P_{jt} = 1$; including an *outside option* score s_{0t} in the softmax allows $\sum_{j \in \mathcal{S}_t} P_{jt} \leq 1$. Because the scoring function uses symmetric aggregation (sum) over competitors and the normalization is applied set-wise, the resulting probabilities remain permutation invariant.

D 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 A1: Hyperparameter Space for Tuning Non-parametric Estimator Benchmark

E Distribution of Features and Coefficients in Numerical Experiments

	MNL	RCL
p_{jt}	$U[0,4]$	$U[0,4]$
X_{jt}	$N(0,1)$	$N(0,1)$

Table A2: Distribution of Features

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

Notes : $\beta_k \sim N(0, 1/2d)$

Table A3: Distribution of Coefficients

[1] **Note:** In aggregate demand settings, this removes the need to assume or guess the total market size—an ad hoc step when converting observed quantities into market shares—since the model works directly with quantities.

F An Numerical Experiment of Endogeneity

In §3.3, we show how our model handles endogeneity in theory. Here, we provide a simulation that demonstrates the performance of our model handling endogeneity.

Let the utility u_{ijt} that consumer i in market t derives from product j as the following linear function

$$u_{ijt} = \eta_i p_{jt} + \beta_i X_{jt} + \kappa_i \mu_{jt} + \varepsilon_{ijt}, \tag{A32}$$

where μ_{jt} is the unobservable that correlates p_{jt} and ε_{ijt} is i.i.d. Type-I extreme value distributed. Specifically, without loss of generality, we specify the correlation between p_{jt} and μ_{jt} as

$$p_{jt} = \delta_1 IV_{jt} + \delta_2 \mu_{jt}, \tag{A33}$$

where IV_{jt} is the exogenous instrumental variable.

In this simulation, we first generate X_{jt} and μ_{jt}, following the same distribution as the X_{jt} in our baseline model, and IV_{jt}, following the same distribution as our as p_{jt} in our baseline model, as discussed in Web Appendix E. We next generate p_{jt} using Equation A33. Finally, we generate market shares assuming that the true model is RCL. For simplicity, we let $\delta_1 = \delta_2 = 1$. In addition, η_i and β_i follow the same distribution as the baseline simulations and κ_i follows the same distribution as β_i, as stated in Web Appendix E. We consider a case with 10 products, 100 markets, and 10 non-price features (in addition to price), the same as our baseline simulation in §5.1.1. In the estimation, we let μ_{jt} be an unobservable. We use the OLS linear regression as our first-stage regressor.

We consider two cases for comparison:

- **Exogeneous Benchmark**: Uses the exact same DGP but treat μ_{jt} as observables to researchers in estimation. Since the coefficients of price and market shares are the same as the endogeneity case, true elasticities are the same for these two DGPs. This gives a benchmark performance with the same data under the assumption that endogeneity is not present.

- **Ignoring Endogeneity**: Directly trains the model using only the observed features X_{jt} and p_{jt} without considering the endogeneity problem. Note that, even though the endogeneity problem is ignored, this does not mean the predictive performance in market shares would be low for this case because the model could be overfitted. However, the elasticities will be biased when the endogeneity is ignored.

Following the routine in our main text, we simulate 20 times for the same DGP with different parameters and features and report the predictive performance on market share ($\hat{\pi}_{jt}$), own price elasticity ($\frac{\partial \hat{\pi}_{jt}/\pi_{jt}}{\partial p_{jt}/p_{jt}}$), and cross-price elasticity ($\frac{\partial \hat{\pi}_{jt}/\pi_{jt}}{\partial p_{k\neq jt}/p_{k\neq jt}}$). We report the performance of our model in Table A4. When we use the control function to correct for endogeneity (Row 1: Our Method), our model underperforms slightly compared to the case where μ_{jt} is treated as an observable (Row 2: Exogenous Benchmark), on all three metrics. However, when we ignore the endogeneity issue and apply our model (Row 3: Ignore Endogneity), although the predictive performance of market shares is not bad, the estimation of own- and cross-elasticities are significantly biased, with the MAEs being 10–25 times higher than the cases where we account for endogeneity. This demonstrates both the importance of accounting for endogeneity as well as the effectiveness of our method in handling this problem in real settings.

G Simulations with Product Fixed-Effects

We now present a series of simulations to illustrate how our model performs when a product-level fixed effect is included in the data-generating process.

Specifically, we add product-level fixed effects into the data generation processes considered in §5.1.1,

	Market shares		Own-elasticity		Cross-elasticity	
	MAE	RMSE	MAE	RMSE	MAE	RMSE
Our Method	0.0256	0.0194	0.2307	0.3516	0.0669	0.1033
Exogeneous Benchmark	0.0254	0.0195	0.2290	0.3469	0.0657	0.1039
Ignore Endogeneity	0.0263	0.0202	2.3832	5.1901	1.9751	4.4306

Table A4: Model Performance of Endogeneity Case

Table A5: Predictive Performance in Fixed-Effect Models: Market Shares, Own-Elasticity, and Cross-Elasticity

(a) Market Shares ($\hat{\pi}_{jt}$)

#	True Model	J	T	d	Our Model		MNL		RCL		NP		MP		No. Obs.
					MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	
0	MNL with FE	5	100	0	0.0048	0.0614	0.0040	0.0898	0.0044	0.0667	0.0047	0.0615	0.1022	0.2859	100
1	RCL with FE	5	100	0	0.0028	0.0442	0.0249	0.1433	0.0034	0.0504	0.0091	0.0743	0.0756	0.2406	100
2	RCL-log() with FE	5	100	0	0.0148	0.0762	0.0498	0.2385	0.0386	0.1680	0.0673	0.1650	0.1540	0.3134	100
3	RCL-sin() with FE	5	100	0	0.0042	0.0539	0.0471	0.1933	0.0345	0.1509	0.0142	0.0981	0.0921	0.2674	100
4	RCL-Inattention with FE	5	100	0	0.0159	0.0946	0.0233	0.1348	0.0218	0.1292	0.0273	0.1395	0.0835	0.2541	100

(b) Own-Elasticity ($\frac{\partial \hat{\pi}_{jt}/\pi_{jt}}{\partial p_{jt}/p_{jt}}$)

#	True Model	J	T	d	Our Model		MNL		RCL		NP		No. Obs
					MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	
0	MNL with FE	5	100	0	0.0239	0.0560	0.0045	0.0062	0.0012	0.0023	0.0118	0.0464	8000
1	RCL with FE	5	100	0	0.0488	0.1408	0.4500	1.2841	0.0289	0.0737	0.1016	0.2884	8000
2	RCL-log() with FE	5	100	0	0.0952	0.2971	5.0892	5.1166	0.9638	1.6061	0.1545	0.4658	8000
3	RCL-sin() with FE	5	100	0	0.0558	0.2689	0.5330	1.0612	0.2467	0.6299	0.1459	0.3655	8000
4	RCL-Inattention with FE	5	100	0	0.1694	4.6926	0.7193	2.0145	0.4044	2.0275	0.4047	2.0149	8000

(c) Cross-Elasticity ($\frac{\partial \hat{\pi}_{jt}/\pi_{jt}}{\partial p_{k\neq jt}/p_{k\neq jt}}$)

#	True Model	J	T	d	Our Model		MNL		RCL		NP		No. Obs
					MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	
0	MNL with FE	5	100	0	0.0151	0.0328	0.0009	0.0017	0.0003	0.0006	0.0118	0.0454	32000
1	RCL with FE	5	100	0	0.0246	0.0649	0.0652	0.1869	0.0065	0.0215	0.0501	0.1136	32000
2	RCL-log() with FE	5	100	0	0.0378	0.1124	0.0789	0.2053	0.0693	0.5645	0.1137	0.3719	32000
3	RCL-sin() with FE	5	100	0	0.0291	0.0954	0.0690	0.2633	0.0640	0.2472	0.0660	0.1836	32000
4	RCL-Inattention with FE	5	100	0	0.0842	4.8431	0.0700	6.0230	0.0434	6.0321	0.1886	5.7391	32000

§5.1.2, and §5.1.3. For each data generation process, we simulate $u_{ijt,fe}$ as:

$$u_{ijt,fe} = u_{ijt} + \delta_j, \tag{A34}$$

where δ_j is the product-level fixed effect. We simulate the fixed effect of each product following a distribution the same as how we simulate coefficients of non-price features as shown in Table A3. For estimation, we include dummy variables for each product, treating these dummies as additional features. Hence, the total number of features becomes $d + J + 1$. To isolate the effect of incorporating a fixed effect from that of increasing the number of features, we include only one feature (price) in this simulation (i.e., $d = 0$).

Table A5 shows the simulation results for 20 random draws for a setting that has 5 products and 200 markets. We see that our model performs consistently better than benchmark models that are not true models. Interestingly, we observe that our model performs slightly better than the true model in the market share prediction when the true model is RCL (row 0 in Table A5a). This is because adding fixed effects also increases the difficulty in the estimation for traditional parametric methods.

Nevertheless, incorporating product fixed effects limits our model's ability to mitigate the curse of dimensionality caused by the number of products. Therefore, we see that the performance of our model is somewhat worse than the setting without fixed effects. For example, when incorporating consumer inattention behavior in the data generation, our model demonstrates superior results compared to other benchmark models for predicted market shares and own-elasticity (row 4 in Tables A5a and A5b) in terms of both MAE and RMSE, but only outperforms RCL for predicted cross-elasticity (row 4 in Table A5c) in terms of RMSE but not MAE.

H Logit Model with Bimodal Mixture Normal Distributed Random Coefficients

In §5.1.1, we show the performance of our model against RCL and MNL when the true model is either RCL or MNL. However, in many practical settings, researchers do not know the true distribution of the unobserved heterogeneity, and the underlying distributions may not be fixed (MNL) or normally distributed (like RCL). Therefore, we now present a simulation where the individual heterogeneity is not normally distributed. We then compare the performance of our approach against standard MNL and RCL as well as a fully NP model. This mimics a situation where the researcher does not know the true underlying distribution of individual heterogeneity and either assumes a standard parametric model (MNL/RCL) or adopts our approach. For our simulations, we consider the individual coefficient of price to follow a bimodal normal mixture distribution. Specifically, we let η_i to be drawn from a bimodal normal distribution as follows: $\eta \sim \chi\mathcal{N}(-0.1, 0.1) + (1-\chi)\mathcal{N}(-2, 2)$, where χ is Bernoulli distributed with a probability of 0.5.

The results from this exercise are shown in Table 2. We employ 20 random draws and consider a setting that has 10 products, 200 markets, and zero non-price features to keep the focus on capturing the ability of the model to capture the distribution of heterogeneity.

We see that our model performs consistently better than the benchmark models in predicting market shares, own- and cross-elasticities. This suggests that in cases where the researcher does not know the true distribution of individual-level heterogeneity, our flexible approach is better than fully parametric approaches such as MNL/RCL or naive non-parametric approaches such as NP.

I Price Variation and Elasticity Estimates

In our main simulations, prices are drawn from a uniform distribution, which provides high variance and ensures sufficient observations across the full range of possible prices. This broad coverage may benefit the estimation of price elasticity, particularly for non-parametric methods. In this appendix, we present a simulation study with reduced price variance. Specifically, we use the same data-generating process as in Appendix H, but only randomly draw prices for products in 20% markets. The same prices are then repeated for products across the remaining 80% markets. In other words, only 20% markets exhibit unique prices,

Table A6: Predictive Performance in Random Coefficient Logit Models with Bimodal Mixture Distribution: Market Shares, Own-Elasticity, and Cross-Elasticity

(a) Market Shares ($\hat{\pi}_{jt}$)

#	True Model	J	T	d	Our Model		MNL		RCL		NP		MP		No. Obs.
					MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	
0	RCL-Bimodal Mixture	10	200	0	0.0010	0.0297	0.0204	0.1779	0.0056	0.0663	0.0185	0.1141	0.0308	0.1455	400

(b) Own-Elasticity ($\frac{\partial \hat{\pi}_{jt}/\pi_{jt}}{\partial p_{jt}/p_{jt}}$)

#	True Model	J	T	d	Our Model		MNL		RCL		NP		No. Obs
					MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	
0	RCL-Bimodal Mixture	10	200	0	0.0624	0.1014	0.6058	0.8244	0.2310	0.2985	0.3938	0.5702	32000

(c) Cross-Elasticity ($\frac{\partial \hat{\pi}_{jt}/\pi_{jt}}{\partial p_{k\neq jt}/p_{k\neq jt}}$)

#	True Model	J	T	d	Our Model		MNL		RCL		NP		No. Obs
					MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	
0	RCL-Bimodal Mixture	10	200	0	0.0184	0.0396	0.0618	0.1132	0.0309	0.0511	0.1595	0.2332	288000

resulting in substantially lower price variation across the dataset.

We find that while our model's performance in estimating price elasticity is slightly worse under reduced price variation compared to the uniform pricing simulation, it still outperforms other baseline models, as results presented in Table A7.

Table A7: Predictive Performance for Reduced Price Variation in Random Coefficient Logit Models with Bimodal Mixture Distribution: Market Shares, Own-Elasticity, and Cross-Elasticity

(a) Market Shares ($\hat{\pi}_{jt}$)

#	Price Distribution	Our Model		MNL		RCL		NP		MP		No. Obs.
		MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	
0	Uniform Distribution	0.0010	0.0297	0.0204	0.1779	0.0056	0.0663	0.0185	0.1141	0.0308	0.1455	400
1	Reduced variation	0.0008	0.0268	0.0203	0.1778	0.0055	0.0662	0.0011	0.0281	0.0310	0.1465	400

(b) Own-Elasticity ($\frac{\partial \hat{\pi}_{jt}/\pi_{jt}}{\partial p_{jt}/p_{jt}}$)

#	Price Distribution	Our Model		MNL		RCL		NP		No. Obs
		MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	
0	Uniform distribution	0.0624	0.1014	0.6058	0.8244	0.2310	0.2985	0.3938	0.5702	32000
1	Reduced variation	0.0659	0.1070	0.6075	0.8268	0.2283	0.2942	0.6295	0.8725	32000

(c) Cross-Elasticity ($\frac{\partial \hat{\pi}_{jt}/\pi_{jt}}{\partial p_{k\neq jt}/p_{k\neq jt}}$)

#	Price Distribution	Our Model		MNL		RCL		NP		No. Obs
		MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	
0	Uniform distribution	0.0184	0.0396	0.0618	0.1132	0.0309	0.0511	0.1595	0.2332	288000
1	Reduced variation	0.0217	0.0435	0.0619	0.1136	0.0306	0.0506	0.2500	0.3660	288000

Note: Table A7 shows the simulation results from 20 random draws of a setting that has 10 products, 200 markets, and zero non-price features. Specifically, we consider the individual coefficient of price to follow a bimodal normal mixture distribution. We specify η_i to be $\chi\mathcal{N}(-0.1, 0.1) + (1-\chi)\mathcal{N}(-2, 2)$, where χ is Bernoulli distributed with a probability of 0.5. We consider two types of price distribution: 1) Uniform distribution with a range of [0,4]; 2) Reduced variation: we randomly draw prices from $\mathcal{U}[0, 4]$ for products in 20% markets and the same prices are then repeated for products across the remaining 80% markets.

J A Different Training Method of the Standard Non-parametric Estimator

In the simulations shown so far, we train the standard non-parametric model (NP) by treating each market as one observation. This limits the size of the training sample and adversely affects the predictive accuracy of the non-parametric method. We now examine whether the performance of this method can be improved by increasing the sample size (keeping the number of markets fixed) by permuting the order of the products. training sample. Specifically, we create multiple observations per market by permuting the order of the products. For example, for each market that has J products, we create $J!$ observations for each market by exhausting all potential permutations of J products.

In the simulations, we set $J = 5$, $K = 100$, and $M = 100$, and the true DGP is set to RCL. This is the exact same data generation as Row 5 in Table 2. We show the results from this exercise in Table A8. We find that when we train the NP model on the $J! = 120$ permutations of each market, its performance becomes comparable (though still slightly worse) to that of our proposed method, which suggests that it can be a feasible strategy for improving the performance of naive NP methods in small J settings.

While this permutation-based approach is tractable for small J, it quickly becomes infeasible for larger J. The factorial $J!$ grows extremely rapidly. For example:

- $20! \approx 2.43 \times 10^{18}$,
- $100! \approx 9.33 \times 10^{157}$, which is vastly larger than the number of atoms in the observable universe.

Even storing these expanded datasets becomes prohibitive. Assuming each additional observation requires just 1 byte, creating all permutations for just 20 products would already occupy over 2,265,817 GB of storage. The computational costs of training on such an enormous dataset—both in terms of memory and processing time—also grow prohibitively large. In contrast, our proposed approach remains scalable for large product assortment situations.

Table A8: Comparison with NP trained with permutated data

(a) Market Shares ($\hat{\pi}_{jt}$)

#	True Model	J	T	d	Our Model		NP - Permutation		No. Obs.
					MAE	RMSE	MAE	RMSE	
0	RCL	5	100	10	0.0289	0.1511	0.0326	0.1602	100

(b) Own-Elasticity ($\frac{\partial \hat{\pi}_{jt}/\pi_{jt}}{\partial p_{jt}/p_{jt}}$)

#	True Model	J	T	d	Our Model		NP - Permutation		No. Obs.
					MAE	RMSE	MAE	RMSE	
1	RCL	5	100	10	0.1076	0.2587	0.1155	0.2951	8000

(c) Cross-Elasticity ($\frac{\partial \hat{\pi}_{jt}/\pi_{jt}}{\partial p_{k\neq jt}/p_{k\neq jt}}$)

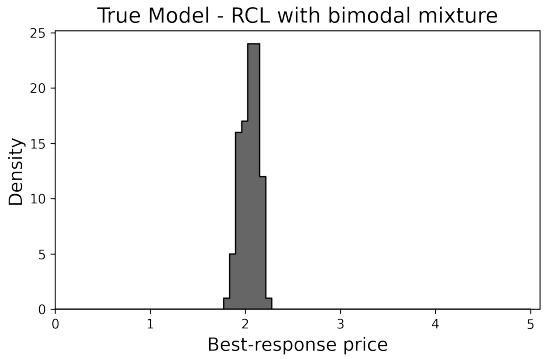
#	True Model	J	T	d	Our Model		NP - Permutation		No. Obs.
					MAE	RMSE	MAE	RMSE	
0	RCL	5	100	10	0.0337	0.0742	0.0435	0.0950	32000

K Simulations to Recover Best-Response Prices

We now present a simulation that demonstrates how our model can recover the best response price for a given product, conditional on its other characteristics and the prices and characteristics of competing products. We simulate 100 markets, and each market has 10 products with one non-price characteristic.

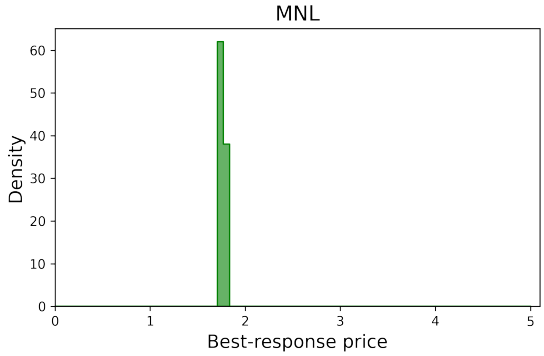
We focus on the pricing problem of a firm that offers one product. In the data generation process, we fix the characteristics of the product the same across markets while the price in each market is randomly drawn from a uniform distribution $\mathcal{U}[0, 4]$. In addition, the characteristics and prices of competing products

also vary across markets. We use a random coefficient logit model with a bimodal mixture distribution of random coefficients (same model parameters as the data generation process in Web Appendix H) as the true model to demonstrate how our models perform against other mis-specified parametric models (MNL and RCL).

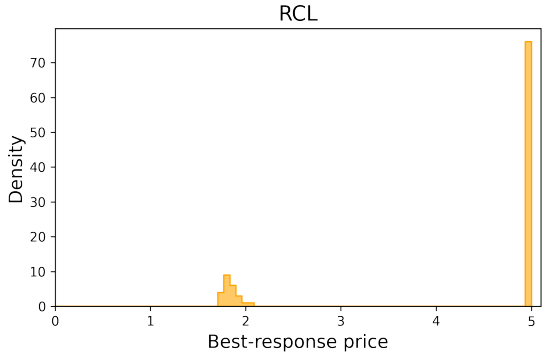


(a) True - RCL with bimodal mixture

(b) Our Model



(c) MNL



(d) RCL

Figure A1: Distributions of the best-response price

Let the focal product be j in every market, and assume that it has a fixed marginal cost of c. Then, the firm's profit in t^{th} market is given by:

$$(p_{jt} - c) \cdot \hat{\pi}_{jt}, \tag{A35}$$

where $\hat{\pi}_{jt}$ is the predicted market share of product j in market t

$$\hat{\pi}_{jt} = \hat{\pi}(p_{jt}, X_{jt}, \{p_{kt}, X_{kt}\}_{k \neq j, k \in S_t}). \tag{A36}$$

Then, the firm's best-response price in market t is defined by the solution to the profit maximization problem given the competitors' prices, characteristics, and its own product characteristics

$$p^*_{jt} = \arg\max_{p_{jt}} \quad \hat{\pi}_{jt} \cdot (p_{jt} - c). \tag{A37}$$

Note that p^*_{jt} varies across markets because the competitors' prices and characteristics vary across markets.

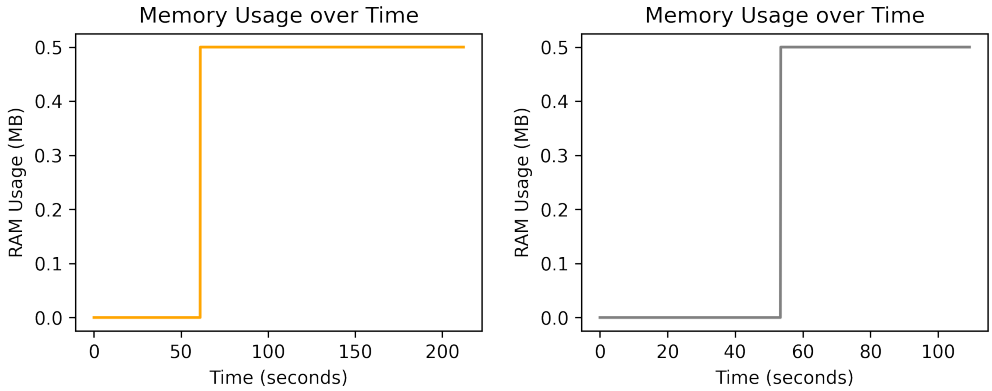
Then, the F.O.C of the profit maximization problem gives us the optimal price as:

$$p_{jt}^{*} = c + \frac{\hat{\pi}_{jt}}{\frac{\partial \hat{\pi}_{jt}}{\partial p_{jt}}} \tag{A38}$$

To efficiently solve this problem and find the best-response price, we use a grid search method[2] that exhaustively evaluates the profit function (Equation (A35)) over possible prices. We further set the marginal cost to $c = 1$.

The distribution of best-response prices across all markets, as recovered by each model, is presented in Figure A1. We see that our model aligns closely with the true best-response prices. In contrast, the distribution of best-response prices recovered from both the RCL and MNL models are quite different from the true distribution of best-response prices.

L Computational Efficiency Comparison



(a) BLP Model (b) Our Model

Figure A2: Memory Footprint

	Estimation Time
Our model	109.32 seconds
BLP model	212.17 seconds

Table A9: Comparison of estimation time

We now conduct an experiment to systematically compare the computational performance of our proposed method against the standard BLP approach. Specifically, we re-estimate the automobile dataset from Section 6 using two methods: (i) our proposed approach with debiased procedures and (ii) the BLP model with the BFGS optimization routine implemented in `pyblp`. The results, summarized in Table A9 and Figure A2, indicate that our method runs approximately two times faster while exhibiting a similar memory footprint compared to the standard BLP approach. Our analysis was run on a Linux-based system, equipped with a 64-thread processor (32 physical cores) and 754.54 GB of total RAM. We used pytorch to implement our approach, and its code was run on an NVIDIA RTX A6000 GPU.

[2]We search over the interval $[0, 5]$ with a 0.01 step size.

M 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, $\mathcal{D}$, into three separate partitions of markets, each denoted as D_l. Each market is exclusively assigned to only one partition. For each partition, we define its complement set, D_l^c, as the subset of data in $\mathcal{D}$ that is not included in D_l.
- **Step 2: Cross-fitting** For each partition l, we first estimate a linear regression model on the complement data set, D_l^c, using the Lasso method with hyperparameters tuned by 3-fold cross-validation. As discussed in section 3.4, we need the estimator of γ to converge atleast at $n^{-1/4}$ rate, a similar result that bounds the 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 D_l. We denote the fitted value as $\hat{f}_l$, which is essentially the MLIV.
- **Step 3: First-stage Regression** We estimate the first-stage estimator γ and residual μ using the MLIV as the only predictor following step 1 in Section 4.

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 A3. In Figure A3b, 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 A3c, 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 A4.

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 A10. 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 A10: Weak Instrument Test

References

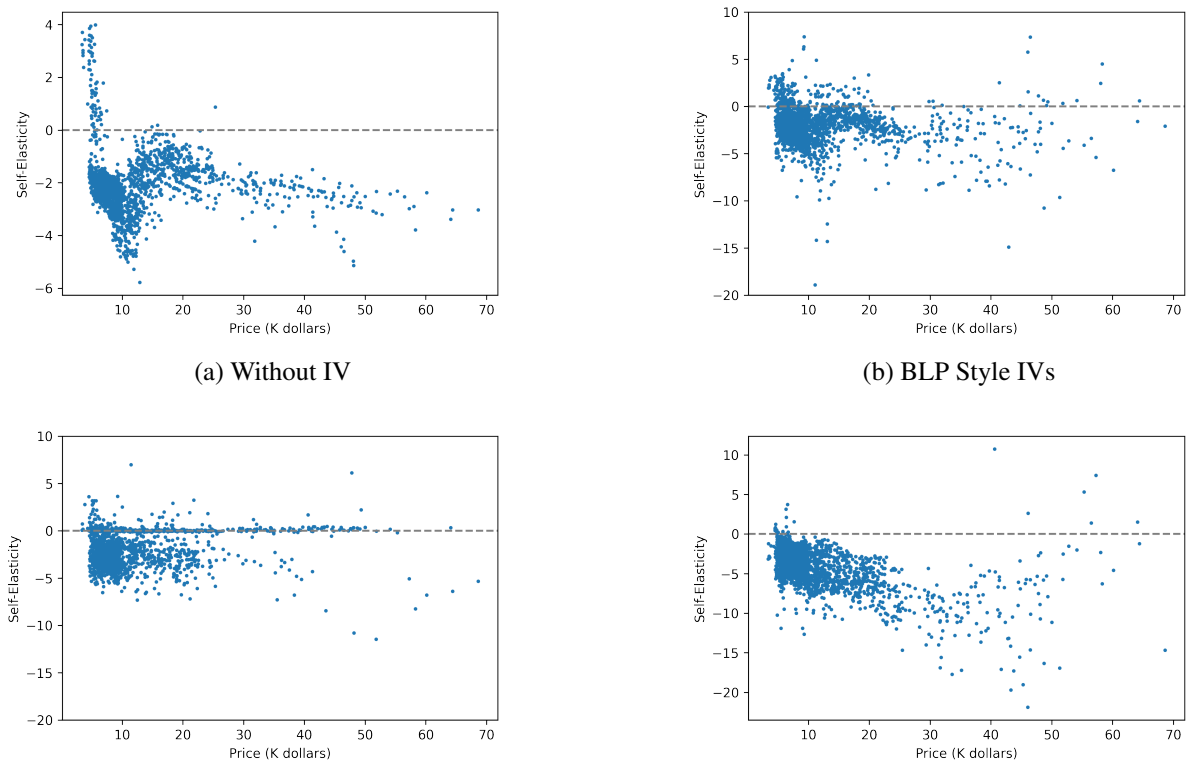
J. Blanchet, G. Gallego, and V. Goyal. A markov chain approximation to choice modeling. *Operations Research*, 64(4):886–905, 2016.

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 riesz 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, 2022.

P. K. Chintagunta. Endogeneity and heterogeneity in a probit demand model: Estimation using aggregate data. *Marketing Science*, 20(4):442–456, 2001.



(a) Without IV

(b) BLP Style IVs

(c) Differentiation IVs

(d) MLIV

Figure A3: Elasticity Estimation Comparison

Figure A3 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).

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

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

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

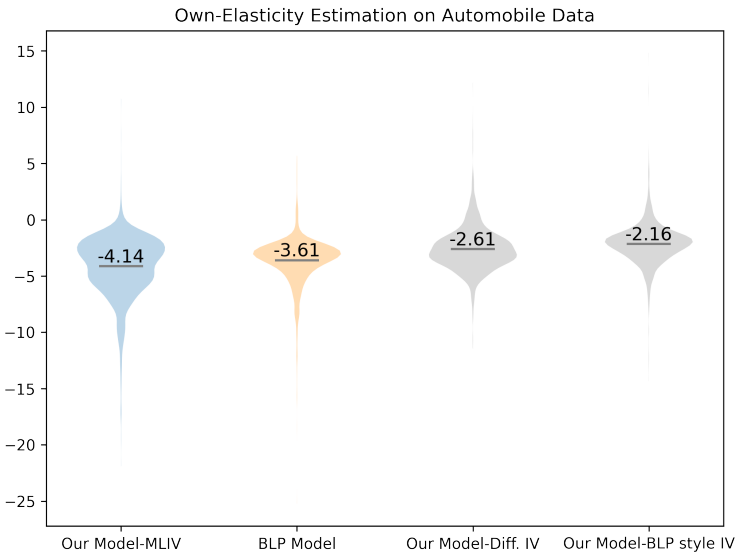
M. S. Goeree. Limited information and advertising in the us personal computer industry. *Econometrica*, 76 (5):1017–1074, 2008.

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.

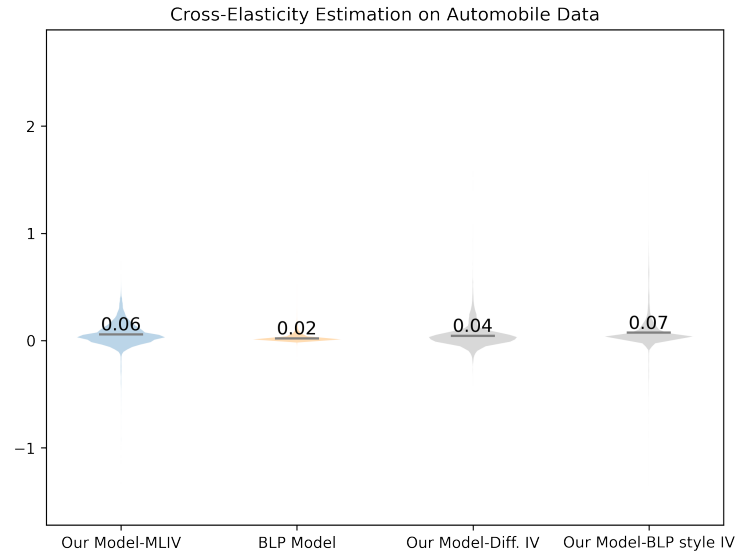
J. Joo. Rational inattention as an empirical framework for discrete choice and consumer-welfare evaluation. *Journal of Marketing Research*, 60(2):278–298, 2023.

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.

A. Singh, K. Hosanagar, and A. Gandhi. Machine learning instrument variables for causal inference. In



(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 (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.

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.

S. Turlo, M. Fina, J. Kasinger, A. Laghaie, and T. Otter. Discrete choice in marketing through the lens of rational inattention. 2023.

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