Data-Driven Nests*

Milena Almagro†, Kenneth H. Lai‡, Elena Manresa§

November 2026

Abstract

Nested demand models allow for flexible substitution patterns while preserving tractability. However, assuming a known nesting structure can be problematic, as misspecification may lead to biased estimates. We propose a two-step estimation strategy that first recovers the nesting structure, and then structural parameters using market-level data on product choices and characteristics. In the first step, we employ the grouped fixed-effects estimator of Bonhomme and Manresa (2015) to identify the nesting structure. In the second step, we estimate preference parameters conditional on the estimated nests. Our framework accommodates longitudinal observations of products across multiple markets and allows for endogenous prices. Monte Carlo simulations demonstrate the accuracy of our approach. We apply the method to the U.S. beer market using NielsenIQ data, identifying four distinct beer groups and estimating a demand system with rich substitution patterns. We validate our estimates out of sample using a demand shock following the Bud Light boycott during 2023. Finally, we benchmark our model against alternative demand specifications, including a naively grouped nested logit and a mixed logit, and show that it more accurately captures substitution patterns in our out-of-sample validation.

JEL classification: C23, C25, C26, C51, C52, D12

Keywords: Demand estimation, nested demand, nested logit, grouped-patterns of heterogeneity.

* We thank audiences at NYU, Cornell, Microsoft, Berkeley, UCL, Caltech, ML Workshop Groningen, MSU Advances in High-Dimensional Inference Workshop, Midwest Econometrics Workshop, IO Montreal Summer Conference, Exeter, EARIE 2025, Chicago Booth, Minnesota, FDIC, FGV, UCLA, and MIT, Princeton, Virginia, BSE Forum, Boston College, SDLS conference, UPenn, IAAE Conference, and Cornell for helpful comments. We are grateful to JF Houde for his valuable discussion. We also thank Rodrigo Adão, Giovanni Compiani, Chris Conlon, JP Dubé, Jacob Leshno, Chris Rauh, and Francesca Molinari for their feedback and suggestions. Marek Bojko, Junbiao Chen, Juan Felipe Herrera, and Sriram Tolety provided outstanding research assistantship. Disclaimer: Researcher(s)' own analyses calculated (or derived) based in part on data from Nielsen Consumer LLC and marketing databases provided through the NielsenIQ Datasets at the Kilts Center for Marketing Data Center at The University of Chicago Booth School of Business. The conclusions drawn from the NielsenIQ data are those of the researcher(s) and do not reflect the views of NielsenIQ. NielsenIQ is not responsible for, had no role in, and was not involved in analyzing and preparing the results reported herein.

† University of Chicago, Booth and NBER, email: milena.almagro@chicagobooth.edu

‡ Ludwig-Maximilians-Universität Munich, email: H.Lai1@lmu.de

§ Princeton University, email: em7118@princeton.edu

1 Introduction

Discrete choice models of random utility are widely used to model and quantify demand across a broad range of applications. A common approach specifies demand according to a nesting structure, which can be interpreted as consumers making sequential choices: first of a category of products, and then, on a product itself. This model was first put forward by (McFadden, 1978). The appeal of nested demand models lies in their ability to remain parsimonious—preserving closed-form expressions of demand—while accommodating flexible substitution patterns. A well-known limitation, however, is that the nesting structure must be specified prior to estimation, which is often not obvious to the researcher. Misspecification of product groups can lead to unreliable parameter estimates (Kovach and Tserenjigmid, 2022; Fogel et al., 2025).

In this paper, we propose a method to estimate the nesting structure and the utility parameters of a nested demand model. The identification strategy exploits the log-linear structure of market shares implied by the Nested Logit (NL), under which products belonging to the same nest share a common model-implied demand curve. When indirect utility is linear, log shares are also linear with nest-specific slopes on characteristics and nest-market intercepts. We then identify the nests by searching over partitions of products that best fit the data. We accommodate endogenous prices by projecting them on exogenous price shifters interacted with group-specific variables, building on the framework of Choi and Okui (2024).

Our proposed method consists of two steps. The first uses the grouped fixed effect estimator of Bonhomme and Manresa (2015) on demand curves implied by the NL model to recover the nesting structure among products. In the second step, using this estimated nesting structure as if it was the true one, we estimate preference parameters by GMM on the inversion of Berry (1994).

We establish two main results. First, we show that the classification step consistently recovers the true nesting structure as the number of products and markets grows. Specifically, under suitable regularity conditions, the probability of misclassifying at least one product converges to zero faster than any polynomial rate. Our proof builds on the ideas of Bonhomme and Manresa (2015), extending their framework to accommodate group-market intercepts that are drifting random variables rather than fixed parameters. A key element of

our analysis is an appropriate re-centering of the random intercept as the number of products grows, establishing conditions of separation in the limit, and not for every J.

Second, we characterize the asymptotic distribution of the estimator of the utility parameters. We show that the second step estimator has the same asymptotic distribution as the infeasible estimator that applies the Berry inversion using the true nesting structure. In doing so, we establish the asymptotic properties of the Berry inversion estimator when both the number of products and markets grow large, a result of independent interest. Notably, thank to our asymptotic equivalence result, inference on utility parameters is conducted ignoring any error that may arise in the first step. We derive sufficient conditions on market structure, exploiting the presence of multiproduct firms, under which price elasticities remain identified, hence avoiding the challenge in identification of demand with many products first raised by Armstrong (2016).

Our asymptotic framework is particularly relevant given the increasing availability of large-scale market data, such as retail scanner data (Hitsch et al., 2021; Atalay et al., 2026; Döpper et al., 2025), housing and labor markets (Bayer et al., 2007; Ahlfeldt et al., 2015; Dingel and Tintelnot, 2026), and trade with a large number of countries and firms (Eaton and Kortum, 2002; Bernard et al., 2019; Dhyne et al., 2021). Moreover, the large-number-of-products asymptotic framework provides a tractable way for establishing the theoretical properties of the proposed estimator, while its practical applicability extends to settings with substantially fewer products. Indeed, our Monte Carlo simulations demonstrate good performance even with as few as 100 products and 10 markets.

We also show how to estimate the number of groups by drawing on clustering heuristics from the machine learning literature. Specifically, we apply the Elbow method within a cross-validation exercise to determine the optimal number of partitions.

We consider several extensions to the baseline model. First, we extend the baseline to a setting without exogenous price shifters, where prices are exogenous conditional on product fixed effects; in this case, endogeneity is handled by demeaning within product across markets in both steps. Second, we incorporate consumer heterogeneity in a setting where both consumer characteristics and individual choices are observed. Third, we show how the algorithm extends to a model with several layers of nesting. Finally, we show that our method extends to models where idiosyncratic shocks follow a Type II Extreme Value (Fréchet) distribution leading to nested CES demand, which is widely used in trade and spatial eco-

nomics applications.

We conduct a Monte Carlo exercise to assess the ability of our method to recover both the true nesting structure and the underlying preference parameters. The experiment considers products observed across multiple markets, with one exogenous characteristic and one endogenous variable — prices. We find that with as few as ten markets, our method correctly assigns approximately 99% of products to their true groups. This match rate increases monotonically with the number of markets, reaching 100% in larger samples. We also validate our heuristic approach to selecting the number of groups.

We apply our estimator to demand for beer across U.S. states using store-level panel data from NielsenIQ. We first use the method with a simplified indirect utility, with continuous covariates only—prices, units, and liquid content. We recover four well-differentiated groups, which we label *Regular Lager*, *Light Beer*, *Special/Craft Beer*, and *Boutique Craft*. These groups exhibit substantial heterogeneity in product characteristics, including prices, beer type, packaging volume and size, and the likelihood of being a domestic product or a top-ten national brand. However, the estimated nesting structure is far from fully explained by observable product characteristics, suggesting that unobserved heterogeneity plays an important role in shaping substitution patterns. Notably, using store-level geographic information, we document substantial geographic variation in the relative popularity of different beer nests: *Light Beer* is disproportionately consumed in the South, while *Boutique Craft* is more prevalent in the Northeast. We also find that these consumption patterns are correlated with local political partisanship. Importantly, neither geographic location nor political affiliation is used in the classification step, highlighting the ability of our approach to uncover economically meaningful patterns of heterogeneity without relying on these observable characteristics.

We also perform an out-of-sample validation exercise to assess the substitution patterns implied by the groups recovered in the first step. The exercise exploits the Bud Light boycott as an exogenous negative demand shock. Using a difference-in-differences design, we define the treatment group as all *Light Beer* products other than Bud Light — i.e., products belonging to the same estimated nest — and the control group as products from all remaining groups. Following the boycott, *Light Beer* products experienced significantly larger sales increases in 2023 relative to controls. These reduced-form results suggest that consumers perceive *Light Beer* products as closer substitutes to Bud Light than products from other

groups, that we take as suggestive evidence that our estimated nesting structure provides reasonable substitution patterns. We confirm the robustness of these findings through several exercises, including group-specific treatment effect estimates and a placebo test using randomly assigned groups.

When we move to the second step estimation of structural parameters, we find an average elasticity of -3.561, similar to previous work studying beer ranging from -6.1 to -3.3 (Pinkse and Slade, 2004; Slade, 2004; Romeo, 2016; Miller and Weinberg, 2017). Moreover, we also find that all the nesting parameters are consistent with utility maximization and lie between 0 and 1 without the need to impose any additional restrictions in the structural estimation. We estimate a high degree of heterogeneity in substitution patterns across different groups, ranging from $\sigma = 0.042$ for *Boutique Craft* to $\sigma = 0.807$ for *Light Beer*. In terms of own-price elasticities, we find that *Light Beer* is the most inelastic group and *Boutique Craft* the most elastic. On the other hand, we also find that *Boutique Craft* has the largest within-group substitution effects following a price shock.

We also perform several robustness checks. First, we use an alternative price instrument, finding that the optimal number of groups remains four and that the average characteristics within each group are similar to those in the baseline. Second, we examine the grouping structure for three and five groups, describing a hierarchical pattern: with three groups, *Regular Lager* and *Special/Craft Beer* merge into a single group, while with five groups, *Regular Lager* splits into two. Third, augmenting the covariate set with categorical information on beer type—one of the variables provided by NielsenIQ—yields one fewer group but marginal improvements in in-sample predictive power.

We finally compare the results of our baseline model to two common approaches to demand estimation. First, we estimate second-step parameters using the "naive" classification that leverages the categorical variable "Beer Type" provided by NielsenIQ as the nesting structure. Second, we also estimate a random coefficient model—the most flexible model of random utility. When doing so, we find that the Data-Driven NL outperforms these two alternative approaches in the out-of-sample validation test in the ability to predict substitution patterns. When looking at the results of the structural estimation, we find that some of the parameters of the Naive NL are not consistent with utility maximization and that the substitution patterns implied by the mixed logit resembled those of a logit model.

Literature review. This paper contributes to the literature on discrete choice demand models by proposing a data-driven approach to estimate the nesting structure of the Nested Logit (NL) model. Introduced by McFadden (1978), the NL model relaxes the IIA restriction of the standard logit by allowing correlated unobserved utility within groups of alternatives. It is best understood as an intermediate case between the standard logit and the Random Coefficients (RC) model of Berry et al. (1995): it retains computational tractability while allowing for within-group preference correlation, and can be interpreted as a particular RC specification in which a random coefficient loads on group dummies (Cardell, 1997).[1] The NL is a special case of the RC logit, in which random coefficients with a particular distribution multiply nest dummies (Berry, 1994; Cardell, 1997). Our model does not try to approximate the distribution of the random coefficients. By contrast, its goal is to uncover the dummies, that is, which products share a nest. Importantly, the special-case result requires the researcher to know these dummies. Without this knowledge, the nested logit is no longer a special case of a random coefficients model on the observed characteristics.

Thanks to their tractability, nested demand models have been widely applied in industrial organization (Verboven, 1996; Parry and Small, 2005; Einav, 2007; Grennan, 2013; Ciliberto and Williams, 2014; Berry et al., 2016; Björnerstedt and Verboven, 2016; Allcott et al., 2024), international trade (Goldberg, 1995; Broda and Weinstein, 2006; Verhoogen, 2008; Atkeson and Burstein, 2008; Khandelwal, 2010; Fajgelbaum et al., 2011), urban and spatial economics (Davis et al., 2019; Couture et al., 2024; Bordeu, 2026; Couture et al., 2025), and labor economics (Azar et al., 2022; Lamadon et al., 2022; Berger et al., 2022; Deb et al., 2024). Crucially, however, all of these applications *assume the nesting structure is known and specified a priori*. In many empirical contexts the appropriate grouping of products is far from obvious. Our paper departs from this assumption by treating the nesting structure as an object to be estimated from the data rather than being imposed by the researcher. In particular, our identification argument does not incorporate additional significant assumptions beyond the structure of the model.

This paper addresses this limitation by estimating the nesting structure from the data. Along these lines, Hortacsu et al. (2023) is the closest paper to ours. They propose a method that exploits non-negativity and sparsity constraints to estimate the nesting structure of an Inverse Generalized Nested Logit model (Fosgerau et al., 2024) using aggregate data on

[1] More generally, any discrete choice model can be approximated by an RC model (McFadden and Train, 2000).

market shares, prices, and product characteristics. Their approach allows for overlapping nests, which can generate flexible substitution patterns, but it is designed for settings with a small number of products and an asymptotic sequence in the number of markets. In contrast, we study a panel data setting with many products and focus on recovering non-overlapping nests as the product space grows large. A related contribution is Kovach and Tserenjigmid (2022), who provide a revealed-preference characterization of nested logit and show how the nesting structure can be recovered from individual-level stochastic choice data. Our paper instead focuses on estimating the nesting structure in empirical demand models using market-level data, where researchers observe shares, prices, and product characteristics and the number of products can be large.

This paper also contributes to the econometric theory literature of demand estimation. In particular, our results extends the analysis of the NL Berry inversion estimator (Berry, 1994) with many products and many markets. Two non-standard issues arise when both dimensions grow. First, shares drift to zero as the number of products grows. The outcome, the log share relative to the outside option, and the log within-nest share both drift to $-\infty$ in proportion to $\log J$. Our levels estimator exploits this drift, and the nest parameters converge at the faster rate $\log J\sqrt{JM}$. Second, identification may fail under Bertrand competition with single-product firms (Armstrong, 2016), and we provide conditions under which instruments based on characteristics of other products are still relevant when firms are multi-product with non-vanishing shares. In addition, our paper is related to Lu et al. (2023), who also exploit large-product-market asymptotics, in the tradition of Berry et al. (2004), to estimate demand models with market-specific nuisance effects that depend on unobserved demand shocks. However, the asymptotic arguments differ. Lu et al. (2023) allow these effects to remain random and establish consistency through uniform identification and convergence conditions. By contrast, we show that the nest-market effects concentrate around nonrandom quantities as the number of products grows, allowing us to treat their stochastic components as negligible perturbations when establishing consistent classification.

This paper also contributes to the literature on grouped fixed-effect estimators (Hahn and Moon, 2010; Bonhomme and Manresa, 2015). Specifically, we adapt the framework proposed by Bonhomme and Manresa (2015) that allows for group-specific time patterns and individual group membership in panel data settings. Grouped patterns of heterogeneity arise naturally as a structural feature of the nested logit model, and we re-interpret the problem

of finding unknown groups/nests as uncovering the structure of unobserved heterogeneity. We leverage Choi and Okui (2024) to accommodate endogenous variables such as prices. We also extend the theoretical analysis of Bonhomme and Manresa (2015) to the nested logit model, which has a non-standard feature: the group-market intercepts of the regression are stochastic and drift to $-\infty$. After re-centering, they concentrate around non-random objects as the number of products grows, and we treat the deviation as a perturbation that vanishes uniformly.[2]

Finally, our empirical application focuses on the beer industry, which has been a frequent setting for structural demand estimation. Previous work has used brand-level demand systems, distance-metric models, nested logit models and random-coefficient logit models, with average own-price elasticities between -3.3 and -6.1 (Hausman et al., 1994; Pinkse and Slade, 2004; Slade, 2004; Hellerstein, 2008; Rojas, 2008; Goldberg and Hellerstein, 2013; Asker, 2016; Romeo, 2016; Miller and Weinberg, 2017; Miller et al., 2021). The choice of nesting structure is a recognized difficulty in this market. Slade (2004) fixes nests by product type and Miller and Weinberg (2017) use a single nest for all inside goods. Hellerstein (2008) sets the nested logit aside because, in her words, it is difficult to define clear nests or stages in beer consumption: a consumer who buys an imported light beer may classify beers first by origin and then by type, or the reverse. We depart from this literature in three ways. First, our estimator handles a large number of products, so we estimate demand at the UPC level, with close to three thousand products and no aggregation. Second, we let the data determine the grouping of products instead of fixing it in advance. Third, because the nests are estimated rather than imposed, they can group products along dimensions the researcher does not observe, as long as those dimensions can be described by a nesting structure. In our application, the recovered nests reproduce the light-beer segment that the literature imposes by hand. They also capture a dimension that fixed nests cannot, since the nest-by-market intercepts absorb differences in nest popularity across markets, such as a stronger taste for light beer in Republican states.

[2] The two-step approach mimics Bonhomme et al. (2022). By contrast, the groups in this paper are structural objects so the classification does not intend to approximate the underlying heterogeneity but rather recover it.

2 The Data-Driven Nested-Logit Model

We consider a model with J products and M independent markets representing different geographies or time periods.[3] Agent i in market m receives utility from product j given by:

$$U_{ijm} = \delta_{jm} + \epsilon_{ijm} \quad j \in \{0, 1, ..., J\}, m \in \{1, ..., M\},$$

where δ_{jm} is the average utility for choice j in market m and ϵ_{ijm} is an idiosyncratic taste shock. The probability that agent i chooses product j is given by the probability that product j maximizes its indirect utility:

$$\begin{aligned}\mathbb{P}_{jm} &= \mathbb{P}(U_{ijm} > U_{ij'm} \, \forall j' \neq j) \\ &= \mathbb{P}(\delta_{jm} - \delta_{j'm} > \epsilon_{ij'm} - \epsilon_{ijm} \, \forall j' \neq j).\end{aligned}$$

Products are grouped into K nests to capture heterogeneity in substitution patterns. Taste shocks $(\epsilon_{i1m}, ..., \epsilon_{iJm})$ have a joint distribution that depends on the partition and parameters $\sigma_k \in (0, 1]$, that capture the different degrees of substitution across nests, for $k = 1, \ldots, K$.[4] Denoting $k(j) \in \{1, \ldots, K\}$ the nest where product j belongs to, the probability of choosing product j is given by:

$$\mathbb{P}_{jm} = \frac{e^{\frac{\delta_{jm}}{\sigma_{k(j)}}} \left(\sum_{\{j'|k(j')=k(j)\}} e^{\frac{\delta_{j'm}}{\sigma_{k(j)}}} \right)^{\sigma_{k(j)}-1}}{1 + \sum_{l=1}^{K} \left(\sum_{\{j'|k(j')=l\}} e^{\frac{\delta_{j'm}}{\sigma_l}} \right)^{\sigma_l}}, \tag{1}$$

and the outside option has share $\mathbb{P}_{0m} = 1/\left(1 + \sum_{l=1}^{K} (\sum_{\{j'|k(j')=l\}} e^{\delta_{j'm}/\sigma_l})^{\sigma_l}\right)$. Implied price

[3] While we present our model with a balanced panel for simplicity, the method extends to unbalanced panels in which products appear in different subsets of markets, as long as the number of product-market observations grows to infinity.

[4] In Section 6.4, we extend our model to the case with multiple layers of sequential nests.

elasticities are:

$$\mathcal{E}^{j}_{j',m} = \begin{cases} \dfrac{1}{\sigma_{k(j)}}\Big(1 - \sigma_{k(j)}\,\mathbb{P}_{jm} - (1-\sigma_{k(j)})\,\mathbb{P}_{j|k(j),m}\Big)\dfrac{\partial \delta_{jm}}{\partial p_{jm}} p_{jm} & \text{if } j' = j, \\ -\,\mathbb{P}_{j'm}\,\dfrac{\partial \delta_{j'm}}{\partial p_{j'm}}\,p_{j'm} & \text{if } k(j') \neq k(j), \\ \Big((\sigma_{k(j)} - 1)\,\dfrac{\mathbb{P}_{j'|k(j),m}}{\sigma_{k(j)}} - \mathbb{P}_{j'm}\Big)\dfrac{\partial \delta_{j'm}}{\partial p_{j'm}}\,p_{j'm} & \text{if } k(j') = k(j), \end{cases} \tag{2}$$

where $\mathbb{P}_{j|k(j),m} \equiv \mathbb{P}_{jm} / \sum_{\{j'|k(j')=k(j)\}} \mathbb{P}_{j'm}$ is the within-nest share. When $\sigma_k \to 0$, products become more elastic and there is more substitution within nests. By contrast, when $\sigma_k = 1$, we are back to the standard logit model. It immediately follows that $\mathcal{E}^{j}_{j',m} \geq \mathcal{E}^{j}_{j'',m}$ if with $k(j) = k(j')$ and $k(j) \neq k(j'')$: products within the same nest are closer substitutes than products across nests.[5]

The nested logit model was first introduced by McFadden (1978). Empirical applications commonly assume that the product partition, $\{k(j)\}_{j=1}^{J}$, is known to the researcher, yet often this assumption is debatable. Here, we relax this assumption and the allow the partition to be unknown and it will be estimated with the rest of the parameters of the model.

Nest misspecification biases preference estimates and σ_k (Kovach and Tserenjigmid, 2022; Fogel et al., 2025). In Appendix A.2, we show a simplified example in which we give intuition on the nature of this bias.

2.1 Empirical Model: Two Linear Equations

Taking logs in (1) yields the empirical structural equation of demand with nested logit:

$$\log \mathbb{P}_{jm} = \frac{\delta_{jm}}{\sigma_{k(j)}} + \underbrace{(\sigma_{k(j)} - 1)\log\Big(\sum_{\{j'|k(j')=k(j)\}} e^{\frac{\delta_{j'm}}{\sigma_{k(j)}}}\Big) - \log\Big(1 + \sum_{l=1}^{K}\Big(\sum_{\{j'|k(j')=l\}} e^{\frac{\delta_{j'm}}{\sigma_l}}\Big)^{\sigma_l}\Big)}_{:=\lambda_{k(j)m}}. \tag{3}$$

The Berry inversion (Berry, 1994) is also a useful equation to estimate the structural parameters of a Nested-Logit model:

$$\log \mathbb{P}_{jm} - \log \mathbb{P}_{0m} = \beta_p p_{jm} + \beta_x' \mathbf{x}_{jm} + (1 - \sigma_{k(j)}) \log \mathbb{P}_{j|k(j),m} + \xi_{jm}. \tag{4}$$

[5] For a full derivation of the substitution patterns see Appendix A.1.

We now proceed to describe the different components of our model.

2.1.1 Utility

We assume that average utility component δ_{jm} is linear and given by

$$\delta_{jm} = \beta_p p_{jm} + \beta_x' \mathbf{x}_{jm} + \xi_{jm}, \tag{5}$$

where p_{jm} denotes the price variable that enters utility, $\mathbf{x}_{jm}$ is a vector of exogenous product characteristics that may include higher-order terms and interactions, and ξ_{jm} represents unobserved demand shocks. We assume that $\mathbf{x}_{jm}$ is orthogonal to ξ_{jm}, while prices can be potentially correlated with ξ_{jm}, hence endogenous.

Substituting the expression for δ_{jm} in (3) yields:

$$\log \mathbb{P}_{jm} = \beta_{k(j),p} p_{jm} + \beta_{k(j),x}' \mathbf{x}_{jm} + \lambda_{k(j)m} + \upsilon_{jm}, \tag{6}$$

where $\beta_{k,p} \equiv \beta_p/\sigma_k$, $\beta_{k,x} \equiv \beta_x/\sigma_k$ and $\upsilon_{jm} \equiv \xi_{jm}/\sigma_{k(j)}$.

2.1.2 Prices

We start by presenting the structural equilibrium pricing equation, which takes the general form:

$$p_{jm} = m_j(z_{jm}, \mathbf{x}_{jm}, \upsilon_{jm}), \tag{7}$$

where $m(\cdot)$ is an unknown function that may be nonlinear in its arguments, υ_{jm} is an unobserved price component potentially correlated with ξ_{jm}, and z_{jm} is an exogenous price shifter such that:

$$\mathbb{E}[\xi_{jm}|\mathbf{x}_{jm}, z_{jm}] = 0.$$

In order to deal with endogeneity of prices, we consider a linear projection on exogenous covariates with nest-specific coefficients and nest-market intercepts:

$$p_{jm} = \gamma_x^{g(j)\prime} \mathbf{x}_{jm} + \gamma_z^{g(j)\prime} z_{jm} + \kappa_{g(j)m} + \omega_{jm}, \tag{8}$$

where ω_{jm} is a projection error, orthogonal to $\mathbf{x}_{jm}$, $\mathbf{z}_{jm}$, and $\kappa_{g(j)m}$ by construction.

Pricing projections are standard ways of dealing with endogeneity of prices in linear models. The above pricing equation projects on all exogenous covariates, x_{jm} and the price shifter, z_{jm}, with group-specific interactions, as well as group-market intercepts. In two-step estimators, the first-step projection has to include every exogenous regressor of the structural equation. The demand equation (6) has group-specific slopes and group-market intercepts, so the projection needs them too. If they are omitted, the fitted prices retain group-specific variation that is correlated with the demand intercepts, and the second step is inconsistent. Given that groups are unknown, we consider the most general case in which the groups in prices do not coincide with the ones of the nesting structure.

A projection with group-specific coefficients arises in standard pricing models. Under Bertrand or Cournot competition with nested logit demand, markups inherit the nesting structure of demand. In this case the price equation has the same grouping as demand, $g(j) = k(j)$. A grouping can also come from the cost side. When a firm owns many products that share a common production line, marginal costs have common components across the products of that firm. The price equation then has a grouping by firm, which need not coincide with the nests.

3 Identification of Nests: An Intuition

The novelty in this paper is the possibility to learn the nesting structure from the data. Under δ_{jm} as in in equation (5), log shares are linear functions of prices and characteristics, with group-specific intercept, λ_{km}, and group-specific slopes, β_k. The main insight from equation (3) is that *a product's log share, net of the indirect utility component, does not vary across products that belong to the same nest*.

The expression in (6) implies that products belonging to the same nest share a common demand curve. When σ_k is closer to 0, the group-specific slope increases and log-shares become more responsive to changes in characteristics. This insight is in line with the expression in (2), where own-price elasticities increase as σ_k decreases. In fact, under a linear indirect utility specification, own-price elasticities are proportional to $\beta_k = \beta/\sigma_k$.[6]

[6] Equivalently, the elasticities can be written in terms of the slope and the intercept of the log-share equation a follows:

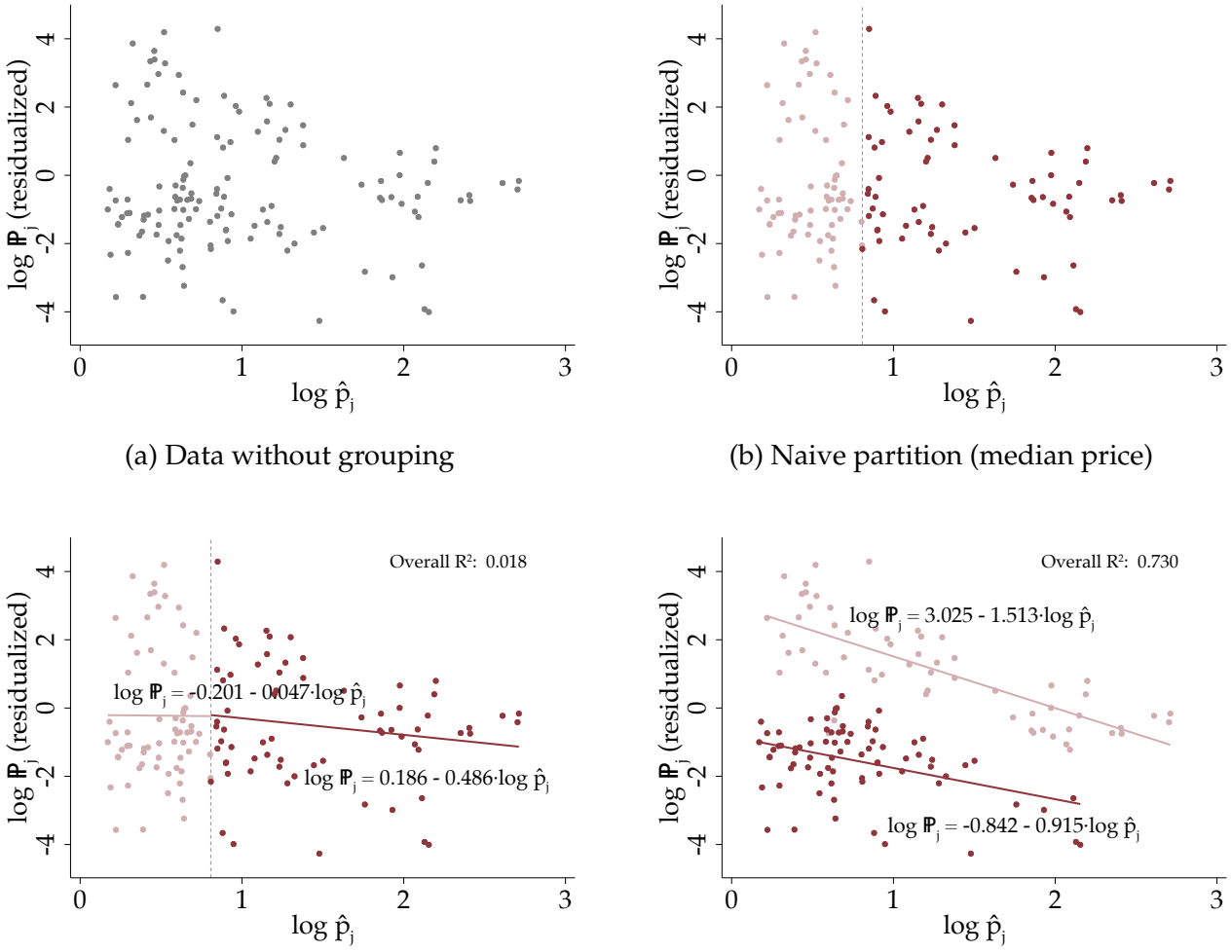
$$\mathcal{E}^{j}_{j',m} = \left(\beta_{p,k(j)}\ \mathbb{1}\{j' = j\} + \frac{\partial \lambda_{k(j)m}}{\partial p_{j'm}}\right) p_{j'm}.$$

λ_{km}, is a function of the inclusive value of the nest of product j, which can be understood as the baseline attractiveness of nest k in market m. Its influence on log shares is again mediated through σ_k, with σ_k closer to 0 implying larger co-movements of levels of log-shares across products in the same group.

Being able to recover groups requires that the pairs (β_k, λ_{km}) are different across k. Differences in σ_k implies differences in β_k. However, in the case they are equal, groups can be separated through intercepts λ_{km}. Note that λ_{km} depends on the prices and characteristics of the products in nest k through the inclusive value, so nests with different distributions of covariates have different intercepts in at least some markets. This would happen if some nests may group larger or more expensive products.

Figure 1 illustrates the intuition behind the identification for the case $K = 2$. We use the data from our empirical application provided by NielsenIQ (see Section 8.1 for more details). The top-left panel shows a scatter plot with predicted log prices (purged of endogeneity) in the x-axis, and log shares (residualized) on the y-axis, and each point corresponds to a product. The identification of the nests amounts to find a partition of this cloud of points that can best explain the data with a group-specific linear regression. In the top-right panel, we show a potential partition, driven by prices: cheaper products and more expensive products belong to different nests. We show the regressions of this candidate nest structure in the bottom left panel, with a resulting R-square of 0.018. This indicates that the candidate partition is unable to explain the variability in the data. The final panel shows the partition that minimizes within-sample error, achieving an $R^2 = 0.73$, and overlays the corresponding demand curves.

Figure 1: Graphical representation of group identification



(a) Data without grouping

(b) Naive partition (median price)

(c) Naive partition demand curves

(d) Estimated demand curves

Notes: Each point represents a product plotted in the space defined by price and residualized log market shares. The top left first shows the raw cloud of observations. The top right panel shows a naive partition obtained by splitting products at the median price, corresponding to a high-price and a low-price nest. The bottom left panel corresponds to the best linear fit following from such price partition. The bottom right panel shows the partition and linear fit that minimizes the objective function.

4 Estimation

Our estimation follows a two-step procedure. In the first step, the *classification step*, we estimate the nesting structure. In the second step, *the estimation step*, we recover the utility parameters conditional on the classification of the first step. Throughout, K is treated as known. Section 6.1 describes how we select it. In what follows, we presents the two estima-

tion steps in turn, followed by a discussion on alternative estimation procedures.

4.1 First Step: Classification

Equation (6) is the basis for identification of the nesting structure. It will also be the basis of the estimation of the partition that constitutes the products' nests. The classifier searches for the partition of products that best explains the log shares through a nest-specific regression on product characteristics. With this procedure, we do not recover the structural parameters β and $\sigma_1, \ldots, \sigma_K$. Only the partition is of interest here.

We estimate the nesting structure by least squares over partitions, slopes and group-market intercepts, in the spirit of the grouped fixed-effects estimator of Bonhomme and Manresa (2015). This estimator is a generalization of the k-means clustering estimator, where points are classified in groups that minimize mean squared error. In our setting, we want to cluster *demand curves*, so the relevant objects are lines rather than points. Products are assigned to the same group when its demand curve best fits their observed relationship between prices and market shares.

Prices are endogenous, so we follow Choi and Okui (2024): We first purge the endogenous part of prices with the projection (8), and then classify on fitted prices. The procedure has two stages.

Price stage. We estimate the nests, the coefficients and the intercepts of the price projection (8) by least squares over partitions:

$$\big(\hat{g}(1), \ldots, \hat{g}(J)\big) = \underset{g(1),\ldots,g(J)}{\operatorname{argmin}} \left\{ \min_{\gamma,\kappa} \sum_{j=1}^{J} \sum_{m=1}^{M} \big(p_{jm} - \mathbf{x}_{jm}\gamma_x^{g(j)} - \gamma_z^{g(j)} z_{jm} - \kappa_{g(j)m}\big)^2 \right\}. \tag{9}$$

We denote the estimated partition by $\hat{g}$ to distinguish it from the estimate of the demand stage. The fitted price is

$$\hat{p}_{jm} \equiv \mathbf{x}_{jm}\hat{\gamma}_x^{\hat{g}(j)\prime} + \hat{\gamma}_z^{\hat{g}(j)\prime} z_{jm} + \hat{\kappa}_{\hat{g}(j)m}.$$

Demand stage. We replace p_{jm} by $\hat{p}_{jm}$ in (6) and estimate the nests, slopes and intercepts

by least squares over partitions:

$$\left(\hat{k}(1), \ldots, \hat{k}(J)\right) = \underset{k(1),\ldots,k(J)}{\operatorname{argmin}} \left\{ \min_{\beta,\lambda} \sum_{j=1}^{J} \sum_{m=1}^{M} \left(\log \mathbb{P}_{jm} - \beta_{k(j),p} \hat{p}_{jm} - \mathbf{x}_{jm} \beta_{k(j),x} - \lambda_{k(j)m} \right)^2 \right\}. \tag{10}$$

Only $\hat{k}$ is of interest. The slopes and intercepts of the classification step are nuisance parameters. Utility parameters are recovered in the estimation step. In Appendix C.1 we describe the algorithms we use to solve (9) and (10).[7]

Remark 4.1 (Reduced-form classification). The price projection (8) allows the grouping of the price equation to differ from the nesting structure. It nests two simpler cases: a price grouping equal to the nesting structure, $g(j) = k(j)$, and a price equation with no groupings. In both cases, a one-stage procedure is available. Substitute (8) into (6) and classify on the exogenous variables $(\mathbf{x}_{jm}, z_{jm})$ in one stage. Appendix B.4 treats this one-stage reduced-form classifier.

Remark 4.2 (Classifying on prices without a projection). Classifying directly on $(p_{jm}, \mathbf{x}_{jm})$, ignoring the endogeneity of prices, recovers the nests in one case. Suppose

$$\mathrm{E}[\xi_{jm} \mid p_{jm}, \mathbf{x}_{jm}] = \phi_{g'(j)}\, p_{jm} + c_{g'(j)m},$$

where the partition g' is coarser than the nests, so that every nest k is contained in one group of g'. Then (6) holds with the price slope $(\beta_p + \phi_{g'(j)})/\sigma_{k(j)}$ in place of $\beta_p/\sigma_{k(j)}$, the intercept $\lambda_{k(j)m} + c_{g'(j)m}/\sigma_{k(j)}$ in place of $\lambda_{k(j)m}$, and an error that is mean independent of the covariates. The slopes and intercepts are biased, but the first step does not use them. However, we still require a price projection in the second step to recover unbiased prefernce parameters.

4.2 Second Step: Structural Parameters

This second step conditions on the estimated classification from the first step and applies the inversion of Berry (1994). Log shares relative to the outside good satisfy (4), which we recall

[7] Note that the classification does not depend on the size of the market or on the definition of the outside option, as long as the outside option forms its own nest. Both enter the log share of every inside product in market m through the same additive term, which is absorbed by the nest-market intercepts λ_{km}.

here:

$$\log \mathbb{P}_{jm} - \log \mathbb{P}_{0m} = \beta_p p_{jm} + \beta_x' \mathbf{x}_{jm} + (1 - \sigma_{k(j)}) \log \mathbb{P}_{j|k(j),m} + \xi_{jm}.$$

The within-nest share is not observed directly, but it can be constructed from the shares and the partition, $\mathbb{P}_{j|k(j),m} = \mathbb{P}_{jm} / \sum_{\{i|k(i)=k(j)\}} \mathbb{P}_{im}$. With estimated nests we construct it with $\hat{k}$ in place of k.

Instruments. Equation (4) has two sources of endogeneity. First, prices may be correlated with ξ_{jm}. Second, the within-nest share is also a function of ξ_{jm}. To address this second source of endogeneity, we use an instrument h_{jm} for the within-nest share in addition to the price instrument z_{jm}. The instrument has to move the relative attractiveness of product j within nest $k(j)$ without being correlated with ξ_{jm}. When x_{jm} are exogenous, several common choices are available: the number of products in the nest, averages of the characteristics of the other products in the nest (BLP-type instruments, Berry et al., 1995), deviations of characteristics relative to the other products in the nest (differentiation instruments, Gandhi and Houde, 2019), or the within-nest share predicted by exogenous characteristics alone. In our empirical application, we use the last one:

$$h_{jm} = \log\Big(\frac{e^{\mathbf{x}_{jm}}}{J_{k(j)}^{-1} \sum_{\{i|k(i)=k(j)\}} e^{\mathbf{x}_{im}}}\Big),$$

where J_k is the number of products inside nest k. Because the coefficient on $\log \mathbb{P}_{j|k,m}$ is nest-specific, h_{jm} enters interacted with the nest indicators. Write e_k for the k-th unit vector of $\mathbb{R}^K$. The instrument and regressor vectors are

$$\tilde{z}_{jm} \equiv \big(\mathbf{x}_{jm}',\ z_{jm}',\ h_{jm}\, e_{k(j)}'\big)', \qquad R_{jm} \equiv \big(p_{jm},\ \mathbf{x}_{jm}',\ \log \mathbb{P}_{j|k(j),m}\, e_{k(j)}'\big)',$$

and the vector of parameters is

$$\theta \equiv \big(\beta_p,\ \beta_x',\ 1 - \sigma_1, \ldots, 1 - \sigma_K\big)'.$$

Estimator. Define $y_{jm} \equiv \log \mathbb{P}_{jm} - \log \mathbb{P}_{0m}$. The GMM estimator solves

$$\hat{\theta} = \arg\min_{\theta}\ \bar{g}(\theta)' \hat{W} \bar{g}(\theta), \qquad \bar{g}(\theta) \equiv \frac{1}{JM} \sum_{j=1}^{J} \sum_{m=1}^{M} \tilde{z}_{jm}\big(y_{jm} - R_{jm}'\theta\big), \tag{11}$$

where $\hat{W}$ is a positive definite weight matrix. Since $\bar{g}(\theta)$ is linear in θ, the estimator has the closed form

$$\hat{\theta} = \left(\hat{G}'\hat{W}\hat{G}\right)^{-1}\hat{G}'\hat{W}\,\frac{1}{JM}\sum_{j=1}^{J}\sum_{m=1}^{M}\tilde{z}_{jm}\,y_{jm}, \tag{12}$$

where $\hat{G} \equiv (JM)^{-1}\sum_{j=1}^{J}\sum_{m=1}^{M}\tilde{z}_{jm}R'_{jm}$ is the sample Jacobian. We call $\hat{\theta}$ with the optimal weight matrix the two-step GMM estimator. The feasible estimator replaces k by $\hat{k}$ in the within-nest share, in $\tilde{z}_{jm}$ and in R_{jm}. We write $\hat{\theta}(\hat{k})$ for the feasible estimator and $\hat{\theta}(k)$ for the infeasible oracle estimator that uses the true nests.[8]

5 Statistical Properties

In this section, we provide conditions to our two main results: first, that the classification step recovers the nesting structure with probability tending to one as J and M grow, and second, that the second-step estimator computed with the estimated nests has the same asymptotic distribution as the oracle estimator computed with the true nests. As a consequence, standard errors for $\hat{\theta}$ need no adjustment for the classification. In addition, we also provide conditions for the oracle estimator when M and J grows, which to our best of our knowledge, was absent from the literature.

Asymptotic Experiment. We consider a triangular array setting, with a sequence of nested logit models indexed by J, with $J, M \to \infty$ and K fixed. The partition $k(\cdot)$ is fixed along the sequence and every nest holds a non-vanishing fraction of the products. Prices may depend on J because equilibrium prices are functions of the characteristics of all products in the market. Consequently, instruments may also depend on J. Finally, σ_k, may also depend on J, to capture variation in the degree of substitution within nest as products are sampled. We treat the observed market shares as population shares and abstract from sampling error in their observation.[9] For the exogenous variables of market m write $\mathcal{X}_m \equiv \sigma(\{\mathbf{x}_{jm}, z_{jm}\}_{j\leq J})$.

[8] Appendix B.6 studies an alternative estimator that demeans (4) within products. It allows product fixed effects, which absorb the level, at the cost of the faster rate.

[9] Observed shares are usually computed from a sample of N consumers per market. With estimated shares, the asymptotic distribution of the second-step estimator acquires additional bias and variance terms unless the number of consumers grows fast enough relative to the sample (Berry et al., 2004; Freyberger, 2015).

5.1 Intuition of Statistical Justification

In this section, we give insights about the arguments that provide justification of our methodology. We start by providing intuition on the validity of the classification step.

5.1.1 Classification Step

A key ingredient in bounding the misclassification probability uniformly over products is a separation condition. In other words, products can be classified only if the nests are distinguishable from each other. Here this means that the slopes and intercepts $(\beta_k; \lambda_{k1}, \ldots, \lambda_{kM})$ differ across nests. The intercepts λ_{km}, however, are not fixed parameters but random variables that drift to $-\infty$ as $J \to \infty$. The first step of the proof splits the intercept of nest k in market m, λ_{km}, into two parts. The first part, $\lambda^0_{km,J}$, is a nest-specific term that stays bounded in probability as J grows and it depends on the distribution of mean utilities within the nest. The second part is a drift, $-\log J$, which is common to all nests:

$$\begin{aligned}\lambda_{km,J} &= (\sigma_k - 1)\log\Big(\sum_{\{j\in k\}} e^{\frac{\delta_{jm}}{\sigma_k}}\Big) - \log\Big(1 + \sum_{l=1}^{K}\Big(\sum_{\{j\in l\}} e^{\frac{\delta_{jm}}{\sigma_l}}\Big)^{\sigma_l}\Big) \\ &= \underbrace{-\log\Big(\frac{1}{J}\sum_{\{j\in k\}} e^{\frac{\delta_{jm}}{\sigma_k}}\Big) + \log\frac{\Big(\frac{1}{J}\sum_{\{j\in k\}} e^{\frac{\delta_{jm}}{\sigma_k}}\Big)^{\sigma_k}}{J^{-\sigma_k} + \sum_{l=1}^{K} J^{\sigma_l-\sigma_k}\Big(\frac{1}{J}\sum_{\{j\in l\}} e^{\frac{\delta_{jm}}{\sigma_l}}\Big)^{\sigma_l}}}_{\lambda^0_{km,J}} - \log J.\end{aligned}$$

If the law of large numbers applies to each sample mean above— and the sigmas are not too distinct, so that $J^{\sigma_l-\sigma_k}$ does not drift for any pair (k,l)—$\lambda^0_{km,J}$ concentrates around a finite non-random center, λ^0_{km}, as J grows. Hence, $\lambda^0_{km,J} = \lambda^0_{km} + e_{km}$, with e_{km} being a residual that vanishes as J grows. Separation can then be stated in terms of the limit λ^0_{km}, which does not depend on J.[10] Let us now re-write (6) as follows:

$$\log \mathbb{P}_{jm} = \beta_{k(j),p} p_{jm} + \beta'_{k(j),x}\mathbf{x}_{jm} + \underbrace{\lambda^0_{k(j)m} + e_{km} - \log J}_{\lambda_{km,J}} + v_{jm}.$$

[10] The common drift $-\log J$ cancels in every comparison between nests, so it plays no role in the classification as the proofs allow the centers of all nests to drift together

With this re-parametrization of the model, we now study the misclassification probability in a simplified set-up. Suppose there are two nests and drop the covariates, so that (6) reads $\log \mathbb{P}_{jm} = \lambda_{k(j)m,J} + v_{jm}$. For given candidate intercepts, the classifier assigns product j to nest 2 when the residual under nest 2 is smaller than the residual under nest 1. Take the true centres with their drift, $\lambda^0_{km} - \log J$, as the candidates, and suppose product j belongs to nest 1 but it is missclassified inside nest 2. The misclassification probability is therefore given by:

$$
\begin{aligned}
&\mathbb{P}\left(\sum_{m=1}^{M}\left(\log \mathbb{P}_{jm} - \lambda^0_{2m} + \log J\right)^2 < \sum_{m=1}^{M}\left(\log \mathbb{P}_{jm} - \lambda^0_{1m} + \log J\right)^2\right) \\
&\quad = \mathbb{P}\left(\frac{1}{M}\sum_{m=1}^{M}(\lambda^0_{2m} - \lambda^0_{1m})\, v_{jm} + \frac{1}{M}\sum_{m=1}^{M}(\lambda^0_{2m} - \lambda^0_{1m})\, e_{1m} > \frac{1}{2M}\sum_{m=1}^{M}(\lambda^0_{2m} - \lambda^0_{1m})^2\right) \\
&\quad \leqslant \mathbb{P}\left(\frac{1}{M}\sum_{m=1}^{M}(\lambda^0_{2m} - \lambda^0_{1m})\, v_{jm} > \frac{1}{4M}\sum_{m=1}^{M}(\lambda^0_{2m} - \lambda^0_{1m})^2\right) \\
&\qquad + \mathbb{P}\left(\frac{1}{M}\sum_{m=1}^{M}(\lambda^0_{2m} - \lambda^0_{1m})\, e_{1m} > \frac{1}{4M}\sum_{m=1}^{M}(\lambda^0_{2m} - \lambda^0_{1m})^2\right),
\end{aligned}
$$

where we use that $\log \mathbb{P}_{jm} + \log J = \lambda^0_{1m} + e_{1m} + v_{jm}$ for a product in nest 1.

The term $M^{-1}\sum_m (\lambda^0_{2m} - \lambda^0_{1m})^2 \geq c > 0$ is bounded away from zero when the two nests are separated. Notice how this condition does not depend on J. We are now going to show that each of the probabilities on the right of the inequality go to zero as J and M go to infinity.

The second term vanishes as $J \to \infty$, regardless of M because the perturbation e_{1m} converges to zero for every m. For the first term, suppose that the v_{jm} are independent across markets and normal with mean zero and variance s^2. Since the centers are non-random, the first term is normal with mean zero and variance $s^2 M^{-2}\sum_m (\lambda^0_{2m} - \lambda^0_{1m})^2$. The probability of misclassifying product j is then

$$
\begin{aligned}
&\mathbb{P}\Big(\frac{1}{M}\sum_{m=1}^{M}(\lambda^0_{2m} - \lambda^0_{1m})\, v_{jm} > \frac{1}{2M}\sum_{m=1}^{M}(\lambda^0_{2m} - \lambda^0_{1m})^2\Big) = \\
&1 - \Phi\Big(\frac{1}{2s}\Big(\sum_{m=1}^{M}(\lambda^0_{2m} - \lambda^0_{1m})^2\Big)^{1/2}\Big) \leq 1 - \Phi\Big(\frac{\sqrt{cM}}{2s}\Big) \;\to\; 0 \quad \text{as } M \to \infty.
\end{aligned}
$$

which goes to 0 when M grows large as long as $c > 0$. Because both probabilities goes to

zero as M grows, it follows that the probability of misclassification hence goes to zero when there is separation.[11]

5.1.2 Estimation Step

The second step is non-standard in two way. First, the regressor that identifies σ_k, $\log \mathbb{P}_{j|k(j),m}$, drifts. Concretely, it satisfies

$$\log \mathbb{P}_{j|k(j),m} = -\log J_{k(j)} + \frac{\delta_{jm}}{\sigma_{k(j)}} - \log\Big(\frac{1}{J_{k(j)}} \sum_{\{i|k(i)=k(j)\}} e^{\delta_{im}/\sigma_{k(j)}}\Big).$$

Therefore, the variation that identifies σ_k grows with J at rate $\log J_k$ and as a result, the larger J is, the more precisely the level of the outcome in a nest pins down σ_k. We will show later that this drift implies that the nest parameters converge at rate $\log J\sqrt{JM}$, faster than β, which converges at rate $\sqrt{JM}$.

A second non-standard aspect of IV estimation with many products is documented by Armstrong (2016). With single-product firms competing in prices, markups converge to constants as the number of products grows. Instruments that move prices only through markups, such as functions of the characteristics of rival products, then become weak. However, Appendix A.3 shows that with multiproduct firms with non-vanishing market shares markups vary by firm and nest and do not collapse to nest constants. In our application, the main brewers hold large within-nest shares in every nest. This supports the relevance of instruments built from the characteristics of rival products, which move prices through markups.

5.2 Conditions

We impose the following conditions on the nested logit model. For a random variable Y, we write $\|Y\|_2 \equiv (\mathrm{E}\, Y^2)^{1/2}$, and $\|\cdot\|$ is the Euclidean norm. For a symmetric matrix, $\lambda_{\min}(\cdot)$ is its smallest eigenvalue. Let J_k be the number of products in nest k and

$$S_{km} \equiv \frac{1}{J_k} \sum_{\{i|k(i)=k\}} e^{\delta_{im}/\sigma_k} \tag{13}$$

[11] This second term in the misclassification probability is already present in the argument of Bonhomme and Manresa (2015), while the second term is new.

the nest average of the exponentiated utilities, with non-random centring $\mu_{km} \equiv \mathrm{E}\, S_{km}$. The inclusive value of nest k is $J_k S_{km}$.

(C1) *Nests.* $k(\cdot)$ is deterministic, and there is $\underline{\pi} > 0$ with $J_k/J \geq \underline{\pi}$ for all k and J.

(C2) *Demand.*

(a) markets are independent across m;

(b) $\mathrm{E}[\xi_{jm} \mid \mathcal{X}_m] = 0$ for all j, m and J, and there is $L_\xi < \infty$ such that

$$\sup_{J,m} \frac{1}{J} \sum_{j=1}^{J} \sum_{i=1}^{J} \left\| \mathrm{E}[\xi_{jm}\xi_{im} \mid \mathcal{X}_m] \right\|_2 \leq L_\xi, \qquad \sup_{J,m} \frac{1}{J^2} \sum_{j,i,l,n=1}^{J} \left\| \mathrm{E}[\xi_{jm}\xi_{im}\xi_{lm}\xi_{nm} \mid \mathcal{X}_m] \right\|_2 \leq L_\xi;$$

(c) $\sup_{j,m,J} \|(\mathbf{x}_{jm}, \xi_{jm}, p_{jm})\|_{\psi_2} < \infty$, where $\|\cdot\|_{\psi_2}$ is the subgaussian norm;[12]

(d) for every $q \geq 2$ there is C_q with $\mathrm{E} \left| \sum_{\{j|k(j)=k\}} \left(e^{\delta_{jm}/\sigma_k} - \mathrm{E}\, e^{\delta_{jm}/\sigma_k} \right) \right|^q \leq C_q J_k^{q/2}$ for all k, m, J.

(e) *Demand parameters.* β lies in a bounded set and $\sigma_k \in [\underline{\sigma}, \bar{\sigma}] \subset (0, 1)$ for all k and J.

(f) *Drift.* $(\sigma_k - \sigma_\ell) \log J = O(1)$ for all k, ℓ.

(C3) *Rates.* $J = O(M^{q_1})$ for some finite q_1.

(C4) *Price projection.*

(a) $\mathrm{E}[\omega_{jm}] = 0$, $\mathrm{E}[\omega_{jm}\mathbf{x}_{jm}] = 0$ and $\mathrm{E}[\omega_{jm}z_{jm}] = 0$ for all j, m and J, and there is $L_\omega < \infty$ such that

$$\sup_{J,m} \frac{1}{J} \sum_{j=1}^{J} \sum_{i=1}^{J} \left(\left| \mathrm{E}[\omega_{jm}\omega_{im}] \right| + \left| \mathrm{E}[\omega_{jm}\xi_{im}] \right| \right) \leq L_\omega;$$

(b) $\sup_{j,m,J} \|z_{jm}\|_{\psi_2} < \infty$.

(c) with $g = k$ in (8) and $\gamma^k \equiv (\gamma_x^{k\prime}, \gamma_z^{k\prime})'$, the coefficients γ^k and the intercepts κ_{km} are non-random and lie in a bounded set for all k, m and J.

(C5) *Separation.* There is $c > 0$ such that, for all $k \neq \ell$ and all J, M large,

[12] For a scalar random variable Y, $\|Y\|_{\psi_2} \equiv \inf\{t > 0 : \mathrm{E}\, e^{Y^2/t^2} \leq 2\}$, and for a random vector V, $\|V\|_{\psi_2} \equiv \sup_{\|u\|=1} \|u'V\|_{\psi_2}$. A bounded subgaussian norm means that the tails are no heavier than those of a normal distribution; see Vershynin (2018, Section 2.5).

(a) in demand,

$$\frac{1}{M}\sum_{m=1}^{M}\Big[(1-\sigma_k)\log\big(J_k\mu_{km}\big)-(1-\sigma_\ell)\log\big(J_\ell\mu_{\ell m}\big)\Big]^2\geq c.$$

(b) in prices, for all j,

$$\frac{1}{M}\sum_{m=1}^{M}\mathrm{E}\left[\big((\gamma^\ell-\gamma^k)'(\mathbf{x}'_{jm},z'_{jm})'+\kappa_{\ell m}-\kappa_{km}\big)^2\right]\geq c.$$

(C6) *Rank condition in the first step.* There is $c_R>0$ such that, for every nest k, every set $A\subset\{j\mid k(j)=k\}$ with $|A|\geq\underline{\pi}J/K$, and all J,M large, the matrix

$$\frac{1}{M}\sum_{m=1}^{M}\mathrm{E}\Big[\frac{1}{|A|}\sum_{j\in A}\big(a_{jm}-\bar{a}_{Am}\big)\big(a_{jm}-\bar{a}_{Am}\big)'\Big],\qquad \bar{a}_{Am}\equiv\frac{1}{|A|}\sum_{j\in A}a_{jm},$$

has smallest eigenvalue at least c_R for

(a) $a_{jm}=(\mathbf{x}'_{jm},z'_{jm})'$, the exogenous variables;

(b) $a_{jm}=(p_{jm}-\omega_{jm},\mathbf{x}'_{jm})'$, the population fitted price and the characteristics.

(C7) *Rank and variance in the second step.* With $k=k(j)$, let

$$F_{jm}\equiv\big(p_{jm},\ \mathbf{x}'_{jm},\ \big(\sigma_k^{-1}\delta_{jm}-\log\mu_{km}-\log(J_k/J)\big)\,e_k'\big)',$$

let d_x be the dimension of $\mathbf{x}_{jm}$, let $E_K\equiv[\,0_{K\times(1+d_x)}\ \ I_K\,]$ select the nest components of θ, and let

$$D_J\equiv\begin{pmatrix}I_{1+d_x}&0\\0&(\log J)^{-1}I_K\end{pmatrix},\qquad \Pi_{JM}\equiv\frac{1}{JM}\sum_{j=1}^{J}\sum_{m=1}^{M}\mathrm{E}\left[\tilde{z}_{jm}e'_{k(j)}\right].$$

Define the scaled population Jacobian and the variance of the moment function at the truth,

$$\Gamma^L_{JM}\equiv\frac{1}{JM}\sum_{j=1}^{J}\sum_{m=1}^{M}\mathrm{E}\left[\tilde{z}_{jm}F'_{jm}\right]D_J-\Pi_{JM}E_K,\qquad \Omega^L_{JM}\equiv\frac{1}{JM}\sum_{m=1}^{M}\sum_{j=1}^{J}\sum_{i=1}^{J}\mathrm{E}\left[\tilde{z}_{jm}\tilde{z}'_{im}\,\xi_{jm}\xi_{im}\right].\tag{14}$$

There is $c_\Gamma > 0$ with $\lambda_{\min}(\Gamma_{JM}^{L\prime}\Gamma_{JM}^{L}) \geq c_\Gamma$ for all J, M large, and $\liminf_{J,M} \lambda_{\min}(\Omega_{JM}^{L}) > 0$. In addition, $\sup_{j,m,J} \mathrm{E}\,\|z_{jm}\|^8 < \infty$ and $\sup_{j,m,J} \mathrm{E}\,|h_{jm}|^8 < \infty$.

We now comment on some of the conditions and the economic implications behind them. (C2)(b) makes the exogenous variables of a market strictly exogenous for demand within the market. While it is mostly standard in the empirical literature of industrial organization, it rules out, for instance, that firms reposition their products in response to demand shocks to competitors. It is what makes instruments built from the characteristics of other products valid in the second step. While we do not require that products to be independent within a market, (C2)(d) restricts the dependence of characteristics across products. For example, products that share a brand may have correlated characteristics, but this correlation needs to be weak when averaged across products in the same market. (C4) is the counterpart of (C2) for the price projection. (C4)(a) only requires orthogonality and not conditional zero means, because ω_{jm} is a projection error. (C4)(a) also restricts how the price residual of a product co-moves with the price residuals and demand shocks of other products. It is the counterpart of the covariance bound in (C2)(b). (C4)(b) asks subgaussian tails for the instrument and covers instruments built from own characteristics and from averages over rivals.

The two residuals are treated differently for two reasons. First, ξ_{jm} is a structural error. The validity of instruments built from the characteristics of rivals rests on $\mathrm{E}[\xi_{jm} \mid \mathcal{X}_m] = 0$. ω_{jm} is a projection error. Conditional mean zero of ω_{jm} would require $\mathrm{E}[p_{jm} \mid \mathcal{X}_m]$ to be linear in $(\mathbf{x}_{jm}, z_{jm})$, which fails when markups depend on the characteristics of rivals. Second, the second step multiplies ξ_{jm} by random instruments, so its covariance bound is conditional on $\mathcal{X}_m$. ω_{jm} enters only the first step, where an unconditional bound is enough.

(C7) is the condition for the second step. Π_{JM} collects the nest means of the instruments. The rescaling D_J reflects the faster rate of the nest parameters, which Section 5.4 explains. (C7) requires the scaled population Jacobian to have full column rank, the variance of the moment function to be non-degenerate, and the instruments to have eight moments. The central limit theorem uses the eight moments to bound the fourth moments of sums of dependent terms. Ω_{JM}^{L} includes the covariances between products of the same market, which (C2)(b) allows. A variance estimator therefore has to cluster by market. For large J, the nest columns of Γ_{JM}^{L} are $-\Pi_{JM} + O(1/\log J)$. So the rank condition asks that the nest means of the instruments identify the nest coefficients. With the exogenous share projection h_{jm}, the relevant block of Π_{JM} is diagonal with entries $(J_k/J)\ \mathrm{E}[h_{jm} \mid k(j) = k]$. It has full rank when

these nest means are bounded away from zero. The moment conditions on the instruments are weaker than the subgaussian tails in (C4)(b). The differentiation instruments are of the order of x_{jm}^2 and have all moments, but they are not subgaussian, so they satisfy (C7) but not (C4)(b).

(C2)(f) bounds the differences between nest parameters at a rate $1/\log J$. It makes the intercepts of all nests drift to $-\infty$ at the same rate, $\log J$, so that their differences stay bounded. Every nest then keeps a positive share of the market in the limit. Without (C2)(f), the nests with the largest σ_k take the whole market as $J \to \infty$, and the intercepts of the other nests drift at nest-specific rates. (C2)(f) is not needed for the classification result. Without it, nests with different σ_k are separated at rate $\log J$, so separation becomes easier, and the proof only needs a parameter set for the intercepts that grows with J. We keep (C2)(f) because it describes the empirically relevant case, in which no nest takes over the market, and because it keeps the proofs short.

5.3 Results

In this section we state our main results formally. Proofs can be found in Appendix B.1.

Theorem 5.1 (Consistent classification)**.** *Under (C1)–(C6), for every $q \geq 1$ and $\delta > 0$, the two-stage classifier (9)–(10) satisfies*

$$\mathbb{P}\big(\hat{k} \neq \tau \circ k \text{ for every permutation } \tau \text{ of } \{1, \dots, K\}\big) = O\big(J^{-q}\big) + O\big(JM^{-\delta}\big),$$

and the price-stage partition satisfies $\mathbb{P}\big(\hat{g} \neq \tau \circ g \text{ for every permutation } \tau\big) = O\big(JM^{-\delta}\big)$. Both right sides tend to zero for $\delta > q_1$.

The labels of the groups carry no information, so the partition is recovered up to a permutation of the labels. The bound has two terms with different sources. The first term is the probability that the perturbation e_{km} is large. By (C2)(d), e_{km} has all moments of order $J_k^{-1/2}$, so this probability decays faster than any power of J. The second term collects the probability that the estimated slopes and intercepts are far from the truth, that the fitted prices are far from the population projection, and that a product is misclassified given slopes and intercepts close to the truth. Each is an average over markets of independent terms with subgaussian tails, so its probability decays faster than any power of M. The factor J is the

cost of a union bound over products; the union bounds over pairs of products in the design and criterion steps cost a further factor J, which (C3) absorbs into $M^{-\delta}$. Under (C3), J grows at most polynomially in M, so the second term vanishes once $\delta > q_1$. Products are therefore classified correctly even when the number of markets is much smaller than the number of products. This result is the basis for the use of estimated nests in the second step.

5.4 Asymptotic Properties of the Second Step

The second-step estimator uses the partition in three places: the within-nest share, the nest-specific columns of the regressor and instrument vectors, and possibly the weight matrix. The first step recovers the partition only up to a permutation of the labels, and a permutation of the labels permutes the nest-specific components of θ. We therefore fix a labelling rule: the groups of any partition are numbered in the order of their smallest product index. With this rule, two partitions that split the products into the same groups give the same labels, and the labels of the true partition are non-random. From here on, $\hat{\theta}(\hat{k})$ and $\hat{\theta}(k)$ are computed with this rule.

The oracle estimator. We first study the oracle estimator $\hat{\theta}(k)$, computed with the true nests. Two features of the within-nest share matter. First, it depends on the other products of the nest through the nest average $S_{k(j)m}$ of (13), which fluctuates around its centring $\mu_{k(j)m}$. Second, it contains the level $-\log J_{k(j)} = -\log J - \log(J_{k(j)}/J)$, which grows with J. The scaled population Jacobian Γ^L_{JM} of (C7), in (14), replaces the nest average by its centring and separates the level $-\log J$. The level makes the Jacobian of order $\log J$ in the nest directions. The matrix D_J therefore rescales the nest components of θ by $\log J$. We assume that $\hat{W} - W_{JM} \to_p 0$ for non-random matrices W_{JM} with eigenvalues in $[c_W, C_W]$, $0 < c_W \leq C_W < \infty$.

Theorem 5.2 (Oracle estimator)**.** *Under (C1)–(C3) and (C7),*

$$\sqrt{JM}\, D_J^{-1}\big(\hat{\theta}(k) - \theta_0\big) \to_d N\big(0, V_{JM}\big),$$

where $V_{JM} \equiv (\Gamma^{L\prime}_{JM} W_{JM} \Gamma^L_{JM})^{-1} \Gamma^{L\prime}_{JM} W_{JM}\, \Omega^L_{JM}\, W_{JM} \Gamma^L_{JM} (\Gamma^{L\prime}_{JM} W_{JM} \Gamma^L_{JM})^{-1}$ *is the GMM variance.*

The statement is with finite-sample population matrices. It means $V_{JM}^{-1/2}\sqrt{JM} D_J^{-1}(\hat{\theta}(k) - \theta_0) \to_d N(0, I)$. With the efficient weight $W_{JM} = (\Omega^L_{JM})^{-1}$, $V_{JM} = \big(\Gamma^{L\prime}_{JM} (\Omega^L_{JM})^{-1} \Gamma^L_{JM}\big)^{-1}$.

The proof gives the linear representation $D_J^{-1}(\hat{\theta}(k) - \theta_0) = (\Gamma_{JM}^{L\prime} W_{JM} \Gamma_{JM}^{L})^{-1} \Gamma_{JM}^{L\prime} W_{JM}\, \bar{g}(\theta_0) + o_p((JM)^{-1/2})$, so $\hat{\beta} - \beta_0 = O_p((JM)^{-1/2})$ and $\hat{\sigma}_k - \sigma_k = O_p(1/(\log J\sqrt{JM}))$. The nest parameters converge at rate $\log J\sqrt{JM}$, faster than β, which converges at rate $\sqrt{JM}$. The faster rate comes from the level $-\log J_{k(j)}$ of the regressor, which the nest means of the instruments match. The scaling D_J describes the rates and does not change inference. t-statistics and Wald statistics computed with the conventional GMM variance formula are invariant to a rescaling of the parameter. The level identification requires that the model have no free nest intercepts. A nest intercept would absorb the nest columns $-\Pi_{JM} E_K$ of the Jacobian. No limit of the Jacobian or of the variance is assumed. Only $\mathrm{E}[\xi_{jm} \mid \mathcal{X}_m] = 0$ and the covariance bound in (C2)(b) enter the rate. The fourth-moment bound in (C2)(b) enters once, for the Lyapunov condition in the central limit theorem. The proof is in Appendix B.7. Appendix B.6 gives the corresponding result for the demeaned estimator, Theorem B.4. That estimator allows product fixed effects and converges at rate $\sqrt{JM}$ in all directions.

The feasible estimator. The feasible estimator replaces k by $\hat{k}$. Theorem 5.1 and Theorem 5.2 together give its asymptotic distribution.

Corollary 1 (Feasible estimator)**.** *Let $\hat{k}$ be the classifier of Theorem 5.1 under its conditions, and suppose (C7) holds. Then $\mathbb{P}\big(\hat{\theta}(\hat{k}) \neq \hat{\theta}(k)\big) \to 0$ and*

$$\sqrt{JM}\, D_J^{-1}\big(\hat{\theta}(\hat{k}) - \theta_0\big) \to_d N\big(0, V_{JM}\big).$$

The feasible estimator equals the oracle estimator with probability tending to one. The same holds for any statistic computed from the data and a partition. So every asymptotic property of the oracle transfers, at any rate and to any statistic that depends on the partition, such as standard errors, Wald tests and confidence intervals. Standard errors need no adjustment for the classification step. No sample splitting is needed. The two steps use the same data, and the argument works on the event that the partition is recovered, not through independence between the steps. The result holds along a fixed sequence of data-generating processes, with fixed constants in the separation and rank conditions. It is not uniform over sequences along which the separation constants shrink. When nests are weakly separated relative to M, misclassification can be frequent in finite samples, and the oracle distribution can approximate the feasible one poorly.

5.5 Discussion

5.5.1 Asymptotic Experiment

Our asymptotic analysis considers a sequence of nested logit models in which the number of products J and the number of markets M both grow. This setting is empirically relevant given the growing availability of granular data. Retail scanner data, used in industrial organization, marketing and macroeconomics, record many products across many stores and periods (Chiong and Shum, 2019; Hitsch et al., 2021; DellaVigna and Gentzkow, 2019; Atalay et al., 2026; Döpper et al., 2025; Ershov et al., 2025; Broda and Weinstein, 2010; Hottman et al., 2016; Jaravel, 2019; Redding and Weinstein, 2020). Mobile phone and commuting data record trips between many origins and many destinations within a city (Ahlfeldt et al., 2015; Athey et al., 2021; Kreindler and Miyauchi, 2023; Dingel and Tintelnot, 2026; Miyauchi et al., 2025). Firm-to-firm transaction data record sales between many suppliers and many buyers (Bernard et al., 2019; Dhyne et al., 2021; Bernard et al., 2022). Trade data have a similar structure, with many origins and many destinations (Eaton and Kortum, 2002; Donaldson, 2018; Allen and Arkolakis, 2022). In these data, destinations play the role of products, and origins play the role of markets. A triangular array framework, that is, a setup in which the model changes as the sample size increases, is necessary because prices may respond to all products in the market. When more products are sampled, prices may change. Finally, large J and large M facilitate the statistical analysis, and we discuss the role of each limit in turn.

First, $J \to \infty$ is useful to provide a condition of separation that does not depend on J. The intercepts λ_{km} are random, since they are functions of inclusive values. Invoking the law of large numbers allows us to rely on non-random centers and weaken the conditions of separation to hold only in the limit, and not for every J. With J fixed, three problems arise. First, each intercept is estimated from the J_k products of its nest, so its estimate does not converge. Second, the intercept of product j depends on its own shock ξ_{jm}, through the inclusive value, with weight of order $1/J_k$. The intercept is then correlated with the error of the first-step equation, so it is not exogenous. Third, separation would have to hold for the realized intercepts, not for their centres. The realized differences $\lambda_{\ell m} - \lambda_{km}$ depend on the demand shocks ξ_{jm} and on the characteristics $\mathbf{x}_{jm}$ of the products in both nests. A separation condition on them would therefore restrict the joint distribution of the data, because it would tie the shocks to the characteristics. Such a restriction may conflict with other assumptions

of the model, such as a zero mean of the shocks within each nest or the exogeneity of the characteristics. We avoid this problem by requiring separation only for the limiting centres of the intercepts, which are non-random functions of the primitives.

Second, we need $M \to \infty$ because demand shocks vary across products and markets. A product is classified correctly only if its unobservable demand shocks average out across markets, as Theorem 5.1 shows.

5.5.2 Alternative Estimators

Our estimator uses the structure of the model and proceeds in two linear steps. The first step recovers the partition. The second step recovers the structural parameters given the partition. A natural question is why the partition and the structural parameters cannot be estimated jointly. We discuss two joint estimators and the reason each one fails.

The first is nonlinear least squares on the shares. Let $\mathbb{P}^0_{jm}(\beta, \sigma, k)$ denote the share in (1) when every demand shock is set to zero, $\xi_{im} = 0$ for all i. A joint estimator of (β, σ, k) minimizes $\sum_j \sum_m \left(\mathbb{P}_{jm} - \mathbb{P}^0_{jm}(\beta, \sigma, k)\right)^2$ over parameters and partitions. This estimator is consistent when the regression function is the conditional mean of the observed share, that is, when $\mathrm{E}[\mathbb{P}_{jm} \mid \mathcal{X}_m] = \mathbb{P}^0_{jm}(\beta, \sigma, k)$ at the true values. With demand shocks this equality generally fails. The observed share is a nonlinear function of the shocks of all products in the market, through the exponentials and the inclusive values in (1).[13] So the conditional mean of the share is not the share at zero shocks, even though $\mathrm{E}[\xi_{jm} \mid \mathcal{X}_m] = 0$. The residual $\mathbb{P}_{jm} - \mathbb{P}^0_{jm}(\beta, \sigma, k)$ has a conditional mean that depends on β, σ, the partition and the covariates, so the estimator is inconsistent. The log-share equation (6) avoids this problem. There the shocks enter additively, and the group-market intercept λ_{km} absorbs the nonlinearity. The classifier treats this intercept as a free parameter.

Nonlinear least squares on the shares is consistent when the only error is an additive measurement error in the observed shares, with conditional mean zero. This case requires the observed characteristics to describe demand exactly, with no demand shocks ξ_{jm}. The empirical literature on demand estimation rarely imposes this restriction, and we do not consider this model.

[13] This problem is common in models where errors enter nonlinearly. For example, let $y = \exp(\beta + \xi)$ with $\xi \sim N(0, s^2)$, so that $\mathrm{E}[\xi] = 0$. Nonlinear least squares regresses y on $\exp(b)$, the model with the shock set to zero, and minimizes $\mathrm{E}[(y - e^b)^2]$. The minimizer satisfies $e^{b^*} = \mathrm{E}[y] = e^\beta\ \mathrm{E}[e^\xi] = e^{\beta + s^2/2}$, so $b^* = \beta + s^2/2 \neq \beta$. The shock has mean zero, but it enters inside the exponential, so $\mathrm{E}[y]$ is not the model evaluated at $\xi = 0$.

The second is GMM on the moment conditions $\mathrm{E}[\tilde{z}_{jm}\xi_{jm}(\beta,\sigma,k)] = 0$, with the partition treated as a parameter. Choi and Okui (2024) show that the GMM objective then has no unique minimizer, because the moments do not discriminate between partitions. In the just-identified case, every partition admits coefficients that set the sample moments to zero exactly. In the over-identified case, every partition admits coefficients at which the population moments are zero, so the sample moments converge to zero in probability. In both cases, every partition attains the minimum of the objective in the limit. So the moment conditions alone do not identify the partition.

Our approach requires mean utility δ_{jm} to be linear in parameters. Existing consistency results for classification in nonlinear models (e.g., Su et al., 2016; Bonhomme et al., 2022) assume that the outcome of each unit depends on its own errors. In the nested logit, each share depends on the shocks of all products in the market, so these results do not apply directly. We believe this is an area ripe for more research. Linearity also helps in two other ways. First, replacing the price by its projection is valid only when δ_{jm} is linear in the price. Otherwise the projection error enters the structural error nonlinearly, and the error has a conditional mean different from zero.[14] Second, linearity gives closed-form estimators, while nonlinear estimators could raise computational costs by orders of magnitude.

A separate question is why we do not follow the common practice of classifying products by k-means on observed characteristics. That approach does not use variation in demand across markets. The k-means partition and ours coincide only when the nests are a function of product characteristics, $k(j) = f(\mathbf{x}_j)$. The condition fails when unobserved attributes captured by ξ_{jm}, such as brand reputation or marketing, enter the formation of nests. Outside this case, clustering on characteristics alone has no statistical justification. Section 8.3.4 compares the two classifications in our empirical application.

Finally, either step of our estimator could use data demeaned within products across markets. Demeaning allows product fixed effects, which the levels estimators exclude (Section 6). Its cost differs between the two steps. In the first step, demeaning removes part of the variation that identifies the nests. The size and the average composition of a nest do not vary across markets, so they drop out. In the second step, demeaning (4) removes the level $-\log J_{k(j)}$ of the regressor. The nest parameters are then identified only from the variation of

[14] For instance, let $\delta_{jm} = \exp(\beta_p p_{jm} + \beta_x x_{jm}) + \xi_{jm}$, and let $p_{jm} = \hat{p}_{jm} + \omega_{jm}$ with $\mathrm{E}[\omega_{jm} \mid \mathcal{X}_m] = 0$. Substituting the projection leaves the error $\xi_{jm} + e^{\beta_p \hat{p}_{jm} + \beta_x x_{jm}}\left(e^{\beta_p \omega_{jm}} - 1\right)$. By Jensen's inequality, $\mathrm{E}[e^{\beta_p \omega_{jm}} \mid \mathcal{X}_m] > 1$ when ω_{jm} is non-degenerate, so this error has a non-zero conditional mean.

the within-nest shares across markets. All parameters then converge at rate $\sqrt{JM}$, instead of $\log J\sqrt{JM}$ for the nest parameters. Appendix B.6 defines the demeaned second-step estimator and states its asymptotic distribution in Theorem B.4.

6 Extensions

In this section, we show how to estimate the number of groups and present a model that allows for price endogeneity in the form of product fixed effects when no exogenous price shifters are available.

6.1 Estimating the Number of Nests

So far, the number of groups is assumed to be known and fixed. In this section, we present how to estimate the number of groups. It is well known that choosing the optimal number of groups in a clustering algorithm is a hard problem. We propose to use a cross validation criterion on out-of-sample data. Concretely, we do cross-validation using the "Elbow Method", a widely used heuristic for choosing the right number of groups in clustering algorithms. Other indexes and criterion, such as the Silhouette method, could be applied following a similar logic, but the Elbow Method has the advantage of being computationally simple to implement.

Our proposed algorithm is as follows. First, divide products $\mathcal{P}$ into N equal parts, $\mathcal{P}_1, ..., \mathcal{P}_N$. Second, iterate for all $k \in \{1, ..., K\}$ the following procedure:

1. For every $n = 1, .., N$:

   (a) Remove $\mathcal{P}_n$ products and estimate grouping structure $\mathcal{B}_{-n}(k)$ and grouping parameters $(\beta_{-n}(k), \lambda_{-n}(k))$ in the remaining $\mathcal{P}_{-n}$ products.

   (b) Classify products $j \in \mathcal{P}_n$ using $(\beta_{-n}(k), \lambda_{-n}(k))$:

$$k^*(j) = \underset{k \in \{1,...,K\}}{\arg\min} \sum_{m=1}^{M} (\log \mathbb{P}_{jm} - \log \mathbb{P}_{0m} - \beta_{k,-n} x_j - \lambda_{km,-n})^2.$$

(c) Compute out-of-sample MSE given the previous classification:

$$MSE_n(k) = \frac{1}{J \cdot M} \sum_{m=1}^{M} \sum_{j \in \mathcal{P}_n} (\log \mathbb{P}_{jm} - \log \mathbb{P}_{0m} - \beta_{k,-n} x_{jm} - \lambda_{km,-n})^2.$$

2. Take average across N folds:

$$MSE(k) = \frac{1}{N} \sum_{n=1}^{N} MSE_n(k).$$

Finally, we select k^* as the point at which the elbow occurs, defined by the following criterion:

$$k^* = \arg\max_k \left\{ \Delta^2 \, MSE(k) \;\text{ or }\; d\big(k,\, \overline{k=2,\, k=\bar{K}}\big) \right\},$$

where $\Delta^2 \, MSE(k)$ denotes the second difference (i.e., the reduction in slope) of $MSE(k)$, and $d(\cdot)$ denotes the distance from point k to the line connecting the endpoints $k = 2$ and $k = \bar{K}$. Intuitively, the elbow criterion identifies the number of clusters beyond which adding further groups yields diminishing reductions in $MSE(k)$, trading off fit against model complexity in a manner analogous to information criteria such as AIC or BIC in maximum likelihood estimation. The Elbow method has two main advantages. First, because it relies on out-of-sample fit, it reduces the risk of overfitting. Second, it is straightforward to compute.

In general, this heuristic — or the addition of further selection criteria — may yield multiple candidates for k^*. In such cases, we proceed as follows. First, we discard any candidate k for which the second-step estimation produces parameters inconsistent with utility maximization, namely any estimate of σ falling outside the interval $(0, 1)$.[15] If multiple candidates remain after this pruning step, we select the largest k among them. The rationale is that overestimating the number of groups does not compromise asymptotic consistency, whereas underestimating it does (Moon and Weidner, 2015).

[15] If no candidate survives this elimination step, we report the k^* associated with the lowest p-value rejecting the null hypothesis that all $\hat{\sigma}$ lie within $(0, 1)$.

6.2 Endogenous Product Fixed Effects

Assume that:

$$\xi_{jm} = \mu_j + \nu_{jm},$$

where $\mathrm{E}[\nu_{jm}|p_{jm}, x_{jm}] = 0$ but where there could be correlation between μ_j and p_{jm}. In this case, average utility is given by:

$$\delta_{jm} = \beta_p p_{jm} + \beta_x x_{jm} + \mu_j + \nu_{jm}. \tag{15}$$

We modify the estimation procedure outlined in Sections 4.1 and 4.2 by taking mean differences within products as this transformation differentiates out the time-invariant demand shocks ξ_j from δ_{jm}:

$$\begin{aligned}\nabla\delta_{jm} &= \delta_{jm} - \frac{1}{M}\sum_{m=1}^{M}\delta_{jm} \\ &= \beta_p \nabla p_{jm} + \beta_x \nabla x_{jm} + \nabla\nu_{jm},\end{aligned}$$

where

$$\nabla x_{jm} \equiv x_{jm} - \frac{1}{M}\sum_{m=1}^{M} x_{jm} \quad \text{and} \quad \nabla p_{jm} \equiv p_{jm} - \frac{1}{M}\sum_{m=1}^{M} p_{jm}.$$

It is easy to see from the previous equation and under our assumptions that:

$$\mathrm{E}[\nabla x_{jm}\nabla\nu_{jm}] = 0 \quad \text{and} \quad E[\nabla p_{jm}\nabla\nu_{jm}] = 0.$$

Finally define $\nabla\lambda_{k(j)m} \equiv \lambda_{k(j)m} - \frac{1}{M}\sum_{m=1}^{M}\lambda_{k(j)m}$. The classification algorithm is modified as follows:

1. Let $B_0^1, \ldots, B_0^K$ be a starting group structure.

2. Compute:

$$(\beta^s, \lambda^s) = \underset{(\beta,\lambda)}{\arg\min} \sum_{j=1}^{J}\sum_{m=1}^{M}\left(\nabla \log \mathbb{P}_{jm} - \nabla \log \mathbb{P}_{0m} - \begin{bmatrix}\nabla p_{jm} \\ \nabla x_{jm}\end{bmatrix}\beta_{k(j)} - \nabla\lambda_{k(j)m}\right)^2.$$

3. For θ^s , compute for all $j \in J$:

$$k(j)^{s+1} = \underset{k \in \{1,\ldots,K\}}{\arg\min} \sum_{m=1}^{M} \left(\nabla \log \mathbb{P}_{jm} - \nabla \log \mathbb{P}_{0m} - \begin{bmatrix} \nabla p_{jm} \\ \nabla x_{jm} \end{bmatrix} \beta_k^s - \nabla \lambda_{km}^s \right)^2,$$

to recover grouping structure $\mathcal{B}^{s+1}$.

4. Repeat until convergence of parameters.

Finally, the moment conditions of the second step are modified as follows:

$$\mathrm{E}\left[\begin{bmatrix} \nabla x_{jm} \\ \nabla p_{jm} \\ \nabla w_{jm} \end{bmatrix} \cdot \nabla \nu_{jm} \right] = 0.$$

These are the moment conditions of the demeaned estimator of Appendix B.6, and Theorem B.4 gives its asymptotic distribution. The levels estimator of Section 4.2 does not apply here, because the product fixed effects absorb the level of the within-nest share that identifies the nest parameters.

6.3 Adding Consumer Heterogeneity

We can naturally extend our method to incorporate consumer heterogeneity. Let's start with the case where both consumer heterogeneity and individual choices are observable. Let's assume that consumer's heterogeneity can be described or approximated by types, $\omega^i \in \{\omega^1, ..., \omega^L\}$ where l indexes type. For example, a practitioner could consider multiple income bins as a way to approximate the income distribution. In such case, let's consider a model of utility given by:

$$\delta_{jm}^l = \beta^l x_{jm} + \xi_{jm}^l.$$

Define:

$$\mathbb{P}_{jm}^l = \mathbb{P}_m(j|\omega = \omega^l).$$

Let's first consider a general model where all consumers have a different nesting structures. It follows:

$$\log \mathbb{P}^l_{jm} - \log \mathbb{P}^l_{0m} = \frac{1}{\sigma_{k^l(j)}}\Big(\beta^l x_{jm} + \xi^l_{jm}\Big) + (\sigma_{k^l(j)} - 1)\log IV^l_{k^l(j)m}.$$

Therefore, our objective function to estimate nests becomes:

$$\min_{k(j)\in\{1,\dots,K\}} \min_{(\beta,\lambda)} \sum_{l=1}^{L} \pi_l \sum_{j=1}^{J} \left(\log \mathbb{P}^l_{jm} - \log \mathbb{P}^l_{0m} - \beta^l_{k(j)} x_{jm} - \lambda^l_{k(j)m}\right)^2, \tag{16}$$

where π_l determines type's l weight in the objective function. It is worth noting that in this general model because both the nesting structure as well as errors are type-specific, one could solve the problem type by type and apply the algorithm outlined in Section 4.1 in parallel across types.

It is straightforward to see that equation (16) also applies in the case where the practitioner wants to impose the same nesting structure across consumer types. In such case, it is only necessary to impose such common nesting structure when solving for equation (16).

6.4 Identification with Multiple Levels

In our previous discussion, we assumed just one level of nests. However, NL models can have multiple upper levels of groups. In this section, we extend our discussion to a model with two levels of nests. Generalizing to more levels is straightforward.

Assume we have upstream nests given by $A_1, ..., A_N$, and downstream nests given by $B_1, ..., B_K$, where $N < K$. We denote by $k(j)$ the downstream nest to which product j belongs and by $n(k)$ the upstream nest to which downstream nest k belongs. In this case, choice probabilities can be written as:

$$\mathbb{P}_{jm} = \mathbb{P}_{j|k(j),m}\mathbb{P}_{k(j)|n(k),m}\mathbb{P}_{n(k),m},$$

with conditional probabilities given by:

$$\mathbb{P}_{j|k,m} = \frac{e^{\frac{\delta_{jm}}{\sigma_k}}}{\sum_{j\in B_k} e^{\frac{\delta_{jm}}{\sigma_k}}}$$

$$\mathbb{P}_{k|n,m} = \frac{\left(\sum_{j\in B_k} e^{\frac{\delta_{jm}}{\sigma_k}}\right)^{\sigma_k}}{\sum_{k\in A_n}\left(\sum_{j\in B_k} e^{\frac{\delta_{jm}}{\sigma_k}}\right)^{\sigma_k}}$$

$$\mathbb{P}_{n,m} = \frac{\left(\sum_{n\in A_n}\left(\sum_{j\in B_k} e^{\frac{\delta_{jm}}{\sigma_k}}\right)^{\sigma_k}\right)^{\sigma_n}}{\sum_k\left(\sum_{k\in A_n}\left(\sum_{j\in B_k} e^{\frac{\delta_{jm}}{\sigma_k}}\right)^{\sigma_k}\right)^{\sigma_n}}.$$

Taking logs and under the assumption that $\delta_{0m} = 0$ with the outside option being a singleton nest, it follows:

$$\ln \mathbb{P}_{jm} - \ln \mathbb{P}_{0m} = \frac{\delta_{jm}}{\sigma_k} + (\sigma_k - 1)\ln IV_{km} + (\sigma_n - 1)\ln IV_{nm},$$

with inclusive values defined as

$$IV_{km} = \sum_{j\in B_k} e^{\frac{\delta_{jm}}{\sigma_k}}$$

$$IV_{nm} = \sum_{k\in A_n}\Big(\sum_{j\in B_k} e^{\frac{\delta_{jm}}{\sigma_k}}\Big)^{\sigma_k}.$$

Assume that $\delta_{jm} = \beta x_{jm}$ and define $\beta_k \equiv \frac{\beta}{\sigma_k}$, $\lambda_{km} \equiv (\sigma_k - 1)\ln IV_{km}$, $\lambda_{nm} = (\sigma_n - 1)\ln IV_{nm}$, $\alpha \equiv -\ln IV_m$. Substituting,

$$\ln \mathbb{P}_{jm} - \ln \mathbb{P}_{0m} = \beta_k x_{jm} + \lambda_{km} + \mu_{nm}.$$

Observe that we cannot directly estimate all λ_{km} and μ_{nm} in the previous model because of co-linearity. Therefore, we impose the restriction $\tilde{\lambda}_{k(n)m} = 0$ for some $k \in A_n$, say $k(n)$. In such case, our model becomes:

$$\ln \mathbb{P}_{jm} - \ln \mathbb{P}_{0m} = \beta_k x_{jm} + \tilde{\lambda}_{km} + \tilde{\mu}_{nm},$$

where

$$\tilde{\mu}_{nm} = \lambda_{k(n)m} + \mu_{nm}$$

and

$$\tilde{\lambda}_{km} = \begin{cases} 0 & \text{for } k = k(n) \in A_n \\ \lambda_{km} - \lambda_{k(n)m} & \text{for } k \in A_n, k \neq k(n). \end{cases}$$

We modify our classification algorithm of Section 4.1 as follows:

1. Let $B_0^1, \ldots, B_0^K, A_0^1, ..., A_0^N$ be a starting group structure.

2. For every $n = 1, ..., N$, pick some $k(n)$ such that $B_k \subset A_n$ and set $\tilde{\lambda}_{k(n)m} = 0$ for all $m = 1, .., M$.

3. Compute:

$$(\beta^s, \lambda^s, \mu^s) = \underset{\beta,\lambda,\mu}{\arg\min} \sum_{j=1}^{J} \sum_{m=1}^{M} \left(\log \mathbb{P}_{jm} - \log \mathbb{P}_{0m} - \beta_{k(j)} x_{jm} - \tilde{\lambda}_{k(j)m} - \tilde{\mu}_{n(j)m} \right)^2$$
$$\text{subject to } \tilde{\lambda}_{k(n)m} = 0 \quad \forall n = 1, ..., N, \forall m = 1, ..., M.$$

4. For $(\beta^s, \lambda^s, \mu^s)$, construct $\theta_{km}^s = \lambda_{km}^s + \mu_{nm}^s$, and compute for all $j \in J$:

$$k(j)^{s+1} = \underset{k \in \{1,...,K\}}{\arg\min} \sum_{m=1}^{M} \left(\log \mathbb{P}_{jm} - \log \mathbb{P}_{0m} - \beta_{k,p} p_{jm} - \beta_{k,x} x_{jm} - \theta_{km} \right)^2,$$

   with $n(j)^s = n(k(j)^{s+1})$ to recover grouping structure $\mathcal{B}^{s+1}$ and $\mathcal{A}^{s+1}$.

5. Repeat until convergence of parameters.

6.5 Type II EV error: Nested CES Demand

Under nested CES, a representative consumer maximizes a two-level CES utility function over products $j \in \{1, \ldots, J\}$ organized into nests $B_1, \ldots, B_K$:

$$U_m = \left(\sum_{k=1}^{K} \left(\sum_{j \in B_k} (A_{jm}\, c_{jm})^{(\sigma_k - 1)/\sigma_k} \right)^{\frac{\sigma_k}{\sigma_k - 1} \cdot \frac{\sigma_0 - 1}{\sigma_0}} \right)^{\sigma_0/(\sigma_0 - 1)},$$

where c_{jm} is consumption of product j, A_{jm} is a quality shifter, $\sigma_k > 1$ is the within-nest elasticity of substitution, and $\sigma_0 \geqslant 1$ is the across-nest elasticity. An equivalent microfoundation

arises from a random utility model where idiosyncratic shocks are drawn from a nested Fréchet (Type II EV) distribution with within-nest shape parameter $\sigma_{k(j)} - 1$ and across-nest shape parameter $\sigma_0 - 1$, directly paralleling the nested logit derivation where shocks follow a GEV distribution. Standard expenditure minimization gives the market share:

$$\mathbb{P}_{jm} = \frac{(p_{jm}/A_{jm})^{1-\sigma_{k(j)}}}{\sum_{d \in B_{k(j)}} (p_{dm}/A_{dm})^{1-\sigma_{k(j)}}} \cdot \frac{\left(\sum_{d \in B_{k(j)}} (p_{dm}/A_{dm})^{1-\sigma_{k(j)}}\right)^{(1-\sigma_0)/(1-\sigma_{k(j)})}}{\sum_{l=1}^{K} \left(\sum_{d \in B_l} (p_{dm}/A_{dm})^{1-\sigma_l}\right)^{(1-\sigma_0)/(1-\sigma_l)}} \tag{17}$$

$$= (p_{jm}/A_{jm})^{1-\sigma_{k(j)}} \Lambda_{k(j)m}, \tag{18}$$

where $\Lambda_{k(j)m}$ collects all nest- and market-level terms. Taking logs and setting $\log A_{jm} = \beta_x x_{jm} + \xi_{jm}$:

$$\log \mathbb{P}_{jm} = \underbrace{(\sigma_{k(j)} - 1)\beta_x}_{\beta_{k(j),x}} x_{jm} - \underbrace{(\sigma_{k(j)} - 1)}_{-\beta_{k(j),p}} \log p_{jm} + \lambda_{k(j)m} + \xi_{jm}, \tag{19}$$

which exactly maps to the first-step equation of our baseline model.

7 Monte Carlo

In this section, we assess the performance of our estimator using a Monte Carlo exercise. Across our different Monte Carlo simulations, we vary the number of products J and the number of markets M, while fixing the rest of the fundamentals. In what follows, we describe our data generating process for fixed $J, M,$ and K.

7.1 Data Generating Process for Panel Data

First, we fix the number of groups $K = 3$ and classify products using uniform draws $k(j) \sim \mathcal{U}\{1, ..., K\}$, requiring that each group has at least twenty percent of products.[16] We fix the vector of $\boldsymbol{\sigma} = (\sigma_1, ..., \sigma_K)$ to certain values in the interval $[0, 1]$ and order them in ascending order. In our main Monte Carlo exercise, we fix $\boldsymbol{\sigma} = (0.2, 0.3, 0.6)$.

[16] We need at least two products per group. The reason is that we need within market-group variance in product characteristics to identify the group-specific parameters.

The average utility δ_{jm} is given by:

$$\delta_{jm} = \beta_p p_{jm} + \beta_x x_{jm} + \xi_{jm},$$

where, $\beta_p = -1, \beta_x = 1$, x_{jm} are exogenous characteristics and p_{jm} are prices that can be correlated with unobservable demand shocks ξ_{jm}. Concretely, we assume that prices are given by $p_{jm} = mc_{jm} + \rho \cdot \xi_{jm}$, where mc_{jm} is orthogonal to ξ_{jm} and $\rho = 0.3$.[17]

To construct product characteristics, first we draw the mean of characteristics mc and x as follows:

$$\begin{bmatrix} \chi_{k,x} \\ \chi_{k,mc} \end{bmatrix} \overset{i.i.d.}{\sim} \mathcal{N}\Big(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.3 \\ 0.3 & 1 \end{bmatrix} \Big),$$

across groups k.

Next, for a given vector of means $(\chi_{k,x}, \chi_{k,mc})$, we draw the vector of characteristics as follows:

$$\begin{bmatrix} x_{jm} \\ mc_{jm} \\ \xi_{jm} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \chi_{k,x} \\ \chi_{k,mc} \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.3 & 0 \\ 0.3 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \right),$$

across j and m.

Once we have generated δ_{jm} for all j and m and in combination with the vector $(\sigma_1, ..., \sigma_K)$, we proceed to construct inclusive values according to the following expressions:

$$IV_{km} = \Big(\sum_{d \in B_k} e^{\frac{\delta_{dm}}{\sigma_k}} \Big).$$

Finally, leveraging the closed-form solution of choice probabilities in NL models, we construct choices probabilities using equation (1).

7.2 Estimation

Our estimation proceeds in two steps. In the first step, we take data generated as described above in the form of

$$\{(\log \mathbb{P}_{jm} - \log \mathbb{P}_{0m}, p_{jm}, x_{jm}, z_{jm})\}_{j,m},$$

[17] Monte Carlo simulations under Nash–Bertrand pricing yield very similar results. We nevertheless focus on this specification because price endogeneity persists as J grows large.

and classify products into K groups according to the algorithm described in Section 4.1.

The output of this classification step is a partition of products into K groups $\hat{\mathcal{B}} = \{\hat{B}_k\}_{k=1}^K$, as well as estimated parameters:

$$(\hat{\beta}_{k,p}, \hat{\beta}_{k,x}) \quad \text{and} \quad \hat{\lambda}_{k,m} \qquad \forall k, m,$$

which are non-linear functions of the structural parameters. Given $\hat{\mathcal{B}}$, in the second step we mostly follow Berry (1994). First, we construct the following conditional choice probabilities:

$$\hat{\mathbb{P}}_{j|k(j),m} = \frac{\mathbb{P}_{j,m}}{\sum_{j \in \hat{B}_k} \mathbb{P}_{j,m}}.$$

We then estimate the following linear equation:

$$\log \mathbb{P}_{jm} - \log \mathbb{P}_{0m} = \beta_p p_{jm} + \beta_x x_{jm} + (1 - \sigma_{k(j)}) \log \hat{\mathbb{P}}_{j|k(j),m} + \xi_{j,m}.$$

As mentioned earlier $\log \hat{\mathbb{P}}_{j|k(j),m}$ is endogenous because it is also function of the unobservable component $\xi_{j,m}$ by construction. To overcome this problem, we instrument $\log \hat{\mathbb{P}}_{j|k(j),m}$ using the projection of $\log \hat{\mathbb{P}}_{j|k(j),m}$ on exogenous characteristics as an instrument:

$$w_{j,m} = \log \Big(\frac{\exp(x_{jm})}{\sum_{j' \in k(j)} \exp(x_{j'm})} \Big).$$

7.2.1 Results

In this section we present our Monte Carlo results for various cases. We vary the number of products to be equal to $J = 100, 500, 1000$ to assess the performance of our estimator in markets of different size. We also set the number of markets equal to $M = 10, 50, 100$ to asses the properties of our estimator in short and large panels. We set the number of random initial conditions $I = 1000$. The results of these Monte Carlo simulations can be found in Table 1.

First, we observe that with as few as ten markets, the match rate between the estimated groups and the true groups is no less than 98.8%. Second, when J and M grow the precision of the estimates increases as expected. Also, conditional on J, increasing the number of markets M leads to higher match rates. However, this is not always the case for fixed M and higher J. The reason is that by increasing J two forces are at play that go in oppo-

Table 1: Monte Carlo Results with Endogenous Prices

					β_p	β_c	σ_1	σ_2	σ_3
J	M	Runtime	Matched	True	-1	1	0.2	0.3	0.6
100	10	00:01	0.988	Mean β	-1.001	0.992	0.195	0.298	0.599
				Std β	0.037	0.031	0.013	0.007	0.01
100	50	00:18	1.0	Mean β	-1.0	1.003	0.2	0.3	0.6
				Std β	0.012	0.012	0.001	0.002	0.004
100	100	01:42	1.0	Mean β	-1.001	1.001	0.2	0.3	0.6
				Std β	0.008	0.008	0.001	0.001	0.003
500	10	00:14	0.995	Mean β	-1.0	0.998	0.199	0.3	0.6
				Std β	0.013	0.012	0.006	0.005	0.003
500	50	03:59	1.0	Mean β	-1.001	1.0	0.2	0.3	0.6
				Std β	0.006	0.004	0.001	0.001	0.001
500	100	17:42	1.0	Mean β	-1.0	1.0	0.2	0.3	0.6
				Std β	0.004	0.003	0.0	0.001	0.001
1000	10	00:15	0.995	Mean β	-1.001	1.002	0.2	0.3	0.599
				Std β	0.01	0.011	0.002	0.002	0.002
1000	50	08:08	1.0	Mean β	-1.002	1.001	0.2	0.3	0.6
				Std β	0.003	0.004	0.0	0.0	0.001
1000	100	31:48	1.0	Mean β	-1.0	1.0	0.2	0.3	0.6
				Std β	0.003	0.003	0.0	0.0	0.001

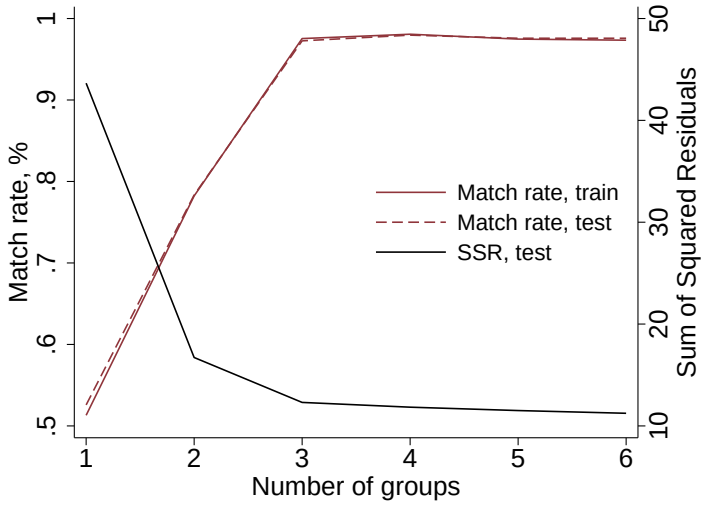
Notes: This table shows average Monte Carlo results over $\max\{k \cdot (P + M) \cdot 50, 2000\}$ bootstrap samples, with $K = 3$ groups, where P is the number of covariates and M is the number of markets. Column 1 reports the number of products $J \in \{100, 500, 1000\}$. Column 2 reports the number of markets $M \in \{10, 50, 100\}$. Column 3 reports the average running time in hours:minutes for one bootstrap sample on a 32-core machine with 16GB of RAM per core. Column 4 reports the average share of products that are correctly matched with their true group. The remaining columns report the average estimates and standard deviations of the structural parameters.

site directions: On the one hand, as J increases the estimated nesting structure is closer to the true nesting structure and thus will exhibit lower mis-classification rate. On the other hand, increasing the number of products also mechanically increases the probability of mis-classification. Overall, we take these results as evidence that our estimation procedure can correctly recover the groups when the market share data is generated using a NL model.

7.2.2 Estimating the Number of Groups

In this section, we show using Monte Carlo simulations how to estimate the number of groups. Our data generating process in this case also follows Section 7.1. The true number of groups is $K = 3$, the total number of products $J = 500$, and the total number of markets is $M = 10$. We do cross-validation exercise described in Section 6.1 on ten folds on data. That is, we train the model in 90% of the data and leaving out 10% of the observations for out of sample testing. The results are presented in Figure 2. We clearly observe very marginal improvements in the predictive power of the model when $K > 3$ as the SSR function becomes almost flat after $K = 4$. Therefore, our cross-validation procedure suggests that the optimal number of groups should be $K^* = 3$, which corresponds to the population number of groups with which the data was generated. We take these results as suggestive evidence that our cross-validation procedure can recover the underlying number of groups.

Figure 2: Average cross-validation results across bootstrap samples



Notes: The figures above show the out-of-sample sum of squared errors on 10% of the data across different number of groups. Parameters are set to be J = 500, M = 10, with true number of groups K = 3. The data generating process follows the procedure described in Section 7.1.

8 Empirical application: Demand for Beer in the US

In this section, we describe how to apply our estimator to the US beer market.

8.1 Data Description

For our main application, we use NielsenIQ store panel data for 2022. This dataset captures purchases from all selling points covered by NielsenIQ scanner data, including supercenters and supermarkets. Each entry records the purchase of a single retail product, including its unit price, the number of units sold, and a store identifier. We define products as Universal Product Codes (UPCs), which combine a retail product with its packaging type.[18] We focus on the beer category, which contains 24,188 unique UPCs. NielsenIQ also provides auxiliary datasets with information on volume size per unit (e.g., whether the product contains 330-ml or 500-ml units), package configuration (e.g., 6-pack cans or 4-pack bottles), beer type (e.g., ale or stout), packaging type (e.g., bottle, can, box), brand, and whether the beer is imported or domestic.

We consider 48 markets, defined as the continental US states. To aggregate sales within a state, we sum all units sold across selling points within that state. We construct a market-specific UPC price by taking a weighted average, where weights are the number of UPCs sold within each store.[19] Finally, because our method relies on longitudinal observations, we retain only products sold in at least ten markets, leaving a total of 2806 unique UPCs.[20] These remaining 11.6% of UPCs cover most of the sales volume, ranging from 61.5% to 99% of total units sold within a state. We combine the sales of all these infrequent UPCs into the outside option. Our final dataset is an unbalanced panel of 64497 observations.[21] Given its unbalanced nature, we adapt our algorithm to allow for the entry and exit of products across markets.

[18] For example, a single 330-ml can of a particular beer has a different UPC than a 6-pack of 330-ml cans of the same beer.

[19] Before state-level aggregation, we drop sales with prices below the 1st or above the 99th percentiles to remove price outliers and potential data entry errors. After state aggregation, we also drop products whose descriptions contain "MALTA GOYA" because these are misclassified as alcoholic beer. We also drop 23 products with missing data on their characteristics.

[20] Many UPCs are sold in very few locations. For example, 11,904 UPCs are sold in only one state and 3,923 UPCs are sold in only two states.

[21] Only 136 UPCs are sold across all 48 states. After applying our filter, the average UPC is sold in 23.32 states, implying an unbalanced panel with approximately 50% of markets missing per product on average.

Table 2 reports summary statistics for our main dataset. These statistics are weighted by sales volume. Even when aggregated at the national level, we observe substantial variation in prices and liquid content.

Table 2: Summary statistics

Panel A: Full Sample Summary Statistics

Variable	Mean	Std. Dev.	10th pct	Median	90th pct
Price ($)	10.618	6.758	2.596	10.298	20.308
Unit quantity (units)	8.838	8.229	1.000	6.000	24.000
Price per ml ($)	0.004	0.001	0.002	0.004	0.005
Unit size (ml)	535.627	450.385	354.882	354.882	739.338
Market share	0.859%	1.152%	0.061%	0.498%	1.967%

Panel B: Median Characteristics of the top ten national brands

Brand	Market share	Unit price/ml	Unit price	Unit quantity	Unit size
Bud Light	12.53%	0.003	8.63	6.00	473.18
Miller	7.81%	0.003	12.47	12.00	354.88
Modelo	7.63%	0.005	10.23	6.00	354.88
Coors	7.11%	0.003	12.62	12.00	354.88
Bud	6.73%	0.003	3.22	1.00	739.34
Michelob	5.96%	0.004	15.20	12.00	354.88
Corona	4.59%	0.004	16.52	12.00	354.88
Busch	4.43%	0.002	10.87	12.00	354.88
Natural	3.63%	0.002	6.49	6.00	473.18
New Belgium	1.98%	0.005	10.28	1.00	567.81

Notes: Panel A summarizes key statistics of product characteristics across beers, weighting each product by its sales volume. Panel B reports the median characteristics of the top ten national brands in 2022.

8.2 Demand Model and Estimation

In our baseline demand specification, we include only three continuous product characteristics—log unit size, log units per UPC, and log prices—and no categorical variables. This setup can be viewed as a conservative benchmark for assessing the performance of our method in uncovering the nesting structure. In Section 8.3, we report a robustness exercise that incorporates categorical variables describing beer type in the estimation of the grouping structure.

We also include market fixed effects to capture demand shocks common to all products within a market (e.g., temperature), as well as market-level variation in the attractiveness

of the outside option, since it is not uniformly defined by the same set of products across markets. The resulting mean utility is:

$$\delta_{jm} = \varphi_m + \beta_s \log \mathrm{s}_j + \beta_n \log \mathrm{n}_j + \beta_p \log \mathrm{p}_{jm} + \xi_{jm},$$

where s_j is unit size, n_j is the number of units per UPC, and p_j is price.

Given that the utility levels are not identified (Train, 2009), we normalize the utility of the outside option $j = 0$ to zero, namely:

$$\delta_{0m} = 0 \text{ for all } m = 1, ..., M.$$

For simplicity, we also assume that the outside option constitutes its own nest for all $m = 1, ..., M$ following many empirical IO applications. Given this assumption, we will classify only on inside goods during our classification step.

To address potential endogeneity of prices due to correlation with ξ_{jm}, we construct Gandhi–Houde instruments (Gandhi and Houde, 2019).[22] For each product, we compute the average squared difference in log unit size and log units between product j and the other products sold in market m:

$$\mathrm{gh}^1_{jm} = \frac{1}{J_m} \sum_{j' \in \mathcal{J}_m} (\log \mathrm{s}_j - \log \mathrm{s}_{j'})^2 \quad \text{and} \quad \mathrm{gh}^2_{jm} = \frac{1}{J_m} \sum_{j' \in \mathcal{J}_m} (\log \mathrm{n}_j - \log \mathrm{n}_{j'})^2,$$

where $\mathcal{J}_m$ is the set of products sold in market m and J_m is its size. The instruments measure how much the other products in the market differ from product j.

Although product characteristics are constant across markets, the set of available products varies. Spatial variation in the assortment therefore provides cross-market variation in these instruments. Relevance relies on the link between product differences and prices. Products that differ more from the other products in the market face weaker competitive pressure and set higher prices. The exclusion restriction requires that the characteristics of the other products in the market do not depend on the demand shock ξ_{jm}. This holds, for

[22] The differentiation instruments are not subgaussian. Each instrument is an average of squared differences of characteristics, so it is sub-exponential when the characteristics are subgaussian. The proofs in Appendix B extend to sub-exponential instruments. The extension replaces the concentration inequalities for subgaussian variables with their counterparts for sub-exponential and sub-Weibull variables, which also give misclassification probabilities that vanish faster than any power of M.

example, if firms choose their products and characteristics before demand shocks realize. It fails if firms adjust product characteristics in response to demand shocks.

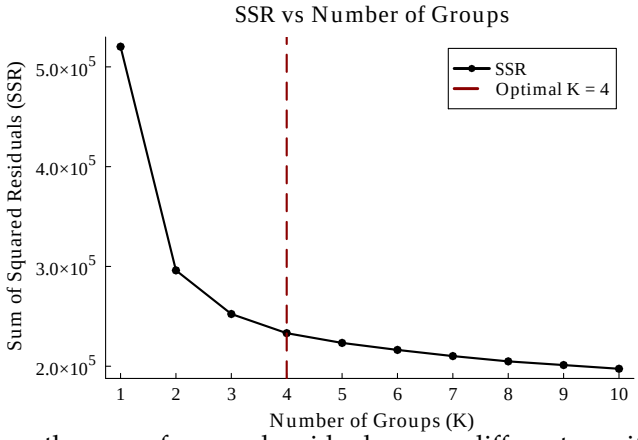
We begin by estimating the first stage. With three product characteristics and 48 market fixed effects, estimating a model with k groups requires $(3 + 48) \cdot k$ parameters. Given the large parameter space, it is important to use many starting values. In general, we set the number of initial conditions to $k \cdot (M + P) \cdot 50$, where M is the number of markets and P is the number of covariates.[23]

8.2.1 First-Step Results

As part of the first stage, we begin by determining the number of groups. To do so, we follow the cross-validation procedure described in Section 6.1 and run the classification algorithm for $k = 1, \ldots, 10$. The results, shown in Figure 3, indicate that improvements in predictive performance are negligible for $k \geqslant 5$.

Applying the Elbow method yields two candidate values, $k = 3$ and $k = 4$. Both specifications produce within-nest independence parameters, σ_k, that are consistent with utility maximization. In this case, we select the largest admissible number of groups as picking more groups does not compromise asymptotic consistency (Moon and Weidner, 2015). Accordingly, our preferred specification sets $k^* = 4$.

Figure 3: Sum of squared residuals across different number of groups



Notes: The figure above shows the sum of squared residuals across different specifications with $k = 2, 3, \ldots, 8$. The results are estimated using Gandhi-Houde instruments. The optimal k^* is the one with the largest distance to the uniform line.

[23] For our baseline classification and K = 2,...,5, results become rather stable after 2,000 initial conditions.

Next, we proceed to described the average product characteristics conditional on every nest for $k^* = 4$. Results can be seen in Table 3. Overall, we find four differentiated groups in terms of median characteristics. Group 1 seems to correspond to cheaper average regular beer (e.g. lager). Group 2 on the other hand, represents a group of light beers with top ten brands representing 75.5% of all their UPCs, significantly above the rest of the other groups. Groups 3 and 4 contain more ales, less light beers, and less as well top-ten brands compared to the other groups. While they look similar, we can see that group 4 is more expensive, both in terms of unit price as well as price per ml.

Table 3: Characteristics of beer across groups

Characteristics	Unconditional values	Conditional values by group 1	2	3	4
Unit size (ml)	354.882	473.176	354.882	354.882	354.882
Unit quantity	6.000	6.000	6.000	6.000	6.000
Unit price	10.209	9.857	10.156	12.367	10.509
Price per ml	0.004	0.003	0.004	0.004	0.005
Share of domestic	0.794	0.820	0.773	0.830	0.861
Share of ale	0.158	0.225	0.098	0.291	0.377
Share of regular beer	0.442	0.469	0.430	0.442	0.326
Share of stout and porter	0.015	0.019	0.011	0.018	0.046
Share of light beer	0.369	0.266	0.446	0.238	0.239
Share of other types	0.016	0.022	0.015	0.010	0.012
Share top-ten brands	0.624	0.460	0.755	0.382	0.334
# of products	2806	659	242	1102	803
Label		Regular Lager	Light Beer	Special/ Craft	Boutique Craft

Notes: Rows 1–4 report median values of unit size, unit quantity, unit price, and price per ml, presented both unconditionally and conditional on group membership. Rows 5–11 report unconditional and conditional mean shares. Groups are estimated using Gandhi-Houde instruments.

Since classification is performed at the UPC level, individual brands may span multiple groups. To obtain a unique group assignment for each brand, we assign it to the group with the highest likelihood. Figure 4 displays the resulting brand classification, where the size of each logo reflects the brand's relative volume within its assigned group. Group 2 is clearly populated by mainstream light beers, including Bud Light, Michelob Ultra, Coors Light, and Miller Lite. Group 1, by contrast, contains many of the non-light counterparts of those same

brands. Group 3 is more heterogeneous, combining some international lagers like Tecate with less mainstream offerings such as Founders. Group 4 consists predominantly of craft beers from local and boutique breweries. Based on the average product characteristics and the most prominent brands within each group, we label Group 1 as *Regular Lager*, Group 2 as *Light Beer*, Group 3 as *Special/Craft*, and Group 4 as *Boutique Craft*.

As an additional validation exercise, we compute the relative popularity of each group across U.S. counties with respect to the national average. Despite the classification incorporating no spatial information, Figures 5 and 6 reveal a geographic patterns that exhibit a strong spatial autocorrelation in the popularity of these groups: *Light Beer* is disproportionately popular in the Southern states, while *Boutique Craft* is concentrated in the Northeast, Oregon, and Washington State. Moreover, across counties, the market share of *Light Beer* has a correlation of 0.39 with the Republican vote share in the 2024 presidential election.[24]

8.2.2 Out-of-sample Validation: Bud Light Boycott

We validate the substitution patterns implied by our estimated nesting structure using the 2023 Bud Light boycott as an exogenous demand shock. In April 2023, a social media promotion of Bud Light featuring Dylan Mulvaney, a transgender actress and influencer, triggered a widespread backlash and calls for a boycott. The episode received extensive media coverage and had measurable effects on Bud Light sales, providing a plausibly exogenous negative demand shock to a single product, namely Bud Light.

If our first-step classification captures substitution patterns reasonably well, we should see more substitution toward products that share a group with Bud Light—*Light Beer*—than toward products in other nests following the boycott. To test this prediction, we augment our state-product panel with data from 2023 and estimate the following difference-in-differences specification:

$$\log \mathbb{P}_{jmt} = \kappa_{jm} + \eta_t + \sum_{k=1}^{K} \alpha_k \, \mathbf{1}\{b_{k(j)} = k\} \, \mathbf{1}\{t = 2023\} + \varepsilon_{jmt}, \tag{20}$$

where κ_{jm} are product-by-market fixed effects, η_t is a year fixed effect, and the interaction terms capture differential changes in log market shares across nests between 2022 and 2023.

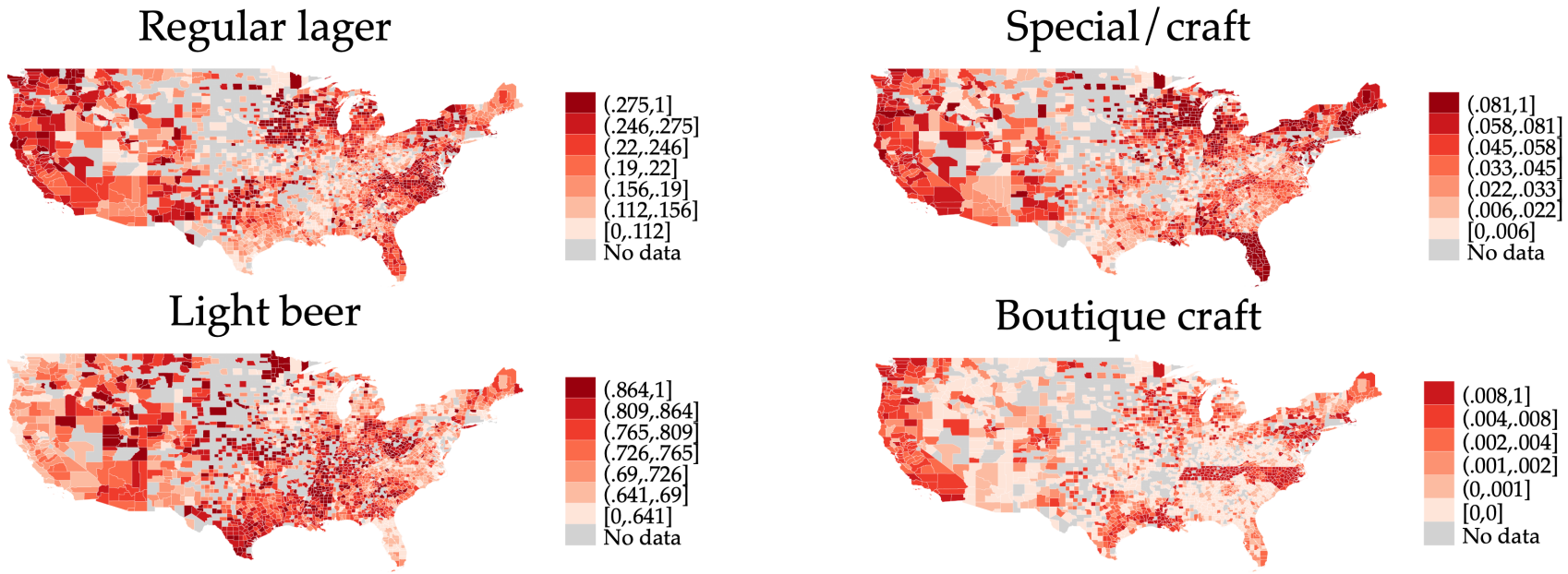
[24] There is a literature documenting that alcohol preferences vary with regional and political characteristics (Yakovlev and Guessford, 2013).

Figure 4: Top 10 brands by group



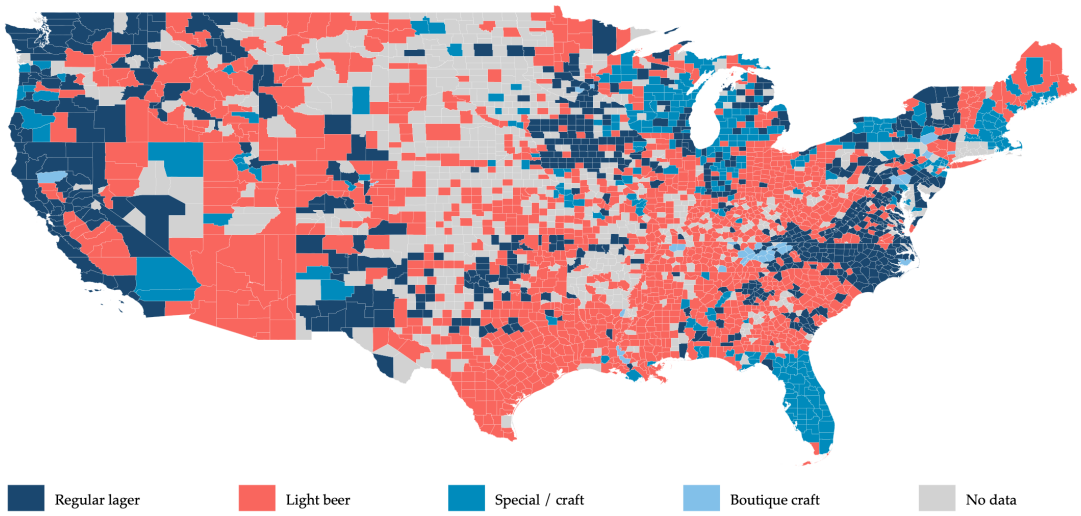
Notes: This figure shows the top ten most national sold beer brands at the national level for each group. The bubble size of each brand reflects the relative size of the market share in each particular group.

Figure 5: Spatial correlation of county market shares by group



Notes: These maps above show the market share of units sold of each beer group over the total sales at the county level.

Figure 6: Most popular group relative to national average



Notes: This map displays the most popular group relative to the national trend in terms of market share of units sales at the county level.

The inclusion of product fixed effects κ_{jm} is important: identification comes entirely from temporal variation within product and market, which was not used in the first-step classification of nests, effectively making this an out-of-sample validation of the estimated nesting structure.

Table 4: Diff-in-Diff results of Bud Light boycott

Dependent Variable: Log Share								
	First stage		DiD				Placebo	
	(1)		(2)		(3)		(4)	
t=2023	-0.364***	(0.005)	-0.408***	(0.005)	-0.449***	(0.008)	-0.361***	(0.010)
Treatment (Budlight)	-0.157***	(0.026)						
Treatment same group			0.422***	(0.008)				
Regular beer × t=2023					0.173***	(0.012)		
Light beer × t=2023					0.463***	(0.010)		
Non-lager/craft × t=2023					-0.001	(0.013)		
Boutique craft × t=2023					0.018	(0.020)		
Placebo Regular beer × t=2023							-0.020	(0.014)
Placebo Light beer × t=2023							-0.003	(0.014)
Placebo Non-lager/craft × t=2023							-0.004	(0.014)
Placebo Boutique craft × t=2023							0.009	(0.015)
Observations	177784		172778		172778		172778	
UPC-Market FE	YES		YES		YES		YES	
Exclude Bud Light	No		YES		YES		YES	
Baseline group	Non-Budlight		Other groups		Outside option		Outside option	

Notes: *** p<0.01, ** p<0.05, * p<0.1

This table reports difference-in-differences estimates of the effect of the April 2023 Bud Light boycott on log market shares, defined as the log of the ratio of units sold for a given UPC in a state to total units sold in that state. Column (1) includes all products and estimates the average effect on Bud Light relative to all other products. Columns (2)–(4) exclude Bud Light from the sample. Column (2) defines treatment as all products in the Light Beer group (excluding Bud Light) and control as products in all remaining groups; the baseline group is other data-driven groups. Column (3) estimates group-specific effects relative to the outside option. Column (4) reports a placebo test in which products are randomly assigned to five groups. All specifications include UPC-by-market fixed effects. Heteroskedasticity-robust standard errors in parentheses.

Table 4 reports the results. The first specification shows that the share of Bud Light fell significantly in 2023 relative to other beers, confirming that the boycott had a statistically significant effect. The second specification sets treated products as those who are in the *Light Beer* group other than Bud Light and control products as those in other groups as well as in the outside option, showing a positive and significant increase in the sales of the treated products relative to control groups. The third specification estimates a different treatment effects for every group, where the control group is defined as those products in the outside options. The coefficient on *Light Beer* is large and statistically significant ($\hat{\alpha}_{\text{Light}} = 0.463$), indicating that products in the same nest as Bud Light experienced a significant increase in market share following the boycott. By contrast, the coefficients for *Regular Lager* ($\hat{\alpha}_{\text{Regular}} = 0.173$), *Special/Craft* ($\hat{\alpha}_{\text{Special}} = -0.001$), and *Boutique Craft* ($\hat{\alpha}_{\text{Boutique}} = 0.018$)

are substantially smaller. These results are consistent with the substitution patterns predicted by the nested logit model: following a negative shock to Bud Light, consumers substituted primarily toward other products in the *Light Beer* group rather than toward products in other nests. Finally, the last specification reports a placebo test using randomly drawn groups that confirms that our results are not driven by anything systematic in the data.

8.2.3 Second-Step Results

Table 5: Estimation Results for Nested Logit with $k^* = 4$

	OLS	IV
log unit size	0.070 (0.006)***	0.339 (0.025)***
log upc quantity	0.080 (0.005)***	0.441 (0.026)***
log price	-0.054 (0.006)***	-0.416 (0.033)***
$\sigma_{\text{regular/lager}}$	0.142 (0.001)***	0.523 (0.014)***
$\sigma_{\text{light beer}}$	0.343 (0.002)***	0.807 (0.016)***
$\sigma_{\text{special/craft}}$	-0.015 (0.001)***	0.336 (0.013)***
$\sigma_{\text{boutique craft}}$	-0.320 (0.001)***	0.042 (0.015)***
Mean Own-price Elasticity	1.351 (0.003)	-3.561 (0.013)
Mean Cross-price Elasticity	-0.003 (0.008)	0.011 (0.033)
Market Fixed Effects	✓	✓
IV Type		$GH_{\text{units}}, GH_{\text{size}}$
Number of Products	2806	
Number of Observations	64497	

Notes: This table reports second-step estimates of the nested logit demand model with $k^* = 4$ groups recovered in the first step. The dependent variable is the log ratio of a product's market share to the outside option share. The OLS column estimates equation (4) without instrumenting for price or the within-group share. The IV column instruments for log price using Gandhi-Houde instruments—the average squared difference in log unit size and log unit quantity relative to competing brands—and for the log within-group share using the exogenous share projection described in Section 4.2. σ_k denotes the within-nest correlation parameter for group k, where $\sigma_k \in (0, 1)$ is consistent with utility maximization. Mean own-price and cross-price elasticities are computed by averaging across products and markets, weighted by national sales volume; standard deviations of the elasticity distribution across products are reported in parentheses. All specifications include market fixed effects. Standard errors in parentheses. $***p < 0.01, **p < 0.05, *p < 0.1$.

We now show the estimation results of preference parameters using $k^* = 4$ as the number of nests. We estimate preference parameters by estimating the second-stage equation:

$$\log \mathbb{P}_{j,m} - \log \mathbb{P}_{0,m} = \varphi_m + \beta_s \log \mathrm{s}_j + \beta_n \log \mathrm{n}_j + \beta_p \log \mathrm{p}_{jm} + (1 - \sigma_{k(j)}) \log \mathbb{P}_{j|k(j),m} + \xi_{jm}.$$

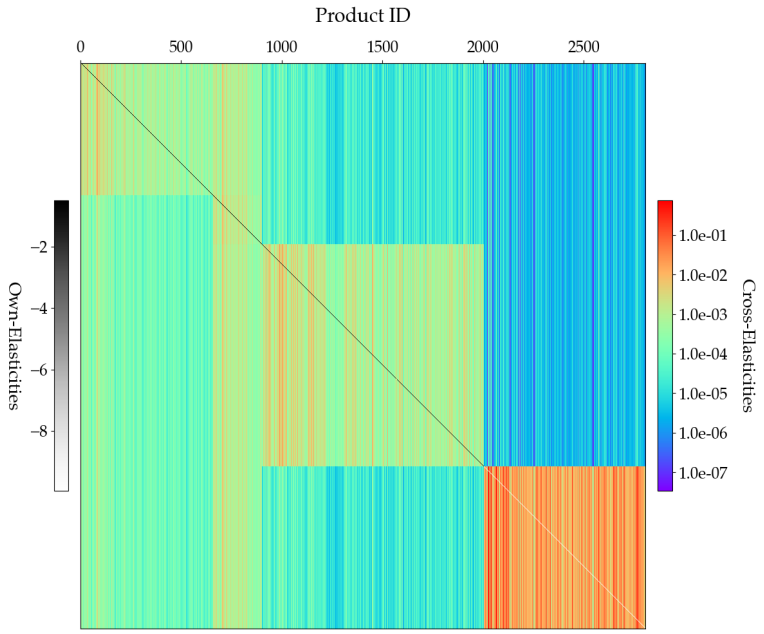
To estimate σ_k we need to instrument $\log \mathbb{P}_{j|k(j),m}$ because those are mechanically correlated with ξ_{jm}. We construct the following instruments:

$$\frac{\exp(\log \mathrm{s}_j)}{\sum_{j \in k(j,m)} \exp(\log \mathrm{s}_j)} \quad \text{and} \quad \frac{\exp(\log \mathrm{n}_j)}{\sum_{j \in k(j,m)} \exp(\log \mathrm{n}_j)},$$

where $k(j,m)$ denotes the products in group $k(j)$ that are also present in market m. Results can be seen in Table 5. As expected, OLS yields an upward-biased price coefficient of -0.054 and some negative within-nest correlations, implying a positive average own-price elasticity. With Gandhi-Houde instruments, the price coefficient increases in magnitude to -0.416, roughly an order of magnitude larger. Notably, the estimates of within-nest correlations all take values that are consistent with utility maximization as they all lie between zero and one. These estimation results were obtained without imposing any constraints on σ_k's. For example, we also find that *Light Beer* has the largest $\hat{\sigma}_k$ equal to 0.807 and that *Boutique Craft* has the lowest sigma equal to 0.042. These results imply that for *Light Beer* independent idiosyncratic shocks dominate choices relatively more than for *Boutique Craft*, where shocks are nearly identical and choices are mostly determined by differences in δ_{jm}. Consequently, *Light Beer* is the most inelastic group and with the lowest within-group substitution as opposed to *Boutique Craft* the most elastic group and with the largest within-group substitution.

Next, we move on to analyze price elasticities. Because we have many products that are not available in all markets, we construct average elasticities by taking averages across markets. When doing so, we obtain an average own-price elasticity of -3.561, similar to previous estimates in the literature studying demand for beer (Pinkse and Slade, 2004; Slade, 2004; Romeo, 2016; Miller and Weinberg, 2017). The average cross-price elasticity is 0.011 with standard deviation 0.033. The full graph of price elasticities can be seen in Figure 8a, where we have ordered products first by group and then by UPC identifier. Following the interpretation on $\hat{\sigma}_k$, we see a lot of within-group substitution for *Boutique Craft*, in contrast to *Light Beer* that behaves close to a simple Logit model with proportional substitution patterns.

Figure 7: Price elasticity matrices of Data-Driven NL



Notes: The figures above display the price elasticity matrices estimated using the nested logit specification and the BLP procedure. Diagonal cells each matrix denote products' own-price elasticities, whereas the off-diagonal cells denote cross-price elasticities. Concretely, the value in row i, column j denotes $\frac{\partial \log \mathbb{P}_i}{\partial \log p_j}$.

8.3 Robustness

In this section, we present a battery of robustness checks.

8.3.1 Hausman instruments

We also estimate the first step using Hausman instruments for prices (Hausman, 1996):

$$\mathrm{h}_{jm} = \frac{1}{N_j - 1} \sum_{m' \neq m} \log p_{jm'},$$

where N_j denotes the number of markets in which product j is sold. Relevance relies on cost shocks that are common across markets, such as input costs. The exclusion restriction requires that ξ_{jm} be uncorrelated with prices in other markets conditional on covariates. This condition holds if demand shocks are independent across markets but may fail if, for example, marketing campaigns are targeted at specific regions.

Results are reported in Appendix C.2.1. The estimated nesting structure is broadly simi-

lar to the baseline: we again recover four groups as the optimal specification, and the conditional group characteristics closely resemble those in the baseline in terms of median product attributes. The structural parameter estimates remain consistent with utility maximization, although we fail to reject that one of the nesting parameters is statistically indistinguishable from one. The implied own-price elasticity is somewhat smaller at -1.612, falling slightly outside the range of -6.1 to -3.3 typically reported in the literature. Overall, while both instrument sets produce qualitatively similar groupings, we adopt the Gandhi-Houde specification as our baseline given the stronger empirical support for that model and the more credible exclusion restriction in this setting.

8.3.2 Results with three and five groups

When changing the number of groups, we find a somewhat hierarchical structure in the grouping and we increase the number of groups. For example, when $k^* = 3$, we find that our baseline *Regular Lager* and *Special/Craft* merge into one group. Overall, we label these three groups as "Light Beer", "Regular Lager", and "Craft Beer" groups. On the other hand, when $k^* = 5$, we obtain similar groups as when $k^* = 4$, but the "Regular Beer" groups splits into two. In terms of the structural parameters, we still find σ estimates that are consistent with utility maximization, however in both cases we cannot reject the null that they are outside the interval $(0, 1)$. For further details see Appendix C.2.2.

8.3.3 Classification with additional categorical variables

We perform a robustness exercise in which we augment the baseline specification with categorical variables describing beer type provided by NielsenIQ. We consider five categories: Light Beer, Regular/Lager, Stout and Porter, Ale, and Others.[25]

Including these categorical variables, the cross-validation procedure selects $k^* = 3$ groups, one fewer than in the baseline. As shown in Table A8, the three groups correspond roughly to *Regular Lager*, *Light Beer*, and *Craft* beer. Second-step estimates are reported in Table A9 and are consistent with utility maximization, yielding an average own-price elasticity of -2.343, somewhat more inelastic than the baseline but within the range of estimates in the

[25] The original data contains nine categories. We consolidate four smaller categories into Others and combine two categories of regular lager into one for simplicity. Results are similar using the nine disaggregated categories.

literature.

That fewer groups emerge when categorical variables are included is intuitive. When beer type categories are excluded from the classification step, the group-by-market fixed effects $\lambda_{k(j)m}$ must absorb not only the unobserved nest-specific demand shifters for which they are designed, but also the systematic variation in demand attributable to beer type. The classifier responds by forming additional groups, each with its own set of market fixed effects, until the residual variation is sufficiently small.

8.3.4 K-means Classification in Product Characteristics

We apply k-means to the same characteristics used in our baseline. The cross-validation procedure again selects $k^* = 4$ groups. Average product characteristics by group are reported in Table A10 and second-step estimates in Table A11. While the number of groups coincides with our baseline, the groups are less heterogeneous, both in conditional product characteristics and in implied substitution patterns. Exploiting market share variation recovers a nesting structure with more heterogeneous substitution patterns than proximity in characteristics space. Section 5.5 discusses when the two classifications coincide.

8.4 Comparison to other demand models

8.4.1 "Naive" Nested Logit

As a benchmark, we estimate a standard nested logit model using the ex-ante product classification provided by NielsenIQ. Specifically, we impose that products are grouped according to *beer type*, yielding nests such as regular/lager, light beer, ale, stout/porter, and a residual category for other types. This approach mirrors common empirical practice, in which researchers select a plausible partition based on observable product categories and estimate the model conditional on that grouping. We refer to this specification as the "naive" nested logit.

Following this naive approach, we find that the naive nested logit performs worse than the data-driven nested logit along several dimensions. First, when we implement the Bud Light boycott validation exercise, the naive specification predicts substantially less substitution toward the group to which Bud Light belongs (labeled *light beer*) compared to the substitution toward the data-driven light beer group using our out-of-sample exercise that

leverages the Bud Light boycott. Concretely, under the data-driven nested logit, products in the same group as Bud Light experience a large and statistically significant increase in demand following the boycott. The coefficient on *Treatment same group* $\times\, t = 2023$ is 0.422, indicating substantial substitution toward products within the same estimated nest. By contrast, the naive nested logit produces much weaker substitution patterns. The coefficient on *Naive light beer* $\times\, t = 2023$ is only 0.13, substantially smaller than the data-driven estimate.

This difference is confirmed in the direct comparison column. In this specification, we construct three treatment subgroups: products that belong only to the Light Beer group identified by the data-driven classification, products that belong only to the Light Beer group defined by NielsenIQ, and products that belong to both groups. We find that the substitution captured by the Naive Light Beer group is entirely driven by products that also belong to the data-driven Light Beer group. By contrast, products that belong only to the Naive Light Beer group do not experience any statistically significant demand response following the boycott.

These results indicate that the substitution patterns implied by the data-driven nesting structure more closely match the observed demand responses following the boycott. One possible explanation is that NielsenIQ classifies products as Light Beer based on low caloric content, which may arise either because the beer is truly light or because it is sold in smaller containers. As a result, the NielsenIQ classification may group together products that are not close substitutes from a demand perspective.

Table 6: Diff-in-Diff results of Bud Light boycott with Naive Nested Logit

Dependent Variable: Log Share										
	Data-driven NL				Naive NL				Interaction	
t=2023	-0.408***	(0.005)	-0.449***	(0.008)	-0.329***	(0.006)	-0.449***	(0.008)	-0.381***	(0.007)
Treatment same group	0.422***	(0.008)								
Regular beer × t=2023			0.173***	(0.012)						
Light beer × t=2023			0.463***	(0.010)						
Non-lager/craft × t=2023			-0.001	(0.013)						
Boutique craft × t=2023			0.018	(0.020)						
Naive light beer × t=2023					0.130***	(0.016)	0.250***	(0.017)		
Naive ale × t=2023							0.028**	(0.012)		
Naive other × t=2023							0.009	(0.043)		
Naive lager × t=2023							0.249***	(0.012)		
Naive stout × t=2023							-0.041*	(0.021)		
DD Light only									0.385***	(0.009)
Naive Light only									0.033	(0.024)
DD and Naive Light									0.422***	(0.012)
Observations	172778		172778		172778		172778		172778	
UPC-Market FE	YES		YES		YES		YES		YES	
Exclude Bud Light	YES		YES		YES		YES		YES	
Baseline group	Other groups		Outside option		Outside option		Outside option		Outside option	

Notes: *** p<0.01, ** p<0.05, * p<0.1

Each column reports OLS estimates from a difference-in-differences specification with log market share as the dependent variable, comparing 2022 to 2023 after the Bud Light boycott. Bud Light products are excluded from all columns. Columns (1)–(2) use the data-driven nesting structure ($k^* = 4$); Columns (3)–(4) use a naive nesting structure based on beer type (light beer, ale, lager, stout/porter, other). In Column (1), the treatment indicator equals one for products in Bud Light's data-driven group (light beer) interacted with $t = 2023$; the baseline comprises all other groups. In Column (2), group-specific post-boycott effects are estimated relative to the outside option. Columns (3) and (4) replicate this design using naive beer-type groups. Column (5) interacts data-driven and naive light-beer indicators to decompose substitution into products classified as light beer by only the data-driven method, only the naive method, or both. All specifications include UPC-by-market (state) fixed effects. Robust (HC1) standard errors are reported in parentheses.

Second, when examining the structural parameter estimates, we find that one group yields a negative $\hat{\sigma}_k$, which lies outside the range consistent with utility maximization. This provides empirical evidence that the nesting structure implied by the naive specification is rejected by the data.

Table 7: Estimation Results for Naive Nested Logit

	OLS	IV
log unit size	0.054 (0.005)***	0.686 (0.041)***
log upc quantity	0.059 (0.004)***	0.694 (0.042)***
log price	-0.039 (0.005)***	-0.753 (0.050)***
σ_{ale}	-0.029 (0.001)***	0.239 (0.012)***
$\sigma_{\text{light beer}}$	0.087 (0.001)***	0.348 (0.010)***
σ_{other}	-0.356 (0.002)***	-0.098 (0.015)***
$\sigma_{\text{regular/lager}}$	0.080 (0.001)***	0.322 (0.010)***
$\sigma_{\text{stout/porter}}$	-0.296 (0.001)***	0.040 (0.017)**
Mean Own-price Elasticity	0.360 (0.003)	-3.752 (0.104)
Mean Cross-price Elasticity	-0.001 (0.003)	0.007 (0.111)
Market Fixed Effects	✓	✓
IV Type		$GH_{\text{units}}, GH_{\text{size}}$
Number of Products	2806	
Number of Observations	64497	

Note: The table reports OLS and IV estimates of a nested logit demand model where products are grouped by beer type (regular beer/lager, light beer, ale, stout/porter, and other), yielding $k = 5$ nests. The dependent variable is $\log \mathbb{P}_{jm} - \log \mathbb{P}_{0m}$. Regressors include log unit size (ml), log UPC quantity, log price, and five group-specific nesting parameters (σ_g). The IV specification instruments for log price using Gandhi-Houde differentiation IVs and for the within-group log share using the exogenous share projection. Market fixed effects are included. Mean own-price and cross-price elasticities are computed from the IV estimates and weighted by national unit sales. $^{***}p < 0.01, ^{**}p < 0.05, ^{*}p < 0.1$.

8.4.2 Mixed Logit

In this section, we estimate a mixed logit model using PyBLP (Conlon and Gortmaker, 2020). Concretely, we specify that the utility of an individual i from purchasing product j in market m is given by

$$u_{ijm} = \delta_{jm} + \mu_{ijm} + \epsilon_{ijm}, \tag{21}$$

where ϵ_{ijm} follows the Type I extreme distribution. The common component of utility is given by

$$\delta_{jm} = \varphi_m + \beta_p \log \mathrm{p}_{jm} + \beta_s \log \mathrm{s}_j + \beta_n \log \mathrm{n}_j + \xi_{jm}, \tag{22}$$

and the individual-specific component μ_{ijm} is given by

$$\mu_{ijm} = (\sigma_p \nu_i) \log \mathrm{p}_{jm}, \tag{23}$$

where σ_p governs the variance of the idiosyncratic component. $\nu_i \sim N(0,1)$ follows a standard normal distribution. We estimate the parameters $\{\boldsymbol{\beta}, \sigma_p\}$ by solving the following GMM problem:[26]

$$\min_{\boldsymbol{\beta},\sigma_p} \quad \bar{g}\left(\boldsymbol{\beta},\sigma_p\right)' W \bar{g}\left(\boldsymbol{\beta},\sigma_p\right), \quad \bar{g}\left(\boldsymbol{\beta},\sigma_p\right) = \frac{1}{N}\sum_{j,m}\left(Z'_{jm}\xi_{jm}(\boldsymbol{\beta},\sigma_p)\right) \tag{24}$$

where $\boldsymbol{\beta} = [\beta_p, \beta_s, \beta_q]$ and Z_{jm} denote instruments. Specifically, we use the Gandhi-Houde instruments to make results as comparable as possible.[27] W is a weighting matrix whose initial values are set by the 2SLS estimate and then updated using a two-step optimal GMM weights.[28] The estimation results are reported in Table 8.

Table 8: Mixed Logit estimates

Panel A: Linear Coefficients	$\boldsymbol{\beta}$
log price	-3.775*** (0.090)
log unit size	3.077*** (0.062)
log unit quantity	3.222*** (0.062)
Panel B: Non-linear Coefficient	σ_p
log price	0.002*** (0.0003)
Own-price elasticity	-4.683 (2.863)
Cross-price elasticity	0.002 (0.005)
Market Fixed Effects	✓
IV Type	Gandhi-Houde
Number of Products	2806
Number of Observations	64497

Notes: Panel A displays the estimates of $\boldsymbol{\beta}$ in equation (24). Panel B displays the estimates of σ_p. Robust standard errors in parentheses.

With these coefficients, the average price elasticity is -4.683 with a standard deviation of 2.863 in the average market. For cross-price elasticities, we find an average value of 0.002 with standard deviation of 0.005. Despite being a less general model, for this particular application, we find that the data-driven NL model delivers more heterogeneity in substitution

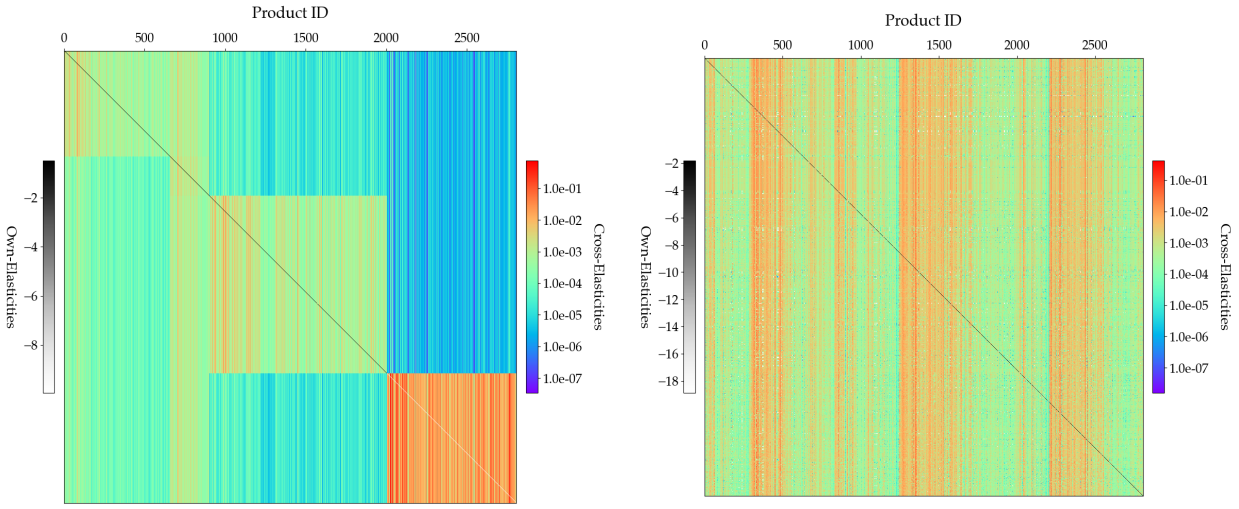
[26] In practice, we use the `slsqp` optimization routine with box contraints in PyBLP.

[27] We also estimate the previous model using optimal instruments as suggested by Conlon and Gortmaker (2020). However, using the same optimization routine and same initial value as in our baseline specification, we obtain an estimate for the variance coefficient that converges to zero.

[28] See Conlon and Gortmaker (2020) for details.

patterns across products with450% more variation in cross-price elasticities.

Figure 8: Price elasticity matrices



(a) Nested logit ($k^* = 4$) (b) Mixed Logit

Notes: The figures above display the price elasticity matrices estimated using the nested logit specification and the BLP procedure. Diagonal cells each matrix denote products' own-price elasticities, whereas the off-diagonal cells denote cross-price elasticities. Concretely, the value in row i, column j denotes $\frac{\partial \log \mathbb{P}_i}{\partial \log p_j}$.

We can also compare the full matrix of price elasticities for both models, as depicted in Figure 8. In the figure, products are ordered first by group and then by UPC identifier, and elasticities are averaged across markets. The left panel plots the estimated elasticities from the data-driven nested logit model, while the right panel reports elasticities from the mixed logit model.

The data-driven nested logit displays substantial variation in substitution patterns with a clear block-diagonal structure, reflecting stronger substitution within nests than across nests. By contrast, the mixed logit produces elasticity patterns that closely resemble those of a simple logit model, with relatively little variation across products.

One possible explanation is that the market shares of many products are extremely small and exhibit limited variation across markets, which makes it difficult to identify substantial heterogeneity in price sensitivities. As a result, $1 - \hat{\mathbb{P}}_{jm}(p^i)$ is close to one for most products, implying that the elasticity expression—which involves an integral over the distribution of random coefficients in the mixed logit—is well approximated by the simple logit formula.

Overall, we conclude that despite being in theory a somewhat restrictive model, the data-driven nested logit model is able to empirically capture very heterogeneous substitution

patterns when applied to this particular setting.

9 Conclusion and Future Work

This paper develops a data-driven approach to estimating nested demand models without imposing an *a priori* grouping of products. Exploiting the log-linear structure of nested logit market shares, we propose a two-step procedure that first recovers the nesting structure by clustering model-implied demand curves and then estimates preference parameters conditional on the estimated nests. We establish consistency of the classification as the number of products and markets grows, allowing for endogenous prices, and show that the resulting structural estimator is asymptotically equivalent to an infeasible estimator that knows the true nesting structure. Consequently, inference on preference parameters can proceed without accounting for classification uncertainty.

Applying our method to the U.S. beer market, we identify four economically meaningful nests with substantial heterogeneity in substitution patterns that cannot be fully explained by observable product characteristics. Using the 2023 Bud Light boycott as an out-of-sample validation, we show that the estimated nests capture observed substitution patterns more accurately than conventional groupings based on product categories. Moreover, our model delivers richer heterogeneity in substitution patterns than a mixed logit benchmark while preserving computational tractability. These findings highlight the importance of recovering economically relevant product groupings from observed demand rather than imposing them *a priori*.

Several questions remain open. One direction is to extend our framework to accommodate unobserved heterogeneity in consumer preferences and nonlinear specifications of indirect utility. Another promising avenue is to investigate whether increasingly rich hierarchical nesting structures can provide discrete approximations to general forms of dependence in idiosyncratic utility shocks across products. More broadly, our findings motivate the development of alternative methods to recover unobserved drivers of demand from observed choices. Ultimately, these approaches may contribute to a more general framework for data-driven model selection in demand estimation, in which both the structure of heterogeneity and the appropriate demand specification are learned from the data.

References

Ahlfeldt, G. M., Redding, S. J., Sturm, D. M., and Wolf, N. (2015). The economics of density: Evidence from the Berlin Wall. *Econometrica*, 83(6):2127–2189.

Allcott, H., Kane, R., Maydanchik, M. S., Shapiro, J. S., and Tintelnot, F. (2024). The effects of "Buy American": Electric vehicles and the Inflation Reduction Act. NBER Working Paper 33032, National Bureau of Economic Research.

Allen, T. and Arkolakis, C. (2022). The welfare effects of transportation infrastructure improvements. *Review of Economic Studies*, 89(6):2911–2957.

Armstrong, T. B. (2016). Large market asymptotics for differentiated product demand estimators with economic models of supply. *Econometrica*, 84(5):1961–1980.

Asker, J. (2016). Diagnosing foreclosure due to exclusive dealing. *Journal of Industrial Economics*, 64(3):375–410.

Atalay, E., Frost, E., Sorensen, A., Sullivan, C., and Zhu, W. (2026). Scalable demand and markups. *Journal of Political Economy*, 134(8):2488–2522.

Athey, S., Ferguson, B., Gentzkow, M., and Schmidt, T. (2021). Estimating experienced racial segregation in US cities using large-scale GPS data. *Proceedings of the National Academy of Sciences*, 118(46):e2026160118.

Atkeson, A. and Burstein, A. (2008). Pricing-to-market, trade costs, and international relative prices. *American Economic Review*, 98(5):1998–2031.

Azar, J. A., Berry, S. T., and Marinescu, I. (2022). Estimating labor market power. NBER Working Paper 30365, National Bureau of Economic Research.

Bayer, P., Ferreira, F., and McMillan, R. (2007). A unified framework for measuring preferences for schools and neighborhoods. *Journal of Political Economy*, 115(4):588–638.

Berger, D., Herkenhoff, K., and Mongey, S. (2022). Labor market power. *American Economic Review*, 112(4):1147–1193.

Bernard, A. B., Dhyne, E., Magerman, G., Manova, K., and Moxnes, A. (2022). The origins of firm heterogeneity: A production network approach. *Journal of Political Economy*, 130(7):1765–1804.

Bernard, A. B., Moxnes, A., and Saito, Y. U. (2019). Production networks, geography, and firm performance. *Journal of Political Economy*, 127(2):639–688.

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

Berry, S., Eizenberg, A., and Waldfogel, J. (2016). Optimal product variety in radio markets. *RAND Journal of Economics*, 47(3):463–497.

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

Berry, S., Linton, O. B., and Pakes, A. (2004). Limit theorems for estimating the parameters of differentiated product demand systems. *Review of Economic Studies*, 71(3):613–654.

Björnerstedt, J. and Verboven, F. (2016). Does merger simulation work? Evidence from the Swedish analgesics market. *American Economic Journal: Applied Economics*, 8(3):125–164.

Bonhomme, S., Lamadon, T., and Manresa, E. (2022). Discretizing unobserved heterogeneity. *Econometrica*, 90(2):625–643.

Bonhomme, S. and Manresa, E. (2015). Grouped patterns of heterogeneity in panel data. *Econometrica*, 83(3):1147–1184.

Bordeu, O. (2026). Commuting infrastructure in fragmented cities. Working paper.

Broda, C. and Weinstein, D. E. (2006). Globalization and the gains from variety. *Quarterly Journal of Economics*, 121(2):541–585.

Broda, C. and Weinstein, D. E. (2010). Product creation and destruction: Evidence and price implications. *American Economic Review*, 100(3):691–723.

Cardell, N. S. (1997). Variance components structures for the extreme-value and logistic distributions with application to models of heterogeneity. *Econometric Theory*, 13(2):185–213.

Chiong, K. X. and Shum, M. (2019). Random projection estimation of discrete-choice models with large choice sets. *Management Science*, 65(1):256–271.

Choi, J. and Okui, R. (2024). Latent group structure in linear panel data models with endogenous regressors. *arXiv preprint arXiv:2405.08687*.

Ciliberto, F. and Williams, J. W. (2014). Does multimarket contact facilitate tacit collusion? Inference on conduct parameters in the airline industry. *RAND Journal of Economics*, 45(4):764–791.

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

Couture, V., Dingel, J. I., Green, A. E., and Handbury, J. (2025). Demographic preferences and income segregation. NBER Working Paper 33386, National Bureau of Economic Research.

Couture, V., Gaubert, C., Handbury, J., and Hurst, E. (2024). Income growth and the distributional effects of urban spatial sorting. *Review of Economic Studies*, 91(2):858–898.

Davis, D. R., Dingel, J. I., Monras, J., and Morales, E. (2019). How segregated is urban consumption? *Journal of Political Economy*, 127(4):1684–1738.

Deb, S., Eeckhout, J., Patel, A., and Warren, L. (2024). Walras–Bowley lecture: Market power and wage inequality. *Econometrica*, 92(3):603–636.

DellaVigna, S. and Gentzkow, M. (2019). Uniform pricing in U.S. retail chains. *Quarterly Journal of Economics*, 134(4):2011–2084.

Dhyne, E., Kikkawa, A. K., Mogstad, M., and Tintelnot, F. (2021). Trade and domestic production networks. *Review of Economic Studies*, 88(2):643–668.

Dingel, J. I. and Tintelnot, F. (2026). Spatial economics for granular settings. *Econometrica*, 94(2):407–464.

Donaldson, D. (2018). Railroads of the Raj: Estimating the impact of transportation infrastructure. *American Economic Review*, 108(4-5):899–934.

Döpper, H., MacKay, A., Miller, N. H., and Stiebale, J. (2025). Rising markups and the role of consumer preferences. *Journal of Political Economy*, 133(8):2462–2505.

Eaton, J. and Kortum, S. (2002). Technology, geography, and trade. *Econometrica*, 70(5):1741–1779.

Einav, L. (2007). Seasonality in the US motion picture industry. *RAND Journal of Economics*, 38(1):127–145.

Ershov, D., Laliberté, J.-W., Marcoux, M., and Orr, S. (2025). Estimating complementarity with large choice sets: An application to mergers. *RAND Journal of Economics*, 56(4):689–707.

Fajgelbaum, P., Grossman, G. M., and Helpman, E. (2011). Income distribution, product quality, and international trade. *Journal of Political Economy*, 119(4):721–765.

Fogel, J., Modenesi, B., and Scuderi, B. (2025). Market definition bias in studies of labor market power. Work in progress.

Fosgerau, M., Monardo, J., and De Palma, A. (2024). The inverse product differentiation logit model. *American Economic Journal: Microeconomics*, 16(4):329–370.

Freyberger, J. (2015). Asymptotic theory for differentiated products demand models with many markets. *Journal of Econometrics*, 185(1):162–181.

Gandhi, A. and Houde, J.-F. (2019). Measuring substitution patterns in differentiated-products industries. NBER Working Paper 26375, National Bureau of Economic Research.

Goldberg, P. K. (1995). Product differentiation and oligopoly in international markets: The case of the US automobile industry. *Econometrica*, 63(4):891–951.

Goldberg, P. K. and Hellerstein, R. (2013). A structural approach to identifying the sources of local currency price stability. *Review of Economic Studies*, 80(1):175–210.

Grennan, M. (2013). Price discrimination and bargaining: Empirical evidence from medical devices. *American Economic Review*, 103(1):145–177.

Hahn, J. and Moon, H. R. (2010). Panel data models with finite number of multiple equilibria. *Econometric Theory*, 26(3):863–881.

Hausman, J., Leonard, G., and Zona, J. D. (1994). Competitive analysis with differentiated products. *Annales d'Économie et de Statistique*, 34:143–157.

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

Hellerstein, R. (2008). Who bears the cost of a change in the exchange rate? Pass-through accounting for the case of beer. *Journal of International Economics*, 76(1):14–32.

Hitsch, G. J., Hortaçsu, A., and Lin, X. (2021). Prices and promotions in U.S. retail markets. *Quantitative Marketing and Economics*, 19(3):289–368.

Hortacsu, A., Lieber, J., Monardo, J., and de Paula, A. (2023). Estimating nesting structures. Ongoing work.

Hottman, C. J., Redding, S. J., and Weinstein, D. E. (2016). Quantifying the sources of firm heterogeneity. *Quarterly Journal of Economics*, 131(3):1291–1364.

Jaravel, X. (2019). The unequal gains from product innovations: Evidence from the U.S. retail sector. *Quarterly Journal of Economics*, 134(2):715–783.

Khandelwal, A. (2010). The long and short (of) quality ladders. *Review of Economic Studies*, 77(4):1450–1476.

Konovalov, A. and Sándor, Z. (2010). On price equilibrium with multi-product firms. *Economic Theory*, 44(2):271–292.

Kovach, M. and Tserenjigmid, G. (2022). Behavioral foundations of nested stochastic choice and nested logit. *Journal of Political Economy*, 130(9):2411–2461.

Kreindler, G. E. and Miyauchi, Y. (2023). Measuring commuting and economic activity inside cities with cell phone records. *Review of Economics and Statistics*, 105(4):899–909.

Lamadon, T., Mogstad, M., and Setzler, B. (2022). Imperfect competition, compensating differentials, and rent sharing in the US labor market. *American Economic Review*, 112(1):169–212.

Lu, Z., Shi, X., and Tao, J. (2023). Semi-nonparametric estimation of random coefficients logit model for aggregate demand. *Journal of Econometrics*, 235(2):2245–2265.

McFadden, D. (1978). Modelling the choice of residential location. In *Spatial Interaction Theory and Planning Models*. North Holland.

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

Miller, N. H., Sheu, G., and Weinberg, M. C. (2021). Oligopolistic price leadership and mergers: The United States beer industry. *American Economic Review*, 111(10):3123–3159.

Miller, N. H. and Weinberg, M. C. (2017). Understanding the price effects of the MillerCoors joint venture. *Econometrica*, 85(6):1763–1791.

Miyauchi, Y., Nakajima, K., and Redding, S. J. (2025). The economics of spatial mobility: Theory and evidence using smartphone data. *Quarterly Journal of Economics*, 140(4):2507–2570.

Moon, H. R. and Weidner, M. (2015). Linear regression for panel with unknown number of factors as interactive fixed effects. *Econometrica*, 83(4):1543–1579.

Nocke, V. and Schutz, N. (2018). Multiproduct-firm oligopoly: An aggregative games approach. *Econometrica*, 86(2):523–557.

Parry, I. W. and Small, K. A. (2005). Does Britain or the United States have the right gasoline tax? *American Economic Review*, 95(4):1276–1289.

Pinkse, J. and Slade, M. E. (2004). Mergers, brand competition, and the price of a pint. *European Economic Review*, 48(3):617–643.

Redding, S. J. and Weinstein, D. E. (2020). Measuring aggregate price indices with taste shocks: Theory and evidence for CES preferences. *Quarterly Journal of Economics*, 135(1):503–560.

Rojas, C. (2008). Price competition in U.S. brewing. *Journal of Industrial Economics*, 56(1):1–31.

Romeo, C. J. (2016). Incorporating prior information into a GMM objective for mixed logit demand systems. *Journal of Industrial Economics*, 64(2):336–363.

Slade, M. E. (2004). Market power and joint dominance in UK brewing. *Journal of Industrial Economics*, 52(1):133–163.

Su, L., Shi, Z., and Phillips, P. C. B. (2016). Identifying latent structures in panel data. *Econometrica*, 84(6):2215–2264.

Train, K. E. (2009). *Discrete Choice Methods with Simulation*. Cambridge University Press, 2nd edition.

Verboven, F. (1996). International price discrimination in the European car market. *RAND Journal of Economics*, 27(2):240–268.

Verhoogen, E. A. (2008). Trade, quality upgrading, and wage inequality in the Mexican manufacturing sector. *Quarterly Journal of Economics*, 123(2):489–530.

Vershynin, R. (2018). *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge.

Yakovlev, P. A. and Guessford, W. P. (2013). Alcohol consumption and political ideology: What's party got to do with it? *Journal of Wine Economics*, 8(3):335–354.

A Appendix

A.1 Substitution patterns and elasticity of demand

For simplicity, let's assume that $\delta_{jm} = \beta_p \log p_{jm} + \beta_x \log X_{jm} + \xi_{jm}$.
Own-price elasticities are given by:

$$\mathcal{E}_j^j = \frac{\beta_p}{\sigma_{k(j)}}\Big(1 - \sigma_{k(j)}\mathbb{P}_{jm} - (1-\sigma_{k(j)})\mathbb{P}_{j|k(j),m}\Big)$$

For cross-price elasticites, recall that if product $q \in B_{k'} \neq B_k$

$$\begin{aligned}\frac{\partial}{\partial p_q}\mathbb{P}_j &= -\frac{e^{\frac{\delta_j}{\sigma_k}}\big(\sum_{d\in B_k} e^{\frac{\delta_d}{\sigma_k}}\big)^{\sigma_k - 1}\sigma_{k'}\big(\sum_{d\in B_{k'}} e^{\frac{\delta_d}{\sigma_{k'}}}\big)^{\sigma_{k'}-1}\Big(-\frac{\beta_p}{\sigma_{k'}}e^{\frac{\delta_q}{\sigma_{k'}}}\Big)}{\Big(\sum_{l=1}^K\big(\sum_{d\in B_l} e^{\frac{\delta_d}{\sigma_l}}\big)^{\sigma_l}\Big)^2} \\ &= -\beta_p \mathbb{P}_j \mathbb{P}_q.\end{aligned}$$

On the other hand, if product $q \in B_k$

$$\begin{aligned}\frac{\partial}{\partial p_q}\mathbb{P}_j &= \frac{e^{\frac{\delta_j}{\sigma_k}}(\sigma_k - 1)\big(\sum_{d\in B_k} e^{\frac{\delta_d}{\sigma_k}}\big)^{\sigma_k - 2}\Big(-\frac{\beta_p}{\sigma_k}e^{\frac{\delta_q}{\sigma_k}}\Big)}{\sum_{l=1}^K\big(\sum_{d\in B_l} e^{\frac{\delta_d}{\sigma_l}}\big)^{\sigma_l}} \\ &\quad - \frac{e^{\frac{\delta_j}{\sigma_k}}\big(\sum_{d\in B_k} e^{\frac{\delta_d}{\sigma_k}}\big)^{\sigma_k - 1}\sigma_k\big(\sum_{d\in B_k} e^{\frac{\delta_d}{\sigma_k}}\big)^{\sigma_k-1}\Big(-\frac{\beta_p}{\sigma_k}e^{\frac{\delta_q}{\sigma_k}}\Big)}{\Big(\sum_{l=1}^K\big(\sum_{d\in B_l} e^{\frac{\delta_d}{\sigma_l}}\big)^{\sigma_l}\Big)^2} \\ &= -\beta_p\Big((\sigma_k - 1)\frac{1}{\sigma_k}\mathbb{P}_j\mathbb{P}_{q|k} + \mathbb{P}_j\mathbb{P}_q\Big).\end{aligned}$$

Therefore, price elasticities are given by

$$\mathcal{E}_q^j = \frac{\frac{\partial}{\partial p_q}\mathbb{P}_j}{\frac{\mathbb{P}_j}{p_q}} = \begin{cases} \frac{\beta_p}{\sigma_{k(j)}}\Big(1 - \sigma_{k(j)}\mathbb{P}_{jm} - (1-\sigma_{k(j)})\mathbb{P}_{j|k(j),m}\Big) & \text{if } j = q \\ -\beta_p\mathbb{P}_q & \text{if } q \in B_{k'} \neq B_k \\ -\beta_p\Big((\sigma_k - 1)\frac{1}{\sigma_k}\mathbb{P}_{q|k} + \mathbb{P}_q\Big) & \text{if } q \in B_k \end{cases}$$

A.2 Bias under mis-specified nests

There are two true nests, k_1 and k_2 with a common σ. The econometrician places all inside goods in a single nest $k_0 = A \cup B$. Let $g(j) \in \{A, B\}$ denote the true nest of product j. Following Berry (1994), the true model is

$$\log\left(\frac{\mathbb{P}_{jm}}{s_{0m}}\right) = x_{jm}\beta + (1-\sigma)\log \mathbb{P}_{j|k(j),m} + \xi_{jm}.$$

The econometrician uses $s_{j|k_0,m}$ as a covariate instead of $s_{j|k(j),m}$.

It follows that

$$\log\left(\frac{\mathbb{P}_{jm}}{\mathbb{P}_{0m}}\right) = x_{jm}\beta + \sigma \log \mathbb{P}_{j|k_0,m} + (1-\sigma)\left(-\log \mathbb{P}_{k(j)|k_0,m}\right) + \xi_{jm},$$

as $\log \mathbb{P}_{j|k(j),m} = \log \mathbb{P}_{j|k_0,m} - \log \mathbb{P}_{k(j)|k_0,m}$. Therefore, the misspecification reduces to one omitted regressor, $-\log \mathbb{P}_{k(j)|k_0,m}$.

Let $W_{jm} = \left(x_{jm}, \log \mathbb{P}_{j|G,m}\right)'$ and $\theta = (\beta, \sigma)'$. Let W stack W_{jm} across products and markets. Suppose $\mathbb{E}[W_{jm}\xi_{jm}] = 0$. The OLS estimator satisfies

$$\operatorname{plim} \hat{\theta} = \theta + (1-\sigma)\operatorname{plim}\left(W'W\right)^{-1}W'[-\log \mathbb{P}_{g(j)|G,m}]_{j,m}.$$

Note that the bias is only zero if $\sigma = 1$ or if there is not correlation between $W_{jm} = \left(x_{jm}, \log \mathbb{P}_{j|G,m}\right)'$. The former case the model is in fact a logit model with no relevant grouping structure. The latter case is unlikely as

$$\log \mathbb{P}_{g(j)|G,m} = \log \frac{\sum_{j\in g}\exp(\beta x_{jm})}{\sum_{j\in G}\exp(\beta x_{jm})},$$

which contains x_{jm}, and therefore is correlated with the regressor W_{jm}.

For this example and more generally, the sign of the bias depends on the covariance of ω with $\log s_{j|G,m}$. We can write

$$\operatorname{Cov}\left(\log s_{j|G,m},\ \omega\right) = -\operatorname{Cov}\left(\log s_{j|g(j),m},\ \log s_{g(j)|G,m}\right) - \operatorname{Var}\left(\log s_{g(j)|G,m}\right).$$

The second term is negative, but the first term is positive when larger nests contain products with lower within-nest shares. The relative size of the two terms depends on the data-generating process.

Finally, we also illustrate the bias with a simulation. We draw $5{,}000$ markets. In each market, the number of products in each true nest is uniform on $\{1, \ldots, 5\}$. We set $x_{jm} \sim N(0,1)$, $\beta = 1$, $\sigma = 0.5$, and $\xi_{jm} = 0$. The mean utility of the outside good is zero. For simplicity, we take $\xi_{jm} = 0$, so OLS under the true nests recovers the parameters exactly. Table A1 reports the estimates.

Table A1: OLS estimates under true and merged nests

	$\hat{\beta}$	$\hat{\sigma}$
True nests (A, B)	1.00	0.50
Merged nest (G)	0.80	0.57
Omitted-variable formula	0.80	0.57

The merged-nest estimates match the omitted-variable formula. In this design, $\hat{\beta}$ is biased downward. $\hat{\sigma}$ is biased upward.

A.3 Multi-Product Firms and Markups with Many Products

A common argument holds that price endogeneity vanishes as the number of products grows. Each product becomes small, so markups converge to constants and prices move only with costs. Armstrong (2016) makes a version of this argument under Bertrand–Nash pricing. He shows that instruments built from the characteristics of rival products then become weak. This appendix derives markups under nested logit with multi-product firms. Markups converge to constants only when the share of every firm within its nest vanishes. When some firm keeps a non-vanishing share, markups differ across firms in the limit.

We consider a single market and drop the market index. For ease of exposition, price enters utility in levels, $\delta_j = \beta_p p_j + \beta_x' \mathbf{x}_j + \xi_j$ with $\beta_p < 0$. The same results hold when price enters utility in logs.

Firm f sells the products in $\mathcal{F}_f \equiv \{j : f(j) = f\}$. Let $\mathbb{P}_k \equiv \sum_{l \in k} \mathbb{P}_l$ be the share of nest k. For firm f, let $\mathbb{P}_{fk} \equiv \sum_{l \in \mathcal{F}_f \cap k} \mathbb{P}_l$ be its share in nest k and $\mathbb{P}_{f|k} \equiv \mathbb{P}_{fk}/\mathbb{P}_k$ its within-nest share.

First-order conditions. Firm f chooses $\{p_j\}_{j \in \mathcal{F}_f}$ to maximize $\sum_{j \in \mathcal{F}_f} (p_j - mc_j)\mathbb{P}_j$. Writing $\eta_j \equiv p_j - mc_j$ for the markup, the first-order condition for p_j is

$$\mathbb{P}_j + \sum_{l \in \mathcal{F}_f} \eta_l \frac{\partial \mathbb{P}_l}{\partial p_j} = 0. \tag{25}$$

Under nested logit, the share derivatives depend on whether product l is product j itself, another product in the nest of j, or a product in another nest:

$$\frac{\partial \mathbb{P}_l}{\partial p_j} = \beta_p \times \begin{cases} \dfrac{\mathbb{P}_j}{\sigma_{k(j)}}\big[1-(1-\sigma_{k(j)})\mathbb{P}_{j|k(j)}-\sigma_{k(j)}\mathbb{P}_j\big] & l=j, \\ -\dfrac{\mathbb{P}_l}{\sigma_{k(j)}}\big[(1-\sigma_{k(j)})\mathbb{P}_{j|k(j)}+\sigma_{k(j)}\mathbb{P}_j\big] & l\in k(j),\ l\neq j, \\ -\mathbb{P}_l\mathbb{P}_j & l\notin k(j). \end{cases} \tag{26}$$

Lemma 1 (Markups). *Under Bertrand–Nash with any ownership partition, the markup of product j is*

$$\eta_j = \eta_{fk} \equiv -\frac{\sigma_k}{\beta_p}\,\frac{1}{(1-\Lambda_f)\big(1-(1-\sigma_k)\mathbb{P}_{f|k}\big)}, \qquad \Lambda_f \equiv \sum_{k'}\frac{\sigma_{k'}\mathbb{P}_{fk'}}{1-(1-\sigma_{k'})\mathbb{P}_{f|k'}}, \tag{27}$$

where $k=k(j)$ and $f=f(j)$. The markup depends on j only through $f(j)$ and $k(j)$.

Proof. Let $T_{fk}\equiv\sum_{l\in\mathcal{F}_f\cap k}\eta_l\mathbb{P}_l$ and $T_f\equiv\sum_{k'}T_{fk'}$. Substitute (26) into (25), divide by $-\beta_p\mathbb{P}_j/\sigma_k$, and use $\mathbb{P}_{j|k}/\mathbb{P}_j = 1/\mathbb{P}_k$:

$$\begin{aligned} 0 &= -\frac{\sigma_k}{\beta_p} - \eta_j\big[1-(1-\sigma_k)\mathbb{P}_{j|k}-\sigma_k\mathbb{P}_j\big] + \Big[\frac{1-\sigma_k}{\mathbb{P}_k}+\sigma_k\Big]\sum_{l\in\mathcal{F}_f\cap k,\,l\neq j}\eta_l\mathbb{P}_l + \sigma_k\sum_{l\in\mathcal{F}_f\setminus k}\eta_l\mathbb{P}_l \\ &= -\frac{\sigma_k}{\beta_p} - \eta_j + \Big[\frac{1-\sigma_k}{\mathbb{P}_k}+\sigma_k\Big]\sum_{l\in\mathcal{F}_f\cap k}\eta_l\mathbb{P}_l + \sigma_k\sum_{l\in\mathcal{F}_f\setminus k}\eta_l\mathbb{P}_l \\ &= -\frac{\sigma_k}{\beta_p} - \eta_j + \frac{1-\sigma_k}{\mathbb{P}_k}\,T_{fk} + \sigma_k T_f. \end{aligned}$$

The second line moves the $l=j$ term into the first sum.

The last line depends on j only through f and k. So all products of firm f in nest k have the same markup, η_{fk}. Then $T_{fk}=\eta_{fk}\mathbb{P}_{fk}$ and $T_f=\sum_{k'}\eta_{fk'}\mathbb{P}_{fk'}$. Substituting T_{fk} and solving for η_{fk},

$$\begin{aligned} \eta_{fk} &= -\frac{\sigma_k}{\beta_p} + (1-\sigma_k)\mathbb{P}_{f|k}\,\eta_{fk} + \sigma_k T_f, \\ \eta_{fk} &= \frac{\sigma_k}{1-(1-\sigma_k)\mathbb{P}_{f|k}}\Big(-\frac{1}{\beta_p}+T_f\Big). \end{aligned}$$

This holds for every nest. Multiply by $\mathbb{P}_{fk'}$ and sum over nests k'. By the definition of Λ_f,

$$T_f = \sum_{k'} \eta_{fk'} \mathbb{P}_{fk'} = \Big(-\frac{1}{\beta_p} + T_f\Big) \sum_{k'} \frac{\sigma_{k'} \mathbb{P}_{fk'}}{1 - (1 - \sigma_{k'}) \mathbb{P}_{f|k'}} = \Big(-\frac{1}{\beta_p} + T_f\Big) \Lambda_f.$$

Since $1 - (1 - \sigma_{k'}) \mathbb{P}_{f|k'} \geq \sigma_{k'}$, we have $\Lambda_f \leq \sum_{k'} \mathbb{P}_{fk'} < 1$. So we can solve for T_f:

$$T_f = -\frac{\Lambda_f}{\beta_p (1 - \Lambda_f)}.$$

Substituting T_f into the expression for η_{fk} gives (27):

$$\begin{aligned} \eta_{fk} &= \frac{\sigma_k}{1 - (1 - \sigma_k) \mathbb{P}_{f|k}} \Big(-\frac{1}{\beta_p} - \frac{\Lambda_f}{\beta_p (1 - \Lambda_f)}\Big) \\ &= -\frac{\sigma_k}{\beta_p} \frac{1}{1 - (1 - \sigma_k) \mathbb{P}_{f|k}} \cdot \frac{(1 - \Lambda_f) + \Lambda_f}{1 - \Lambda_f} \\ &= -\frac{\sigma_k}{\beta_p} \frac{1}{(1 - \Lambda_f)\big(1 - (1 - \sigma_k) \mathbb{P}_{f|k}\big)}. \end{aligned}$$

□

For plain logit, $\sigma_k = 1$ for all k, and (27) reduces to $\eta_j = -1/[\beta_p (1 - \mathbb{P}_{f(j)})]$, where $\mathbb{P}_f \equiv \sum_{l \in \mathcal{F}_f} \mathbb{P}_l$ is the total share of firm f. This is the common markup within a firm of Konovalov and Sándor (2010). Lemma 1 extends it to nested logit, with a common markup within firm and nest.[29]

Revisiting Armstrong (2016). Let the instruments collect own characteristics and functions of rivals' characteristics, the "BLP instruments." Theorem 1 of Armstrong (2016) shows that the IV estimator based on these instruments is inconsistent if markups converge to a constant faster than $J^{-1/2}$, uniformly over products. Rivals' characteristics affect prices only through markups and, if markups converge to a constant, the instruments lose their correlation with price. Under nested logit, the constant can differ by nest, so the condition becomes

$$\sqrt{J} \max_j \big|\eta_j - b_{k(j)}\big| \to_p 0 \quad \text{for some constants } b_1, \ldots, b_K.$$

Lemma 1 shows that, within a nest, markups differ across firms only through their shares.

[29] A common markup within a nest holds for the broader class of demand systems in Nocke and Schutz (2018).

Differences across nests do not matter for this condition, since b_k can differ by nest. The within-nest differences vanish when every firm has a vanishing share (Remark A.1). They do not vanish when a firm keeps a non-vanishing share. Let $\mathbb{P}_{f|k} \geq \underline{s} > 0$ for firm f, and let firm f' in nest k have vanishing shares in every nest. Since $\Lambda_f \geq 0$ and $1/(1-x) \geq 1+x$, (27) gives, for products j of f and j' of f' in nest k,

$$\eta_j - \eta_{j'} \;\geq\; -\frac{\sigma_k(1-\sigma_k)}{\beta_p}\,\underline{s} + o_p(1). \tag{28}$$

The gap does not depend on J, so the condition of Armstrong (2016) fails.

Suppose instead that every firm has vanishing within-nest shares, $\max_{f,k} \mathbb{P}_{f|k} = \mathcal{O}_p(J^{-1})$. This holds with single-product firms. It also holds with firms of bounded size when no product keeps a non-vanishing share of its nest. Then $\Lambda_f = \mathcal{O}_p(J^{-1})$ for every firm, and by (27),

$$\eta_j = -\frac{\sigma_{k(j)}}{\beta_p} + \mathcal{O}_p(J^{-1})$$

uniformly over products. This is the case studied by Armstrong (2016). His condition holds with $b_k = -\sigma_k/\beta_p$, and identification of β_p has to come from cost shifters.

In our application, the main brewers hold large within-nest shares in every nest, and every nest also contains products of smaller brewers, so the critique of Armstrong (2016) is unlikely to apply.

Remark A.1 (Sufficient condition for non-vanishing shares). Note that $\mathbb{P}_{f|k} \geq \underline{s}$ holds under two mild conditions. First, the firm owns a non-vanishing fraction of the products of nest k, $J_{fk}/J_k \geq \underline{\pi}_{f|k} > 0$, where J_{fk} is the number of products of firm f in nest k. Second, mean utilities are uniformly bounded, $|\delta_j| \leq \bar{\delta}$ for all j and J. The second condition says that no product is negligible relative to the others in its nest. Under it, every within-nest share satisfies $e^{-2\bar{\delta}/\sigma_k}/J_k \leq \mathbb{P}_{j|k} \leq e^{2\bar{\delta}/\sigma_k}/J_k$, so

$$\mathbb{P}_{f|k} = \sum_{l \in \mathcal{F}_f \cap k} \mathbb{P}_{l|k} \;\geq\; e^{-2\bar{\delta}/\sigma_k}\,\frac{J_{fk}}{J_k} \;\geq\; e^{-2\bar{\delta}/\sigma_k}\,\underline{\pi}_{f|k} > 0.$$

B Appendix for Statistical Theory of Estimator

This appendix contains the proofs of the results in Section 5. Appendix B.1 collects notation and standard inequalities. Appendix B.2 develops a theory of classification for a grouped-effects regression with a vanishing perturbation. Appendix B.3 maps the nested-logit first step with exogenous prices into the general model and verifies the relevant conditions. Appendix B.4 does the same with endogenous prices for a one-stage classifier that substitutes the price projection. Appendix B.5 treats the two-stage classifier of Section 4.1, allows price groupings that differ from the nests, and proves Theorem 5.1. Appendix B.6 defines the demeaned second-step estimator and states and proves its asymptotic distribution, Theorem B.4. Appendix B.7 proves Theorem 5.2, on the oracle estimator in levels, and Corollary 1.

B.1 Notation and technical facts

For an array $b = (b_m)_{m \le M}$ of scalars or vectors, $\|b\|$ is the Euclidean norm of the stacked vector, so $M^{-1}\|b\|^2 = M^{-1}\sum_m \|b_m\|^2$ is its mean square over markets. Write b_j for the array $(b_{jm})_{m \le M}$. For a symmetric matrix A, $\lambda_{\min}(A)$ is its smallest eigenvalue. For a scalar random variable Y and $r \ge 1$, $\|Y\|_r := (\mathrm{E}\,|Y|^r)^{1/r}$.

Subgaussian norm. For a scalar random variable X, $\|X\|_{\psi_2} := \inf\{t > 0 : \mathrm{E}\, e^{X^2/t^2} \le 2\}$; for a random vector V, $\|V\|_{\psi_2} := \sup_{\|u\|=1} \|u'V\|_{\psi_2}$. Two standard facts (Vershynin, 2018, §2.5):

(F1) *Moments.* If $\|X\|_{\psi_2} \le C$, then $\mathrm{E}\,|X|^r \le 2\,\Gamma(\frac{r}{2}+1)\,C^r$ for every $r \ge 1$.

(F2) *Exponential moments.* If $\|X\|_{\psi_2} \le C$, then $\mathrm{E}\, e^{\lambda|X|} \le 2e^{\lambda^2 C^2/4}$ for every $\lambda > 0$, since $\lambda|x| \le x^2/C^2 + \lambda^2 C^2/4$.

Sub-exponential norm and two inequalities. $\|X\|_{\psi_1} := \inf\{t > 0 : \mathrm{E}\, e^{|X|/t} \le 2\}$. Three standard facts (Vershynin, 2018, §2.6–2.8). If X, Y are subgaussian then $\|XY\|_{\psi_1} \le \|X\|_{\psi_2}\|Y\|_{\psi_2}$. If $W_1, \dots, W_M$ are independent, mean zero, with $\max_m \|W_m\|_{\psi_1} \le K$, then for every $t > 0$ (Bernstein)

$$\mathbb{P}\Big(\Big|\frac{1}{M}\sum_{m=1}^{M} W_m\Big| > t\Big) \le 2\exp\Big(-c_B\, M \min\Big\{\frac{t^2}{K^2}, \frac{t}{K}\Big\}\Big).$$

If $W_1, \dots, W_M$ are independent, mean zero, with $\|W_m\|_{\psi_2} \le K_m$, then for every $t > 0$ (Hoeffding)

$$\mathbb{P}\Big(\Big|\frac{1}{M}\sum_{m=1}^{M} W_m\Big| > t\Big) \le 2\exp\Big(-\frac{c_H\, M^2 t^2}{\sum_m K_m^2}\Big).$$

Here $c_B, c_H > 0$ are universal. Both right sides are $o(M^{-\delta})$ for every $\delta > 0$ whenever the exponent is at least $c'M$ for a fixed $c' > 0$.

Rosenthal's inequality. (Ros) If $\zeta_1, \dots, \zeta_n$ are independent with mean zero and values in $\mathbb{R}^d$, then for every $q \ge 2$ there is C_q with

$$\mathrm{E}\Big\|\sum_{j\le n}\zeta_j\Big\|^q \le C_q\Big(\sum_{j\le n}\mathrm{E}\,\|\zeta_j\|^q + \Big(\sum_{j\le n}\mathrm{E}\,\|\zeta_j\|^2\Big)^{q/2}\Big).$$

Within transformation. For any arrays a_{jm}, b_{jm}, since $\sum_m \nabla a_{jm} = 0$,

$$\sum_m \nabla a_{jm}\,\nabla b'_{jm} = \sum_m \nabla a_{jm}\, b'_{jm} = \sum_m a_{jm} b'_{jm} - M\,\bar{a}_j \bar{b}'_j. \tag{29}$$

B.2 Classification in a grouped-effects model with a vanishing perturbation

Model. We denote products by $j = 1, \dots, J$ and markets by $m = 1, \dots, M$, with $J, M \to \infty$. There is a fixed, but unknown, partition of products into groups, $k^0(j) \in \{1, \dots, K\}$ for $j \in \mathbb{N}$, with K fixed and known.

The data are $(y_{jm,J}, w_{jm,J})$, with $y_{jm,J} \in \mathbb{R}$ and $w_{jm,J} \in \mathbb{R}^d$. They satisfy

$$y_{jm,J} = w'_{jm,J}\vartheta^0_{k^0(j),J} + \alpha^0_{k^0(j)m,J} + e_{jm,J} + v_{jm,J}, \tag{30}$$

where $\vartheta^0_{1,J}, \dots, \vartheta^0_{K,J}$ are group-specific slopes and $\alpha^0_{km,J}$ are group-market centres. The perturbation $e_{jm,J}$ vanishes uniformly, and $v_{jm,J}$ is a product-market error. The index J denotes objects that may change with the number of products. Because both the coefficients and the data contain it, the model is a triangular array. For instance, in the nested logit model, equilibrium prices depend on the characteristics of all products in the market, so the distribution of $w_{jm,J}$ changes with J. In what follows, we drop the subscript J for the ease of exposition. Write $\vartheta^0 := (\vartheta^0_1, \dots, \vartheta^0_K)$.

Classifier. Let $\tilde{\mathcal{K}} := \{1, \dots, K\}^J$ be the set of candidate partitions, written $\tilde{k} = (\tilde{k}(1), \dots, \tilde{k}(J))$. Let Θ and $\mathcal{A}_J^{KM}$ be the parameter sets of (G1) below. The estimator is

$$(\hat{\vartheta}, \hat{\alpha}, \hat{k}) := \arg \min_{(\vartheta, \alpha, \tilde{k}) \in \Theta^K \times \mathcal{A}_J^{KM} \times \tilde{\mathcal{K}}} \sum_{j=1}^{J} \sum_{m=1}^{M} \left(y_{jm} - w_{jm}' \vartheta_{\tilde{k}(j)} - \alpha_{\tilde{k}(j)m} \right)^2. \tag{31}$$

Conditions. Conditions (G1)–(G9) are on the general model (30). The true partition is $k^0(\cdot)$ and candidate partitions are $\tilde{k} \in \tilde{\mathcal{K}}$. In the nested logit, $k^0(j) = k(j)$.

(G1) There are bounded sets $\Theta \subset \mathbb{R}^d$ and $\mathcal{A} \subset \mathbb{R}$, and non-random scalars $a_{m,J}$, $m \in \mathbb{N}$, $J \in \mathbb{N}$, such that

$$\{\vartheta_{k,J}^0 : \ k \le K, \ J \in \mathbb{N}\} \subset \Theta \qquad \text{and} \qquad \{\alpha_{km,J}^0 - a_{m,J} : \ k \le K, \ m \in \mathbb{N}, \ J \in \mathbb{N}\} \subset \mathcal{A}.$$

Write $\mathcal{A}_J^{KM} := \{\alpha : \ \alpha_{km} - a_{m,J} \in \mathcal{A} \text{ for all } k, m\}$ for the set of admissible centres.

(G2) The market blocks $\big((w_{jm}, v_{jm})_{j \le J}\big)$, $m = 1, \dots, M$, are independent across m.

(G3) $\mathrm{E}[w_{jm} v_{jm}] = 0$ for all j, m.

(G4) $\mathrm{E}[v_{jm}] = 0$ for all j, m; $\sup_{j,m,J,M} \|v_{jm}\|_{\psi_2} < \infty$; and there is $L < \infty$ with

$$\sup_{J,M} \ \frac{1}{J} \sum_{j=1}^{J} \sum_{i=1}^{J} \Big| \frac{1}{M} \sum_{m=1}^{M} \mathrm{E}[v_{jm} v_{im}] \Big| \le L.$$

(G5) $\sup_{j,m,J,M} \|w_{jm}\|_{\psi_2} < \infty$.

(G6) For every $\eta > 0$, $\mathbb{P}\big(\max_{j \le J} \frac{1}{M} \sum_{m=1}^{M} e_{jm}^2 > \eta \big) \to 0$.

(G7) There is $c > 0$ such that, for all $k \neq \ell$, all J, M large and all $j \le J$,

$$\frac{1}{M} \sum_{m=1}^{M} \mathrm{E}\left[\left(w_{jm}'(\vartheta_\ell^0 - \vartheta_k^0) + \alpha_{\ell m}^0 - \alpha_{km}^0 \right)^2 \right] \ \ge \ c.$$

(G8) For $\ell, \ell' \leq K$ and $\tilde{k} \in \tilde{\mathcal{K}}$ let $\hat{\rho}_{\ell\ell'}(\tilde{k})$ be the smallest eigenvalue of

$$\frac{1}{J}\sum_{j=1}^{J} \mathbb{1}\{k^0(j) = \ell,\ \tilde{k}(j) = \ell'\} \begin{pmatrix} \frac{1}{M}\sum_m w_{jm}w'_{jm} & \frac{w_{j1}}{\sqrt{M}} & \cdots & \frac{w_{jM}}{\sqrt{M}} \\ \frac{w'_{j1}}{\sqrt{M}} & 1 & & \\ \vdots & & \ddots & \\ \frac{w'_{jM}}{\sqrt{M}} & & & 1 \end{pmatrix}.$$

There is $\rho > 0$ such that $\mathbb{P}\big(\min_{\ell \leq K} \min_{\tilde{k} \in \tilde{\mathcal{K}}} \max_{\ell' \leq K} \hat{\rho}_{\ell\ell'}(\tilde{k}) \geq \rho\big) \to 1$.

(G9) $J = O(M^q)$ for some finite q.

Comments.

- The scalars $a_{m,J}$ in (G1) are a drift of the centres that is common to all groups in a market. They may be unbounded. In the nested logit, $a_{m,J} = -\log J$: the intercepts of all nests drift to $-\infty$ at the same rate. The drift does not affect the classifier, since adding a market-level term to the outcome shifts every group-market intercept by that term and leaves the minimizing partition and slopes unchanged. In the proofs the centres enter only through differences, $\alpha^0_{km} - \alpha_{\ell m}$, in which $a_{m,J}$ cancels, so only the bounded set $\mathcal{A}$ matters.

- The last part of (G4) bounds the dependence of v across products. (G4) and (G5) ask subgaussian tails. These are used because the per-product misclassification bound needs tails of $M^{-1}\sum_m \|w_{jm}\|^2$ and of $M^{-1}\sum_m v_{jm}w_{jm}$ that decay faster than any power of M.

- The perturbation e is new relative to BM15. (G6) only asks its root mean square over markets to vanish uniformly over products, so e may be product-specific and correlated with w and v.

- (G8) implies that every group holds a non-vanishing fraction of the products. Taking $\tilde{k} = k^0$, for $\ell' \neq \ell$ the set $\{j : k^0(j) = \ell,\ k^0(j) = \ell'\}$ is empty, so (G8) implies that $\hat{\rho}_{\ell\ell}(k^0) \geq \rho$ with probability tending to one. Note also that the lower-right block of that matrix is $\pi_{\ell J} I_M$, so $\pi_{\ell J} \geq \hat{\rho}_{\ell\ell}(k^0) \geq \rho$ for J large.

Misclassification. The proof that the probability of misclassification vanishes has four steps.

1. *Criterion.* Lemma 2 shows that the least-squares criterion is uniformly close to a criterion without the errors e and v. It uses (G1)–(G6).

2. *Slopes and centres.* The criterion without errors attains its minimum at the true slopes, centres and groups. Proposition 1 turns this into consistency of the estimated slopes and centres, after a relabelling of the groups. It uses (G8) for the design and (G7) for the separation of groups.

3. *One product.* Fix a product j. Suppose the estimated slopes and centres are close to the truth and the perturbation is small. Then the classifier assigns j to a wrong group only if one of four averages over markets is far from its mean (Lemma 3). Each of these events has probability $o(M^{-\delta})$ for every $\delta > 0$ (Lemma 5 below). This step uses (G1)–(G5) and (G7).

4. *All products.* A union bound over the J products gives $J\, o(M^{-\delta})$, which vanishes under (G9). Theorem B.1 adds two terms. The first is the probability that Step 2 fails. The second is the probability that the perturbation is large, which vanishes by (G6).

For $(\vartheta, \alpha, \tilde{k}) \in \Theta^K \times \mathcal{A}_J^{KM} \times \tilde{\mathcal{K}}$ define the sample objective of (31), normalized by JM, and its counterpart without the errors,

$$\begin{aligned} \check{Q}(\vartheta, \alpha, \tilde{k}) &:= \frac{1}{JM} \sum_{j=1}^{J} \sum_{m=1}^{M} \left(y_{jm} - w_{jm}' \vartheta_{\tilde{k}(j)} - \alpha_{\tilde{k}(j)m} \right)^2, \\ \tilde{Q}(\vartheta, \alpha, \tilde{k}) &:= \frac{1}{JM} \sum_{j=1}^{J} \sum_{m=1}^{M} D_{jm}(\vartheta, \alpha, \tilde{k})^2 + \frac{1}{JM} \sum_{j=1}^{J} \sum_{m=1}^{M} v_{jm}^2, \end{aligned} \tag{32}$$

where

$$D_{jm}(\vartheta, \alpha, \tilde{k}) := w_{jm}'(\vartheta_{k^0(j)}^0 - \vartheta_{\tilde{k}(j)}) + \alpha_{k^0(j)m}^0 - \alpha_{\tilde{k}(j)m}.$$

By (30), the residual is $y_{jm} - w_{jm}' \vartheta_{\tilde{k}(j)} - \alpha_{\tilde{k}(j)m} = v_{jm} + e_{jm} + D_{jm}$.

Lemma 2 (Uniform convergence of the criterion)**.** *Under (G1)–(G6),*

$$\sup_{(\vartheta, \alpha, \tilde{k})} \left| \check{Q}(\vartheta, \alpha, \tilde{k}) - \tilde{Q}(\vartheta, \alpha, \tilde{k}) \right| \to_p 0.$$

Proof. Expand the square of the residual $v_{jm} + e_{jm} + D_{jm}$ and subtract $\tilde{Q}$. For every $(\vartheta, \alpha, \tilde{k})$,

$$\check{Q} - \tilde{Q} = \underbrace{\frac{1}{JM} \sum_j \sum_m e_{jm}^2 + \frac{2}{JM} \sum_j \sum_m e_{jm}\big(v_{jm} + D_{jm}\big)}_{=:A} + \underbrace{\frac{2}{JM} \sum_j \sum_m v_{jm} D_{jm}}_{=:B}. \tag{33}$$

Both A and B depend on $(\vartheta, \alpha, \tilde{k})$.

By (G4), (G5) and (F1), $\mathrm{E}\,\|w_{jm}\|^4 \le C$ and $\mathrm{E}\,v_{jm}^4 \le C$ for some finite constant C. By (G1), $\|\vartheta\| \le C$, and $|\alpha^0_{k^0(j)m} - \alpha_{\tilde{k}(j)m}| \le C$ for every $(\alpha, \tilde{k})$, since both centres lie in $a_{m,J} + \mathcal{A}$ and $\mathcal{A}$ is bounded. Hence $|D_{jm}| \le C(1 + \|w_{jm}\|)$ for every $(\vartheta, \alpha, \tilde{k})$, for some other C. We show that $\sup_{(\vartheta,\alpha,\tilde{k})} |A|$ and $\sup_{(\vartheta,\alpha,\tilde{k})} |B|$ are $o_p(1)$.

Bound on A. Write $\bar{e} := \max_{j \le J} M^{-1/2}\|e_j\|$. The first summand of A equals $\frac{1}{J}\sum_j M^{-1}\|e_j\|^2$, which is at most $\bar{e}^2$. For the second summand,

$$\begin{aligned}
&\Big|\frac{2}{JM} \sum_j \sum_m e_{jm}\big(v_{jm} + D_{jm}\big)\Big| \\
&\quad\le 2\Big(\frac{1}{JM} \sum_j \sum_m e_{jm}^2\Big)^{1/2}\Big(\frac{1}{JM} \sum_j \sum_m \big(v_{jm} + D_{jm}\big)^2\Big)^{1/2} && \text{(Cauchy–Schwarz over } (j,m)) \\
&\quad= 2\Big(\frac{1}{J} \sum_j M^{-1}\|e_j\|^2\Big)^{1/2}\Big(\frac{1}{JM} \sum_j \sum_m \big(v_{jm} + D_{jm}\big)^2\Big)^{1/2} && (\|e_j\|^2 = \textstyle\sum_m e_{jm}^2) \\
&\quad\le 2\,\bar{e}\,\Big(\frac{1}{JM} \sum_j \sum_m \big(v_{jm} + D_{jm}\big)^2\Big)^{1/2} && (M^{-1/2}\|e_j\| \le \bar{e} \text{ for every } j) \\
&\quad\le 2\,\bar{e}\,\Big(\frac{1}{JM} \sum_j \sum_m \big(2v_{jm}^2 + 2D_{jm}^2\big)\Big)^{1/2} && ((x+y)^2 \le 2x^2 + 2y^2) \\
&\quad\le 2\,\bar{e}\,\Big(\frac{1}{JM} \sum_j \sum_m \big(2v_{jm}^2 + 2C^2(1 + \|w_{jm}\|)^2\big)\Big)^{1/2} && (|D_{jm}| \le C(1 + \|w_{jm}\|)) \\
&\quad\le 2\,\bar{e}\,\Big(\frac{1}{JM} \sum_j \sum_m \big(2v_{jm}^2 + 4C^2(1 + \|w_{jm}\|^2)\big)\Big)^{1/2} && ((1+a)^2 \le 2(1+a^2)) \\
&\quad\le C\,\bar{e}\,\Big(\frac{1}{JM} \sum_j \sum_m \big(1 + v_{jm}^2 + \|w_{jm}\|^2\big)\Big)^{1/2} && \text{(new } C).
\end{aligned}$$

The average on the right is $O_p(1)$ by Markov, and $\bar{e} = o_p(1)$ by (G6). Neither bound depends on $(\vartheta, \alpha, \tilde{k})$. So $\sup_{(\vartheta,\alpha,\tilde{k})} |A| = o_p(1)$.

Bound on B. Split B into two terms:

$$B = \underbrace{\frac{2}{JM}\sum_j\sum_m v_{jm}w'_{jm}(\vartheta^0_{k^0(j)} - \vartheta_{\tilde{k}(j)})}_{=:B_1} + \underbrace{\frac{2}{JM}\sum_j\sum_m v_{jm}\big(\alpha^0_{k^0(j)m} - \alpha_{\tilde{k}(j)m}\big)}_{=:B_2}.$$

Bound on B_1. Sum first over markets:

$$|B_1| \le \frac{C}{J}\sum_j \Big\|\frac{1}{M}\sum_m v_{jm}w_{jm}\Big\|.$$

The right side does not depend on $(\vartheta, \tilde{k})$. For each j,

$$\begin{aligned}
\mathrm{E}\Big\|\frac{1}{M}\sum_m v_{jm}w_{jm}\Big\|^2 &= \frac{1}{M^2}\sum_{m,m'}\mathrm{E}\left[v_{jm}v_{jm'}\,w'_{jm}w_{jm'}\right] && \text{(expand the square)}\\
&= \frac{1}{M^2}\sum_{m\neq m'}\mathrm{E}[v_{jm}w_{jm}]'\,\mathrm{E}[v_{jm'}w_{jm'}] && \text{(independence, (G2))}\\
&\quad + \frac{1}{M^2}\sum_m \mathrm{E}\left[v_{jm}^2\|w_{jm}\|^2\right] &&\\
&= \frac{1}{M^2}\sum_m \mathrm{E}\left[v_{jm}^2\|w_{jm}\|^2\right] && (\mathrm{E}[v_{jm}w_{jm}] = 0,\ \text{(G3)})\\
&\le \frac{1}{M^2}\sum_m \big(\mathrm{E}\,v_{jm}^4\big)^{1/2}\big(\mathrm{E}\,\|w_{jm}\|^4\big)^{1/2} && \text{(Cauchy–Schwarz)}\\
&\le \frac{C}{M} && \text{(fourth moments, (G4), (G5), (F1)).}
\end{aligned}$$

Hence

$$\begin{aligned}
\mathrm{E}\Big[\frac{C}{J}\sum_j\Big\|\frac{1}{M}\sum_m v_{jm}w_{jm}\Big\|\Big] &= \frac{C}{J}\sum_j\mathrm{E}\Big\|\frac{1}{M}\sum_m v_{jm}w_{jm}\Big\| && \text{(linearity)}\\
&\le \frac{C}{J}\sum_j\Big(\mathrm{E}\Big\|\frac{1}{M}\sum_m v_{jm}w_{jm}\Big\|^2\Big)^{1/2} && \text{(Jensen)}\\
&\le C\,M^{-1/2} \;\to\; 0 && \text{(previous display).}
\end{aligned}$$

By Markov, this bound on $|B_1|$ is $o_p(1)$. It holds for every $(\vartheta, \tilde{k})$, so $\sup_{(\vartheta,\tilde{k})}|B_1| = o_p(1)$.

Bound on B_2. Write $\tilde{\alpha}_{km} := \alpha_{km} - a_{m,J}$ and $\tilde{\alpha}^0_{km} := \alpha^0_{km} - a_{m,J}$, which lie in the bounded set $\mathcal{A}$ by (G1). The drift cancels in B_2, so $B_2 = \frac{2}{JM}\sum_j\sum_m v_{jm}(\tilde{\alpha}^0_{k^0(j)m} - \tilde{\alpha}_{\tilde{k}(j)m})$. The true values $(\tilde{\alpha}^0, k^0)$ are themselves a candidate. So $\sup_{(\alpha,\tilde{k})}|B_2|$ is at most four times the supremum in

the next display, and it suffices to show that

$$\sup_{(\alpha,\tilde{k})} \Big| \frac{1}{JM} \sum_j \sum_m v_{jm} \tilde{\alpha}_{\tilde{k}(j)m} \Big| \to_p 0.$$

Group the products by their candidate group and apply Cauchy–Schwarz over the pairs (m, ℓ):

$$\begin{aligned} \Big| \frac{1}{JM} \sum_j \sum_m v_{jm} \tilde{\alpha}_{\tilde{k}(j)m} \Big| &= \Big| \frac{1}{M} \sum_m \sum_\ell \tilde{\alpha}_{\ell m} \frac{1}{J} \sum_j \mathbb{1}\{\tilde{k}(j) = \ell\} v_{jm} \Big| \\ &\le C \Big(\frac{1}{M} \sum_m \sum_\ell \Big(\frac{1}{J} \sum_j \mathbb{1}\{\tilde{k}(j) = \ell\} v_{jm} \Big)^2 \Big)^{1/2}. \end{aligned}$$

Let $V_{ji} := M^{-1} \sum_m v_{jm} v_{im}$. Expanding the square,

$$\frac{1}{M} \sum_m \sum_\ell \Big(\frac{1}{J} \sum_j \mathbb{1}\{\tilde{k}(j) = \ell\} v_{jm} \Big)^2 = \frac{1}{J^2} \sum_j \sum_i \mathbb{1}\{\tilde{k}(j) = \tilde{k}(i)\} V_{ji} \le \frac{1}{J^2} \sum_j \sum_i |V_{ji}|.$$

The right side does not depend on $(\alpha, \tilde{k})$. For each pair (j, i), the triangle inequality and Jensen give $\mathrm{E}\,|V_{ji}| \le |\,\mathrm{E}\,V_{ji}| + (\mathrm{var}\,V_{ji})^{1/2}$. The mean is $\mathrm{E}\,V_{ji} = M^{-1} \sum_m \mathrm{E}[v_{jm} v_{im}]$. By (G2), V_{ji} averages M independent terms, so $\mathrm{var}\,V_{ji} \le M^{-2} \sum_m \mathrm{E}[v_{jm}^2 v_{im}^2] \le C/M$. Averaging over (j, i) and using the bound L of (G4),

$$\mathrm{E}\Big[\frac{1}{J^2} \sum_j \sum_i |V_{ji}| \Big] \le \frac{1}{J} \Big(\frac{1}{J} \sum_j \sum_i |\,\mathrm{E}\,V_{ji}| \Big) + \Big(\frac{C}{M} \Big)^{1/2} \le \frac{L}{J} + \Big(\frac{C}{M} \Big)^{1/2} \to 0.$$

By Markov, the supremum is $o_p(1)$, and so is $\sup_{(\alpha,\tilde{k})} |B_2|$.

Since $\check{Q} - \tilde{Q} = A + B_1 + B_2$, the triangle inequality gives $\sup_{(\vartheta,\alpha,\tilde{k})} |\check{Q} - \tilde{Q}| \to_p 0$. □

Proposition 1 (Consistency). *Under (G1)–(G8), there is a permutation $\hat{\tau}$ of $\{1, \dots, K\}$, which depends on the data, such that for every $k \le K$,*

$$\|\hat{\vartheta}_{\hat{\tau}(k)} - \vartheta_k^0\| \to_p 0, \qquad \frac{1}{M} \sum_{m=1}^{M} \big(\hat{\alpha}_{\hat{\tau}(k)m} - \alpha_{km}^0 \big)^2 \to_p 0.$$

Proof. Step 1: the average fit error is $o_p(1)$. By (G1), the true values $(\vartheta^0, \alpha^0, k^0)$ lie in $\Theta^K \times \mathcal{A}_J^{KM} \times \tilde{\mathcal{K}}$. Write $\Delta := \sup_{(\vartheta,\alpha,\tilde{k})} |\check{Q}(\vartheta, \alpha, \tilde{k}) - \tilde{Q}(\vartheta, \alpha, \tilde{k})|$, which is $o_p(1)$ by Lemma 2. Write

also $\hat{D}_{jm} := D_{jm}(\hat{\vartheta}, \hat{\alpha}, \hat{k})$ and $\hat{D} := \frac{1}{JM}\sum_j \sum_m \hat{D}_{jm}^2$. Then

$$
\begin{aligned}
0 \leq \hat{D} &= \tilde{Q}(\hat{\vartheta}, \hat{\alpha}, \hat{k}) - \frac{1}{JM}\sum_j \sum_m v_{jm}^2 && \text{(definition of } \tilde{Q}\text{)} \\
&= \tilde{Q}(\hat{\vartheta}, \hat{\alpha}, \hat{k}) - \tilde{Q}(\vartheta^0, \alpha^0, k^0) && (D_{jm}(\vartheta^0, \alpha^0, k^0) = 0) \\
&\leq \check{Q}(\hat{\vartheta}, \hat{\alpha}, \hat{k}) - \check{Q}(\vartheta^0, \alpha^0, k^0) + 2\Delta && \text{(Lemma 2 at both points)} \\
&\leq 2\Delta && \text{(the estimate minimizes } \check{Q}\text{)}.
\end{aligned}
$$

So $\hat{D} = o_p(1)$.

Step 2: the fit error bounds the distance between parameters. Next, we find a lower bound on the fit error of the set in terms of that common difference. To do so, we bound the error $\hat{D}_{jm}$ for some specific set of products. Concretely, fix $k, \ell \leq K$ and consider the set of products $\{j : k^0(j) = k,\ \hat{k}(j) = \ell\}$, that is, the products of true group k that the classifier assigns to group ℓ. Because all of them share the true parameters and the estimated parameters, the difference between the two is the same.

For each product in the set, write $\hat{D}_{jm} = w_{jm}'(\vartheta_k^0 - \hat{\vartheta}_\ell) + \alpha_{km}^0 - \hat{\alpha}_{\ell m}$. Expanding the square and collecting terms,

$$
\frac{1}{M}\sum_m \hat{D}_{jm}^2 = \begin{pmatrix} \vartheta_k^0 - \hat{\vartheta}_\ell \\ \frac{\alpha_{k1}^0 - \hat{\alpha}_{\ell 1}}{\sqrt{M}} \\ \vdots \\ \frac{\alpha_{kM}^0 - \hat{\alpha}_{\ell M}}{\sqrt{M}} \end{pmatrix}' \begin{pmatrix} \frac{1}{M}\sum_m w_{jm}w_{jm}' & \frac{w_{j1}}{\sqrt{M}} & \cdots & \frac{w_{jM}}{\sqrt{M}} \\ \frac{w_{j1}'}{\sqrt{M}} & 1 & & \\ \vdots & & \ddots & \\ \frac{w_{jM}'}{\sqrt{M}} & & & 1 \end{pmatrix} \begin{pmatrix} \vartheta_k^0 - \hat{\vartheta}_\ell \\ \frac{\alpha_{k1}^0 - \hat{\alpha}_{\ell 1}}{\sqrt{M}} \\ \vdots \\ \frac{\alpha_{kM}^0 - \hat{\alpha}_{\ell M}}{\sqrt{M}} \end{pmatrix}.
$$

Next, we sum over the products in the set:

$$
\begin{aligned}
0_p(1) = \hat{D} &\geq \frac{1}{J}\sum_j \mathbb{1}\{k^0(j) = k,\ \hat{k}(j) = \ell\}\, \frac{1}{M}\sum_m \hat{D}_{jm}^2 \\
&= \begin{pmatrix} \vartheta_k^0 - \hat{\vartheta}_\ell \\ \frac{\alpha_{k1}^0 - \hat{\alpha}_{\ell 1}}{\sqrt{M}} \\ \vdots \\ \frac{\alpha_{kM}^0 - \hat{\alpha}_{\ell M}}{\sqrt{M}} \end{pmatrix}' \left[\frac{1}{J}\sum_j \mathbb{1}\{k^0(j) = k,\ \hat{k}(j) = \ell\} \begin{pmatrix} \frac{1}{M}\sum_m w_{jm}w_{jm}' & \frac{w_{j1}}{\sqrt{M}} & \cdots & \frac{w_{jM}}{\sqrt{M}} \\ \frac{w_{j1}'}{\sqrt{M}} & 1 & & \\ \vdots & & \ddots & \\ \frac{w_{jM}'}{\sqrt{M}} & & & 1 \end{pmatrix} \right] \begin{pmatrix} \vartheta_k^0 - \hat{\vartheta}_\ell \\ \frac{\alpha_{k1}^0 - \hat{\alpha}_{\ell 1}}{\sqrt{M}} \\ \vdots \\ \frac{\alpha_{kM}^0 - \hat{\alpha}_{\ell M}}{\sqrt{M}} \end{pmatrix} \\
&\geq \hat{\rho}_{k\ell}(\hat{k}) \Big(\|\vartheta_k^0 - \hat{\vartheta}_\ell\|^2 + \frac{1}{M}\sum_m \left(\alpha_{km}^0 - \hat{\alpha}_{\ell m}\right)^2 \Big) \geq 0,
\end{aligned}
$$

where the second inequality holds because by (G8) the smallest eigenvalue of the matrix above is $\hat{\rho}_{k\ell}(\hat{k})$, and a quadratic form is at least its smallest eigenvalue times the squared norm of the vector.

Note that the product can be small because parameters may be close, or $\hat{\rho}_{k\ell}(\hat{k})$ may be small. The next step shows that, for each true group k, at least one ℓ has $\hat{\rho}_{k\ell}(\hat{k})$ bounded away from zero, which implies that the estimated parameters are close to the true parameters.

Step 3: for each true group k, some $\hat{\rho}_{k\ell}(\hat{k})$ is bounded away from zero. By (G8), the bound

$$\min_{\ell \le K} \min_{\tilde{k} \in \tilde{\mathcal{K}}} \max_{\ell' \le K} \hat{\rho}_{\ell\ell'}(\tilde{k}) \;\ge\; \rho$$

holds with probability tending to one. Because the estimated partition $\hat{k} \in \tilde{\mathcal{K}}$, for each true group k there is at least one ℓ with $\hat{\rho}_{k\ell}(\hat{k}) \ge \rho$, when the bound in (G8) holds. For each k, let $\hat{\tau}(k)$ be the smallest such ℓ. Taking the smallest such ℓ only makes the choice unique. So $\hat{\tau}$ maps each true group to an estimated group. We show next that this map is injective.

Step 4: an injective map from true groups to estimated groups. To show injectivity, the argument combines two facts. Distinct true groups are far apart, and each true group is close to its match. If two true groups shared a match, they would both be close to it, and hence close to each other.

(a) Distinct true groups are far apart. Fix $k \neq k'$ and any product j. Then

$$\begin{aligned}
c &\le \frac{1}{M} \sum_m \mathrm{E}\left[\left(w'_{jm}(\vartheta^0_{k'} - \vartheta^0_k) + \alpha^0_{k'm} - \alpha^0_{km}\right)^2\right] && \text{(G7)} \\
&\le \frac{1}{M} \sum_m \left(2\,\mathrm{E}\left[\left(w'_{jm}(\vartheta^0_{k'} - \vartheta^0_k)\right)^2\right] + 2\left(\alpha^0_{k'm} - \alpha^0_{km}\right)^2\right) && \text{($(x+y)^2 \le 2x^2 + 2y^2$; centres non-random, (G1))} \\
&\le \frac{1}{M} \sum_m \left(2\,\mathrm{E}\,\|w_{jm}\|^2\,\|\vartheta^0_{k'} - \vartheta^0_k\|^2 + 2\left(\alpha^0_{k'm} - \alpha^0_{km}\right)^2\right) && \text{(Cauchy–Schwarz; slopes non-random, (G1))} \\
&\le C\left(\|\vartheta^0_k - \vartheta^0_{k'}\|^2 + \frac{1}{M} \sum_m \left(\alpha^0_{km} - \alpha^0_{k'm}\right)^2\right) && \text{($\mathrm{E}\,\|w_{jm}\|^2 \le C$, (G5)),}
\end{aligned}$$

where C depends only on $\sup_{j,m} \mathrm{E}\,\|w_{jm}\|^2$.

(b) Each true group is close to its match. For every k,

$$\begin{aligned}\hat{D} &\geq \hat{\rho}_{k\hat{\tau}(k)}(\hat{k})\Big(\|\vartheta_k^0 - \hat{\vartheta}_{\hat{\tau}(k)}\|^2 + \frac{1}{M}\sum_m \big(\alpha_{km}^0 - \hat{\alpha}_{\hat{\tau}(k)m}\big)^2\Big) && \text{(Step 2 with } \ell = \hat{\tau}(k)\text{)}\\ &\geq \rho\Big(\|\vartheta_k^0 - \hat{\vartheta}_{\hat{\tau}(k)}\|^2 + \frac{1}{M}\sum_m \big(\alpha_{km}^0 - \hat{\alpha}_{\hat{\tau}(k)m}\big)^2\Big) && \text{(definition of } \hat{\tau}\text{).}\end{aligned}$$

The second inequality also uses that the term in parentheses is non-negative. Dividing by $\rho > 0$,

$$\|\vartheta_k^0 - \hat{\vartheta}_{\hat{\tau}(k)}\|^2 + \frac{1}{M}\sum_m \big(\alpha_{km}^0 - \hat{\alpha}_{\hat{\tau}(k)m}\big)^2 \leq \frac{\hat{D}}{\rho} \qquad \text{for every } k. \tag{34}$$

(c) No two distinct true groups share a match. Suppose $\hat{\tau}(k) = \hat{\tau}(k')$ for some $k \neq k'$. Write $\vartheta_k^0 - \vartheta_{k'}^0 = (\vartheta_k^0 - \hat{\vartheta}_{\hat{\tau}(k)}) - (\vartheta_{k'}^0 - \hat{\vartheta}_{\hat{\tau}(k)})$, and likewise for the centres in each market. Then

$$\begin{aligned}\|\vartheta_k^0 - \vartheta_{k'}^0\|^2 + \frac{1}{M}\sum_m \big(\alpha_{km}^0 - \alpha_{k'm}^0\big)^2 &\leq 2\Big(\|\vartheta_k^0 - \hat{\vartheta}_{\hat{\tau}(k)}\|^2 + \frac{1}{M}\sum_m \big(\alpha_{km}^0 - \hat{\alpha}_{\hat{\tau}(k)m}\big)^2\Big)\\ &\quad + 2\Big(\|\vartheta_{k'}^0 - \hat{\vartheta}_{\hat{\tau}(k)}\|^2 + \frac{1}{M}\sum_m \big(\alpha_{k'm}^0 - \hat{\alpha}_{\hat{\tau}(k)m}\big)^2\Big)\\ &\leq \frac{4\hat{D}}{\rho} && \text{((34) for } k \text{ and for } k'\text{).}\end{aligned}$$

By (a), the left side is at least c/C. So a shared match implies $c/C \leq 4\hat{D}/\rho$, that is, $\hat{D} \geq c\rho/(4C)$. Since $\hat{D} = o_p(1)$ and $c\rho/(4C)$ is a fixed positive number, a shared match has probability tending to zero.

Hence, with probability tending to one, the bound in (G8) holds and no two distinct true groups share a match. In these samples, $\hat{\tau}$ is defined and injective from $\{1, \ldots, K\}$ to itself, so it is a permutation. In the remaining samples, where the bound in (G8) fails or $\hat{\tau}$ is not injective, set $\hat{\tau}$ equal to the identity. So $\hat{\tau}$ is a permutation in every sample.

Step 5: conclusion. Let E be the event that the bound in (G8) holds and $\hat{\tau}$, as defined in Step 4, is injective. By Step 4(c), $\mathbb{P}(E) \to 1$. Write $\hat{d}_k := \|\hat{\vartheta}_{\hat{\tau}(k)} - \vartheta_k^0\|^2 + \frac{1}{M}\sum_m \big(\hat{\alpha}_{\hat{\tau}(k)m} - \alpha_{km}^0\big)^2$. On E, (34) gives $\max_{k \leq K} \hat{d}_k \leq \hat{D}/\rho$. Fix $\varepsilon > 0$. Then

$$\mathbb{P}\Big(\max_{k \leq K} \hat{d}_k > \varepsilon\Big) \leq \mathbb{P}(E^c) + \mathbb{P}\big(\hat{D} > \rho\varepsilon\big) \to 0,$$

since $\hat{D} = o_p(1)$ by Step 1. Both terms of $\hat{d}_k$ are non-negative, so each is $o_p(1)$, uniformly in k.

□

From here on, we relabel the estimated groups by the permutation $\hat{\tau}$ of Proposition 1, so that $\hat{\tau}$ is the identity, as in BM15. For $\eta > 0$ let

$$\mathcal{N}(\eta) := \Big\{(\vartheta, \alpha) : \ \|\vartheta_k - \vartheta_k^0\|^2 < \eta \text{ and } \frac{1}{M}\sum_{m=1}^{M}(\alpha_{km} - \alpha_{km}^0)^2 < \eta \text{ for all } k \le K\Big\}.$$

After the relabelling, Proposition 1 says that $\mathbb{P}\big((\hat{\vartheta}, \hat{\alpha}) \in \mathcal{N}(\eta)\big) \to 1$ for every $\eta > 0$.

Lemma 3 (Misclassification of one product)**.** *Under (G1)–(G5) and (G7), there is $\eta_0 > 0$ such that for every $\eta \in (0, \eta_0)$ and every $\delta > 0$,*

$$\max_{j \le J} \mathbb{P}\Big((\hat{\vartheta}, \hat{\alpha}) \in \mathcal{N}(\eta),\ \hat{k}(j) \neq k^0(j),\ M^{-1/2}\|e_j\| \le \sqrt{\eta}\Big) = o\big(M^{-\delta}\big).$$

Proof. Fix j and write $k := k^0(j)$. Fix $\ell \neq k$. The event $\{\hat{k}(j) \neq k\}$ is the union over $\ell \neq k$ of the events treated below, and $K - 1$ is fixed.

Step 1: the decision rule. Given $(\hat{\vartheta}, \hat{\alpha})$, the classifier assigns product j to the group ℓ' that minimizes $\sum_m (y_{jm} - w_{jm}'\hat{\vartheta}_{\ell'} - \hat{\alpha}_{\ell' m})^2$. Let

$$\hat{\nu}_m := y_{jm} - w_{jm}'\hat{\vartheta}_k - \hat{\alpha}_{km}, \qquad \hat{u}_m := w_{jm}'(\hat{\vartheta}_\ell - \hat{\vartheta}_k) + \hat{\alpha}_{\ell m} - \hat{\alpha}_{km},$$

so that $y_{jm} - w_{jm}'\hat{\vartheta}_\ell - \hat{\alpha}_{\ell m} = \hat{\nu}_m - \hat{u}_m$. The event $\{\hat{k}(j) = \ell\}$ is contained in $\{\sum_m (\hat{\nu}_m - \hat{u}_m)^2 \le \sum_m \hat{\nu}_m^2\}$, that is,

$$\sum_m \hat{u}_m \hat{\nu}_m \ \ge\ \frac{1}{2}\sum_m \hat{u}_m^2. \tag{35}$$

By (30), $\hat{\nu}_m = v_{jm} + e_{jm} + w_{jm}'(\vartheta_k^0 - \hat{\vartheta}_k) + (\alpha_{km}^0 - \hat{\alpha}_{km})$, so

$$\sum_m \hat{u}_m \hat{\nu}_m = \sum_m \hat{u}_m v_{jm} + \sum_m \hat{u}_m e_{jm} + \sum_m \hat{u}_m w_{jm}'(\vartheta_k^0 - \hat{\vartheta}_k) + \sum_m \hat{u}_m(\alpha_{km}^0 - \hat{\alpha}_{km}).$$

Step 2: reduction to the true contrast. Let

$$u_m := w_{jm}'(\vartheta_\ell^0 - \vartheta_k^0) + \alpha_{\ell m}^0 - \alpha_{km}^0 \tag{36}$$

be the true contrast between groups ℓ and k at product j, and let $d_m := \hat{u}_m - u_m$, so that

$$d_m = w'_{jm}\big((\hat{\vartheta}_\ell - \vartheta^0_\ell) - (\hat{\vartheta}_k - \vartheta^0_k)\big) + (\hat{\alpha}_{\ell m} - \alpha^0_{\ell m}) - (\hat{\alpha}_{km} - \alpha^0_{km}).$$

On $\{(\hat{\vartheta}, \hat{\alpha}) \in \mathcal{N}(\eta)\}$, by the triangle inequality and Cauchy–Schwarz,

$$M^{-1/2}\|d\| \le 2\sqrt{\eta}\, M^{-1/2}\|w_j\| + 2\sqrt{\eta} = \sqrt{\eta}\,\kappa_j, \qquad \kappa_j := 2(1 + M^{-1/2}\|w_j\|),$$

The bound follows from $\|(w'_{jm}a)_{m\le M}\| \le \|w_j\|\,\|a\|$ for any vector a, applied to $a = \hat{\vartheta}_\ell - \vartheta^0_\ell$ and $a = \hat{\vartheta}_k - \vartheta^0_k$, each of norm at most $\sqrt{\eta}$ on $\mathcal{N}(\eta)$, and from $\|(\hat{\alpha}_{\ell m} - \alpha^0_{\ell m})_{m\le M}\| \le \sqrt{M\eta}$ and likewise for k. Since $\hat{u} = u + d$, also $M^{-1/2}\|\hat{u}\| \le M^{-1/2}\|u\| + \sqrt{\eta}\,\kappa_j$.

We now rewrite the decision rule (35) in terms of the true contrast u and the sample average $M^{-1}\sum_m u_m v_{jm}$. Write $a_m := w'_{jm}(\vartheta^0_k - \hat{\vartheta}_k)$ and $b_m := \alpha^0_{km} - \hat{\alpha}_{km}$ for the estimation errors of the slope and of the centre of the true group, so that $\hat{\nu}_m = v_{jm} + e_{jm} + a_m + b_m$. Insert $\hat{u} = u + d$ on both sides of (35). The left side is

$$\sum_m \hat{u}_m\hat{\nu}_m = \sum_m u_m v_{jm} + \sum_m d_m v_{jm} + \sum_m \hat{u}_m e_{jm} + \sum_m \hat{u}_m a_m + \sum_m \hat{u}_m b_m,$$

and the right side is

$$\frac{1}{2}\sum_m \hat{u}_m^2 = \frac{1}{2}\sum_m u_m^2 + \sum_m u_m d_m + \frac{1}{2}\sum_m d_m^2.$$

Divide by M and move every term except $M^{-1}\sum_m u_m v_{jm}$ to the right. The event (35) is

$$\begin{aligned}\frac{1}{M}\sum_m u_m v_{jm} \;\ge\; &\frac{1}{2M}\|u\|^2 + \frac{1}{2M}\|d\|^2 + \frac{1}{M}\sum_m u_m d_m \\ &- \frac{1}{M}\sum_m d_m v_{jm} - \frac{1}{M}\sum_m \hat{u}_m e_{jm} - \frac{1}{M}\sum_m \hat{u}_m a_m - \frac{1}{M}\sum_m \hat{u}_m b_m. \qquad (37)\end{aligned}$$

Each term on the right has a role. The first is the signal: the squared distance between the two groups at product j, which (G7) keeps away from zero. The second is non-negative and is dropped. The remaining five are the nuisance terms that separate this lemma from the case with known slopes and centres and no perturbation. $M^{-1}\sum_m u_m d_m$ is the estimation error of the contrast, projected on the contrast itself. $M^{-1}\sum_m d_m v_{jm}$ is the same estimation error in-

teracted with the noise of product j. $M^{-1}\sum_m \hat{u}_m e_{jm}$ is the perturbation of the group-market intercepts, which is random and may be correlated with everything else. $M^{-1}\sum_m \hat{u}_m a_m$ and $M^{-1}\sum_m \hat{u}_m b_m$ are the estimation errors of the slope and the centre of the true group, which enter the residual $\hat{\nu}_m$ under the true group. Each of the five is a sample average of a product of two arrays over markets. Cauchy–Schwarz bounds it by the product of the two root mean squares, and on $\{M^{-1/2}\|e_j\| \le \sqrt{\eta}\} \cap \{(\hat{\vartheta}, \hat{\alpha}) \in \mathcal{N}(\eta)\}$ one factor in each product is at most $\sqrt{\eta}$ times something:

$$\Big|\frac{1}{M}\sum_m u_m d_m\Big| \le M^{-1/2}\|u\|\, M^{-1/2}\|d\| \le \sqrt{\eta}\,\kappa_j M^{-1/2}\|u\|,$$

$$\Big|\frac{1}{M}\sum_m d_m v_{jm}\Big| \le M^{-1/2}\|d\|\, M^{-1/2}\|v_j\| \le \sqrt{\eta}\,\kappa_j M^{-1/2}\|v_j\|,$$

$$\Big|\frac{1}{M}\sum_m \hat{u}_m e_{jm}\Big| \le M^{-1/2}\|\hat{u}\|\, M^{-1/2}\|e_j\| \le \sqrt{\eta}\,\big(M^{-1/2}\|u\| + \sqrt{\eta}\,\kappa_j\big),$$

$$\Big|\frac{1}{M}\sum_m \hat{u}_m b_m\Big| \le M^{-1/2}\|\hat{u}\|\Big(\frac{1}{M}\sum_m (\alpha^0_{km} - \hat{\alpha}_{km})^2\Big)^{1/2} \le \sqrt{\eta}\,\big(M^{-1/2}\|u\| + \sqrt{\eta}\,\kappa_j\big),$$

$$\Big|\frac{1}{M}\sum_m \hat{u}_m a_m\Big| \le M^{-1/2}\|\hat{u}\|\, M^{-1/2}\|w_j\|\, \|\vartheta^0_k - \hat{\vartheta}_k\|$$

$$\le \sqrt{\eta}\, M^{-1/2}\|w_j\|\big(M^{-1/2}\|u\| + \sqrt{\eta}\,\kappa_j\big).$$

The bounds are of two kinds. Those proportional to $M^{-1/2}\|u\|$ scale with the signal. Summing their coefficients,

$$\sqrt{\eta}\, M^{-1/2}\|u\|\,\big(\kappa_j + 2 + M^{-1/2}\|w_j\|\big) = \sqrt{\eta}\, M^{-1/2}\|u\|\,\big(4 + 3M^{-1/2}\|w_j\|\big)$$

$$\le 4\sqrt{\eta}\,(1 + M^{-1/2}\|w_j\|)\, M^{-1/2}\|u\|.$$

In Step 3 this is absorbed by the signal $\frac{1}{2M}\|u\|^2$ once η is small, because $M^{-1/2}\|u\|$ is bounded below off a tail event. The remaining bounds do not involve u:

$$\sqrt{\eta}\,\kappa_j M^{-1/2}\|v_j\| + 2\eta\,\kappa_j + \eta\,\kappa_j M^{-1/2}\|w_j\|.$$

With $\eta \le 1$, $\kappa_j = 2(1 + M^{-1/2}\|w_j\|)$ and $2ab \le a^2 + b^2$, the three terms are at most $\sqrt{\eta}\,(1 + M^{-1}\|w_j\|^2 + 2M^{-1}\|v_j\|^2)$, $\sqrt{\eta}\,(6 + 2M^{-1}\|w_j\|^2)$ and $\sqrt{\eta}\,(1 + 3M^{-1}\|w_j\|^2)$, so their sum is at most $8\sqrt{\eta}\,\big(1 + M^{-1}(\|w_j\|^2 + \|v_j\|^2)\big)$. In Step 3 these are absorbed by the lower bound on the

signal, because $M^{-1}\|w_j\|^2$ and $M^{-1}\|v_j\|^2$ are bounded off a tail event. Dropping $\frac{1}{2M}\|d\|^2 \geq 0$ from the right side of (37) and inserting the bounds, the event (35) is contained in

$$\Big\{\frac{1}{M}\sum_m u_m v_{jm} \;\geq\; \frac{1}{2M}\|u\|^2 - 4\sqrt{\eta}\,(1+M^{-1/2}\|w_j\|)\,M^{-1/2}\|u\| - C_0\sqrt{\eta}\,\big(1+M^{-1}(\|w_j\|^2+\|v_j\|^2)\big)\Big\}, \tag{38}$$

with $C_0 = 8$. What remains on the left is the one term whose smallness is probabilistic rather than algebraic: the sample covariance between the contrast and the noise of product j. Its mean is zero by (G3) and (G4), and Step 4 bounds its tail. This step uses no bound on Θ or $\mathcal{A}$. The size of the centres enters only through $M^{-1/2}\|u\|$. The containment holds pointwise and for any labelling of the estimated groups. In particular, it holds after the relabelling by $\hat{\tau}$ of Proposition 1, even though $\hat{\tau}$ depends on the data.

Step 3: tail events. Let $\bar{u}^{\,2} := M^{-1}\sum_m \mathrm{E}\,u_m^2$, which is at least c by (G7). Let $\bar{M}$ be the constant of Lemma 5 and consider the events

$$\Omega_1 := \Big\{M^{-1}\|u\|^2 \leq \tfrac{1}{2}\bar{u}^{\,2}\Big\}, \qquad \Omega_2 := \Big\{M^{-1}\|w_j\|^2 \geq \bar{M}\Big\}, \qquad \Omega_3 := \Big\{M^{-1}\|v_j\|^2 \geq \bar{M}\Big\}.$$

Outside $\Omega_1 \cup \Omega_2 \cup \Omega_3$, $M^{-1/2}\|u\| \geq \bar{u}/\sqrt{2} \geq \sqrt{c/2}$, $1 + M^{-1/2}\|w_j\| \leq 1 + \sqrt{\bar{M}}$ and $1 + M^{-1}(\|w_j\|^2 + \|v_j\|^2) \leq 1 + 2\bar{M}$. Choose $\eta_0 \leq 1$ so small that for all $\eta < \eta_0$

$$4\sqrt{\eta}\,(1+\sqrt{\bar{M}}) \leq \tfrac{1}{8}\sqrt{c/2} \qquad \text{and} \qquad C_0\sqrt{\eta}\,(1+2\bar{M}) \leq \tfrac{1}{16}c.$$

Then, outside $\Omega_1 \cup \Omega_2 \cup \Omega_3$, the right side of (38) is at least $\frac{1}{2M}\|u\|^2 - \frac{1}{8M}\|u\|^2 - \frac{1}{8M}\|u\|^2 = \frac{1}{4M}\|u\|^2 \geq \frac{1}{8}\bar{u}^{\,2}$. The first inequality uses $4\sqrt{\eta}(1+\sqrt{\bar{M}})M^{-1/2}\|u\| \leq \frac{1}{8}\sqrt{c/2}\,M^{-1/2}\|u\| \leq \frac{1}{8M}\|u\|^2$. The second uses $\frac{1}{16}c \leq \frac{1}{8M}\|u\|^2$. Hence (38) is contained in

$$\Omega_1 \cup \Omega_2 \cup \Omega_3 \cup \Omega_4, \qquad \Omega_4 := \Big\{\frac{1}{M}\sum_m u_m v_{jm} \;\geq\; \tfrac{1}{8}\bar{u}^{\,2}\Big\}.$$

The choice of η_0 depends on c, $\bar{M}$ and C_0 only. It does not depend on j, k, ℓ, J, M, or the size of $\bar{u}^{\,2}$.

Step 4: tail bounds. By Lemma 5(i)–(iv), each Ω_i has probability $o(M^{-\delta})$, uniformly in j, k, ℓ, and in the size of $\bar{u}^{\,2} \geq c$. The union over the $K-1$ values of ℓ and the maximum over j give the claim. □

Theorem B.1 (Misclassification, general model). *Under (G1)–(G9), with the estimated groups relabelled by $\hat{\tau}$ of Proposition 1, for every $\varsigma > 0$,*

$$\mathbb{P}\big(\hat{k}(j) \neq k^0(j) \text{ for some } j \leq J\big) = o(1) + o\big(JM^{-\varsigma}\big).$$

Proof. Fix $\eta \in (0, \eta_0)$, with η_0 from Lemma 3. Define the events

$$\mathcal{E}_1 := \big\{(\hat{\vartheta}, \hat{\alpha}) \in \mathcal{N}(\eta)\big\}, \qquad \mathcal{E}_2 := \Big\{\max_{j \leq J} M^{-1/2}\|e_j\| \leq \sqrt{\eta}\Big\}, \qquad \mathcal{F}_j := \big\{\hat{k}(j) \neq k^0(j)\big\}.$$

The event in the theorem is $\bigcup_{j \leq J} \mathcal{F}_j$. Then

$$\begin{aligned}
\bigcup_{j \leq J} \mathcal{F}_j &\subset \Big(\bigcup_{j \leq J} \mathcal{F}_j \cap \mathcal{E}_1 \cap \mathcal{E}_2\Big) \cup \mathcal{E}_1^c \cup \mathcal{E}_2^c && \text{(split on } \mathcal{E}_1 \cap \mathcal{E}_2\text{)} \\
&= \bigcup_{j \leq J} \big(\mathcal{F}_j \cap \mathcal{E}_1 \cap \mathcal{E}_2\big) \cup \mathcal{E}_1^c \cup \mathcal{E}_2^c && \text{(intersection distributes over union)} \\
&\subset \bigcup_{j \leq J} \Big(\mathcal{F}_j \cap \mathcal{E}_1 \cap \big\{M^{-1/2}\|e_j\| \leq \sqrt{\eta}\big\}\Big) \cup \mathcal{E}_1^c \cup \mathcal{E}_2^c && (\mathcal{E}_2 \subset \{M^{-1/2}\|e_j\| \leq \sqrt{\eta}\}).
\end{aligned}$$

So

$$\begin{aligned}
\mathbb{P}\Big(\bigcup_{j \leq J} \mathcal{F}_j\Big) &\leq \mathbb{P}(\mathcal{E}_1^c) + \mathbb{P}(\mathcal{E}_2^c) + \sum_{j \leq J} \mathbb{P}\Big(\mathcal{F}_j \cap \mathcal{E}_1 \cap \big\{M^{-1/2}\|e_j\| \leq \sqrt{\eta}\big\}\Big) && \text{(union bound)} \\
&\leq \mathbb{P}(\mathcal{E}_1^c) + \mathbb{P}(\mathcal{E}_2^c) + J \max_{j \leq J} \mathbb{P}\Big(\mathcal{F}_j \cap \mathcal{E}_1 \cap \big\{M^{-1/2}\|e_j\| \leq \sqrt{\eta}\big\}\Big) && \text{(sum} \leq J \times \text{max)} \\
&= o(1) + o(1) + J\, o\big(M^{-\varsigma}\big).
\end{aligned}$$

The first term is $o(1)$ by Proposition 1, since the groups are relabelled by $\hat{\tau}$. The second term is $o(1)$ by (G6), since $M^{-1}\|e_j\|^2 = M^{-1}\sum_m e_{jm}^2$. The maximum in the third term is the probability in Lemma 3. With $\delta = \varsigma$, that lemma makes it $o(M^{-\varsigma})$. The lemma holds for any labelling of the estimated groups (Step 2 of its proof). □

Lemma 4 (Covariates observed with vanishing error). *Let (G1)–(G9) hold for (y_{jm}, w_{jm}). Let Δ_{jm} be random vectors in $\mathbb{R}^d$ with $\max_{j \leq J} M^{-1}\sum_{m=1}^{M}\|\Delta_{jm}\|^2 \to_p 0$, and let $\hat{w}_{jm} := w_{jm} + \Delta_{jm}$. Let $(\hat{\vartheta}, \hat{\alpha}, \hat{k})$ be the classifier (31) computed with $\hat{w}_{jm}$ in place of w_{jm}. Then the conclusion of Propo-*

sition 1 holds for $(\hat{\vartheta}, \hat{\alpha})$, and, after the relabelling by $\hat{\tau}$, for every $\varsigma > 0$,

$$\mathbb{P}\big(\hat{k}(j) \neq k^0(j) \text{ for some } j \leq J\big) = o(1) + o\big(JM^{-\varsigma}\big).$$

Proof. Let $C_\Theta := \sup_{\vartheta \in \Theta} \|\vartheta\|$, finite by (G1), and $\bar{\Delta} := \max_{j \leq J} M^{-1/2} \|\Delta_j\|$, so that $\bar{\Delta} = o_p(1)$ by assumption. For every $(\vartheta, \alpha, \tilde{k})$ the residual with the observed covariates is

$$y_{jm} - \hat{w}'_{jm}\vartheta_{\tilde{k}(j)} - \alpha_{\tilde{k}(j)m} = v_{jm} + \tilde{e}_{jm}(\vartheta, \tilde{k}) + D_{jm}(\vartheta, \alpha, \tilde{k}), \qquad \tilde{e}_{jm}(\vartheta, \tilde{k}) := e_{jm} - \Delta'_{jm}\vartheta_{\tilde{k}(j)},$$

with D_{jm} as in (32), defined with the true covariates. By Cauchy–Schwarz, $|\tilde{e}_{jm}(\vartheta, \tilde{k})| \leq |e_{jm}| + C_\Theta \|\Delta_{jm}\|$, so

$$\max_{j \leq J} M^{-1/2} \|\tilde{e}_j(\vartheta, \tilde{k})\| \leq \bar{e} + C_\Theta \bar{\Delta} =: \bar{e}_\Delta \qquad \text{for every } (\vartheta, \tilde{k}),$$

with $\bar{e} := \max_j M^{-1/2} \|e_j\|$ as in the proof of Lemma 2. By (G6) and the assumption on Δ, $\bar{e}_\Delta = o_p(1)$. The point is that $\tilde{e}$ depends on the candidate parameters, but its bound does not.

Step 1: criterion. Let $\hat{Q}(\vartheta, \alpha, \tilde{k})$ be the sample objective of (31) with $\hat{w}$, normalized by JM. The decomposition (33) holds with $\tilde{e}_{jm}(\vartheta, \tilde{k})$ in place of e_{jm}: $\hat{Q} - \tilde{Q} = A_\Delta + B$, where A_Δ is the term A with $\tilde{e}$ in place of e and B is unchanged, since B involves only v and D. The bound on A in the proof of Lemma 2 uses e only through $\bar{e}$ and holds for every $(\vartheta, \alpha, \tilde{k})$. It therefore holds for A_Δ with $\bar{e}_\Delta$ in place of $\bar{e}$. Hence $\sup_{(\vartheta, \alpha, \tilde{k})} |\hat{Q} - \tilde{Q}| \to_p 0$.

Step 2: slopes and centres. The proof of Proposition 1 uses the sample criterion only in its Step 1, through Lemma 2 at the estimate and at the truth. With $\hat{Q}$ in place of $\check{Q}$, Step 1 of that proof gives $\hat{D} \leq 2 \sup |\hat{Q} - \tilde{Q}| = o_p(1)$, where $\hat{D}$ is defined with the true covariates. Steps 2–5 use $\hat{D}$, (G7), (G8) and the non-randomness of ϑ^0 and α^0. They do not involve the covariates used by the classifier. So the conclusion of Proposition 1 holds for $(\hat{\vartheta}, \hat{\alpha})$, with a permutation $\hat{\tau}$, and we relabel by it.

Step 3: one product. Fix j, $k := k^0(j)$ and $\ell \neq k$. Given $(\hat{\vartheta}, \hat{\alpha})$, the classifier assigns j to ℓ only if (35) holds with

$$\hat{\nu}_m := y_{jm} - \hat{w}'_{jm}\hat{\vartheta}_k - \hat{\alpha}_{km}, \qquad \hat{u}_m := \hat{w}'_{jm}(\hat{\vartheta}_\ell - \hat{\vartheta}_k) + \hat{\alpha}_{\ell m} - \hat{\alpha}_{km}.$$

By (30), $\hat{\nu}_m = v_{jm} + e_{jm} + w'_{jm}(\vartheta^0_k - \hat{\vartheta}_k) + (\alpha^0_{km} - \hat{\alpha}_{km}) - \Delta'_{jm}\hat{\vartheta}_k$, and $\hat{u}_m = u_m + d_m + \Delta'_{jm}(\hat{\vartheta}_\ell - \hat{\vartheta}_k)$,

with u_m the true contrast (36) and d_m as in the proof of Lemma 3. On $\{(\hat\vartheta,\hat\alpha)\in\mathcal{N}(\eta)\}\cap\{M^{-1/2}\|\Delta_j\|\le\sqrt{\eta}\}$,

$$M^{-1/2}\big\|\big(\Delta'_{jm}(\hat\vartheta_\ell-\hat\vartheta_k)\big)_{m\le M}\big\|\le 2C_\Theta\sqrt{\eta},\qquad M^{-1/2}\big\|\big(\Delta'_{jm}\hat\vartheta_k\big)_{m\le M}\big\|\le C_\Theta\sqrt{\eta}.$$

So the bound on d in Step 2 of the proof of Lemma 3 holds for $d_m+\Delta'_{jm}(\hat\vartheta_\ell-\hat\vartheta_k)$ with $\kappa_j:=2(1+C_\Theta+M^{-1/2}\|w_j\|)$, and the additional term $M^{-1}\sum_m\hat u_m\Delta'_{jm}\hat\vartheta_k$ is bounded by $C_\Theta\sqrt{\eta}\,(M^{-1/2}\|u\|+\sqrt{\eta}\,\kappa_j)$, which has the form of the bound on the e term there. The containment (38) therefore holds with a constant C_0 that depends on C_Θ, and Steps 3–4 of that proof are unchanged. Hence there is $\eta_0>0$, depending on c, $\bar M$, C_0 and C_Θ, such that for every $\eta\in(0,\eta_0)$ and $\delta>0$,

$$\max_{j\le J}\mathbb{P}\Big((\hat\vartheta,\hat\alpha)\in\mathcal{N}(\eta),\ \hat k(j)\ne k^0(j),\ M^{-1/2}\|e_j\|\le\sqrt{\eta},\ M^{-1/2}\|\Delta_j\|\le\sqrt{\eta}\Big)=o(M^{-\delta}).$$

Step 4: all products. Repeat the proof of Theorem B.1 with the event $\mathcal{E}_3:=\{\bar\Delta\le\sqrt{\eta}\}$ added to $\mathcal{E}_1\cap\mathcal{E}_2$. The union bound gives $\mathbb{P}(\mathcal{E}_1^c)+\mathbb{P}(\mathcal{E}_2^c)+\mathbb{P}(\mathcal{E}_3^c)+J\,o(M^{-\varsigma})$. The first term is $o(1)$ by Step 2, the second by (G6), the third by assumption, and the last by Step 3 with $\delta=\varsigma$. □

Rates. (G9) enters only through the union bound in Theorem B.1. It allows J to grow like any power of M. Neither J nor M can stay fixed. With J fixed, the classifier estimates each centre α^0_{km} from a fixed number of products, and Proposition 1 fails. With M fixed, it classifies each product from a fixed number of observations, and Lemma 5 fails.

Lemma 5 (Tail bounds over markets)**.** *Assume (G1)–(G5). Fix a product j and, for $k\ne\ell$, let $u_m:=u^{k\ell}_{jm}$ be the contrast (36), with $\bar u^{\,2}:=M^{-1}\sum_m\mathrm{E}\,u_m^2$. Suppose $\bar u^{\,2}\ge c>0$. There is $\bar M<\infty$, depending on the constants in the conditions only, such that for every $\delta>0$ each of the following probabilities is $o(M^{-\delta})$, uniformly in $j\le J$, in $k\ne\ell$, and in the value of $\bar u^{\,2}\ge c$:*

(i) $\mathbb{P}\big(M^{-1}\|w_j\|^2\ge\bar M\big)$;

(ii) $\mathbb{P}\big(M^{-1}\|v_j\|^2\ge\bar M\big)$;

(iii) $\mathbb{P}\big(M^{-1}\|u\|^2\le\frac12\bar u^{\,2}\big)$;

(iv) $\mathbb{P}\big(\frac1M\sum_m u_m v_{jm}\ge\frac18\bar u^{\,2}\big)$.

Proof. Preliminaries. By (G2), the pairs (w_{jm}, v_{jm}) are independent across m. By (G3)–(G5), $\|w_{jm}\|_{\psi_2} \le C$ and $\|v_{jm}\|_{\psi_2} \le C$ uniformly, $\mathrm{E}\, v_{jm} = 0$ and $\mathrm{E}[w_{jm} v_{jm}] = 0$. The contrast $u_m = w_{jm}'(\vartheta_\ell - \vartheta_k) + (\alpha^0_{\ell m} - \alpha^0_{km})$ is independent across m. Write $b_m := \mathrm{E}\, u_m$, non-random, and $\zeta_m := u_m - b_m = (w_{jm} - \mathrm{E}\, w_{jm})'(\vartheta_\ell - \vartheta_k)$, with $\|\zeta_m\|_{\psi_2} \le C_w$ uniformly by (G5). Since $b_m^2 = (\mathrm{E}\, u_m)^2 \le \mathrm{E}\, u_m^2$,

$$\frac{1}{M} \sum_m b_m^2 \le \bar{u}^{\,2}. \tag{39}$$

The means b_m may grow with $\bar{u}$; (39) says they do not grow faster.

(i) $\|w_{jm}\|^2$ is sub-exponential with $\| \|w_{jm}\|^2 \|_{\psi_1} \le dC^2$ and mean bounded by C_1. Bernstein with $t = 1$ and $\bar{M} \ge C_1 + 1$.

(ii) v_{jm}^2 is sub-exponential with norm at most C^2 and mean at most C_2. Bernstein with $t = 1$ and $\bar{M} \ge C_2 + 1$.

(iii) $u_m^2 = b_m^2 + 2 b_m \zeta_m + \zeta_m^2$ and $\bar{u}^{\,2} = M^{-1} \sum_m (b_m^2 + \mathrm{E}\, \zeta_m^2)$, so

$$M^{-1} \|u\|^2 - \bar{u}^{\,2} = \frac{2}{M} \sum_m b_m \zeta_m + \frac{1}{M} \sum_m (\zeta_m^2 - \mathrm{E}\, \zeta_m^2).$$

The event $\{M^{-1}\|u\|^2 \le \frac{1}{2} \bar{u}^{\,2}\}$ is contained in the union of

$$\Big\{ \frac{2}{M} \sum_m b_m \zeta_m \le -\tfrac{1}{4} \bar{u}^{\,2} \Big\} \qquad \text{and} \qquad \Big\{ \frac{1}{M} \sum_m (\zeta_m^2 - \mathrm{E}\, \zeta_m^2) \le -\tfrac{1}{4} \bar{u}^{\,2} \Big\}.$$

For the first, Hoeffding with $B_m = |b_m| C_w$ and $t = \bar{u}^{\,2}/8$ gives the exponent

$$\frac{c_H\, M\, \bar{u}^{\,4}}{64 C_w^2\, M^{-1} \sum_m b_m^2} \ge \frac{c_H\, M\, \bar{u}^{\,2}}{64 C_w^2} \ge \frac{c_H\, c\, M}{64 C_w^2},$$

using (39) and $\bar{u}^{\,2} \ge c$. For the second, Bernstein for the centred sub-exponential ζ_m^2 with $t = \bar{u}^{\,2}/4 \ge c/4$. Both are $o(M^{-\delta})$ uniformly.

(iv) $\frac{1}{M} \sum_m u_m v_{jm} = \frac{1}{M} \sum_m \zeta_m v_{jm} + \frac{1}{M} \sum_m b_m v_{jm}$. The summands of the first average are independent, with

$$\mathrm{E}[\zeta_m v_{jm}] = (\vartheta^0_\ell - \vartheta^0_k)' \big(\mathrm{E}[w_{jm} v_{jm}] - \mathrm{E}[w_{jm}]\, \mathrm{E}[v_{jm}] \big) = 0$$

by (G3) and (G4), and $\|\zeta_m v_{jm}\|_{\psi_1} \le C_w C$; Bernstein with $t = \bar{u}^{\,2}/16 \ge c/16$. The summands of the second are independent, mean zero by (G4), with $\|b_m v_{jm}\|_{\psi_2} \le |b_m| C$; Hoeffding with

$t = \bar{u}^2/16$ gives the exponent $c_H M \bar{u}^4/(256C^2 M^{-1}\sum_m b_m^2) \geq c_H c M/(256C^2)$ by (39). Both are $o(M^{-\delta})$ uniformly in the value of $\bar{u}^2 \geq c$. □

B.3 The nested logit with exogenous prices

This section shows that the nested-logit first step is an instance of the general model of Appendix B.2. Here the price variable is exogenous, so the covariates $w_{jm} = (p_{jm}, \mathbf{x}'_{jm})'$ are $\mathcal{X}_m$-measurable. Appendix B.4 extends the results to endogenous prices. The true partition k^0 of the general model is the nesting structure k.

Conditions. Conditions (C1)–(C3) are those of Section 5.2. With exogenous prices, the separation and design conditions are stated on w_{jm} instead of $\tilde{w}_{jm}$.

(C5$_x$) *Separation.* There is $c > 0$ such that, for all $k \neq \ell$ and all J, M large,

$$\frac{1}{M}\sum_{m=1}^{M}\Big[(1-\sigma_k)\log \mathrm{E}\left[IV_m^k\right] - (1-\sigma_\ell)\log \mathrm{E}\left[IV_m^\ell\right]\Big]^2 \geq c.$$

(C6$_x$) *Rank.* There is $c_R > 0$ such that, for every nest k, every set $A \subset \{j : k(j) = k\}$ with $|A| \geq \underline{\pi} J/K$, and all J, M large,

$$\lambda_{\min}\Big(\frac{1}{M}\sum_{m=1}^{M}\mathrm{E}\Big[\frac{1}{|A|}\sum_{j\in A}\big(w_{jm}-\bar{w}_{Am}\big)\big(w_{jm}-\bar{w}_{Am}\big)'\Big]\Big) \geq c_R, \qquad \bar{w}_{Am} := \frac{1}{|A|}\sum_{j\in A} w_{jm}.$$

(C5$_x$) asks every pair of nests to differ in the population intercept, on average or in how it varies across markets. Under (C2)(f), (C5$_x$) is equivalent to (G7) (Proposition 2(vi)). Lemma 9 gives sufficient conditions in terms of nest sizes and composition. (C6$_x$) asks every large set of products of a nest to show variation of w across its products within markets, on average over markets and in every direction. Proposition 2(vii) derives (G8) from it, and Lemma 10 gives a pairwise sufficient condition.

Mapping objects. Recall the nest average S_{km} of (13), its centring $\mu_{km} = \mathrm{E}\, S_{km}$, and define the log error

$$r_{km} := \log S_{km} - \log \mu_{km}.$$

Write $\pi_{kJ} := J_k/J$ for the share of products in nest k.

Centre and error. Since the inclusive value of nest ℓ is $J_\ell S_{\ell m}$, the intercept in (3) is

$$\lambda_{km} = (\sigma_k - 1)\log(J_k S_{km}) - \log D_m, \qquad D_m := 1 + \sum_\ell (J_\ell S_{\ell m})^{\sigma_\ell}.$$

Replace each nest average $S_{\ell m}$ by its centring $\mu_{\ell m}$:

$$\lambda_{km} = \bar{\lambda}_{km} + e_{km}, \qquad \bar{\lambda}_{km} := (\sigma_k - 1)\log(J_k \mu_{km}) - \log D_m^0, \qquad D_m^0 := 1 + \sum_\ell (J_\ell \mu_{\ell m})^{\sigma_\ell}, \qquad e_{km} := \lambda_{km} - \bar{\lambda}_{km} \tag{40}$$

where $\bar{\lambda}_{km}$ is not random. With these objects, (6) reads

$$\log \mathbb{P}_{jm} = w_{jm}'\beta_{k(j)} + \bar{\lambda}_{k(j)m} + e_{k(j)m} + v_{jm}, \tag{41}$$

which is the general model (30). Its dependent variable is $y_{jm} = \log \mathbb{P}_{jm}$ and its covariates are w_{jm}. Its slopes are $\vartheta_k^0 = \beta_k = \beta/\sigma_k$ and its centres are $\alpha_{km}^0 = \bar{\lambda}_{km}$. Its perturbation is the group-market error, $e_{jm} = e_{k(j)m}$, and its product-market error is $v_{jm} = \xi_{jm}/\sigma_{k(j)}$. The centres drift to $-\infty$ at the common rate $\log J$; Lemma 7 shows that under (C2)(f) they are bounded after removing this drift, which is (G1) with $a_{m,J} = -\log J$. No normalization of the outcome is used.

Results.

Lemma 6 (Inclusive values). *(i) Under (C2)(c),(e), there are constants $0 < c_\mu \le C_\mu < \infty$ with $c_\mu \le \mu_{km} \le C_\mu$ for all k, m, J.*

(ii) Under (C2)(c),(d),(e), for every $q \ge 1$ and $\eta \in (0, 1]$,

$$\sup_{k,m} \mathbb{P}(|r_{km}| > \eta) \le C_q \, \eta^{-2q} J_k^{-q},$$

with C_q depending on q, c_μ and the constants in (C2)(d) only.

(iii) Under (C2)(c),(d),(e), there is $C < \infty$ with $\mathrm{E}\, r_{km}^2 \le C\, J_k^{-1/2}$ for all k, m, J.

Proof. (i) Write $\delta_{jm} = \beta' w_{jm} + \xi_{jm}$. By the triangle inequality for $\|\cdot\|_{\psi_2}$, $\|\delta_{jm}\|_{\psi_2} \le \|\beta\| \, \|w_{jm}\|_{\psi_2} + \|\xi_{jm}\|_{\psi_2}$. By (C2)(e), $\|\beta\|$ is bounded. By (C2)(c), $\|w_{jm}\|_{\psi_2}$ and $\|\xi_{jm}\|_{\psi_2}$ are bounded uniformly in j, m, J. So $C_\delta := \sup_{j,m,J} \|\delta_{jm}\|_{\psi_2} < \infty$. By (F1), $C_1 := \sup_{j,m,J} \mathrm{E}\,|\delta_{jm}| < \infty$.

Since $\mu_{km} = J_k^{-1} \sum_{j \in k} \mathrm{E}\, e^{\delta_{jm}/\sigma_k}$, it suffices to bound each $\mathrm{E}\, e^{\delta_{jm}/\sigma_k}$. By (C2)(e), $\sigma_k \ge \underline{\sigma}$. So $e^{\delta_{jm}/\sigma_k} \le e^{|\delta_{jm}|/\underline{\sigma}}$. By (F2) with $\lambda = 1/\underline{\sigma}$, $\mathrm{E}\, e^{\delta_{jm}/\sigma_k} \le 2e^{C_\delta^2/(4\underline{\sigma}^2)} =: C_\mu$. By Jensen, $\mathrm{E}\, e^{\delta_{jm}/\sigma_k} \ge e^{\mathrm{E}\,\delta_{jm}/\sigma_k} \ge e^{-C_1/\underline{\sigma}} =: c_\mu$. Neither constant depends on k, m, J or π_{kJ}. Also, $c_\mu \le 1 \le C_\mu$.

(ii) Fix $q \ge 1$.

Step 1: tail of S_{km}. By definition, $S_{km} - \mu_{km} = J_k^{-1} \sum_{j \in k} \left(e^{\delta_{jm}/\sigma_k} - \mathrm{E}\, e^{\delta_{jm}/\sigma_k} \right)$. Markov's inequality and (C2)(d) with exponent $2q$ give, for every $t > 0$,

$$\mathbb{P}\big(|S_{km} - \mu_{km}| > t\big) \le \frac{C_{2q} J_k^q}{t^{2q} J_k^{2q}} = C_{2q}\, t^{-2q} J_k^{-q}. \tag{42}$$

Step 2: from S_{km} to r_{km}. Fix $\eta \in (0, 1]$ and let $t := c_\mu \eta / 2$, so $t \le c_\mu/2$. Suppose $|S_{km} - \mu_{km}| \le t$. By (i), $\mu_{km} \ge c_\mu$, so $S_{km} \ge c_\mu/2$. On $[c_\mu/2, \infty)$, $\log$ is Lipschitz with constant $2/c_\mu$. So $|r_{km}| \le (2/c_\mu)|S_{km} - \mu_{km}| \le \eta$. Hence $\{|r_{km}| > \eta\} \subset \{|S_{km} - \mu_{km}| > t\}$. By (42),

$$\mathbb{P}(|r_{km}| > \eta) \le C_{2q} (2/c_\mu)^{2q}\, \eta^{-2q} J_k^{-q}.$$

The bound does not depend on k or m. This is (ii) with $\bar{C}_q := C_{2q}(2/c_\mu)^{2q}$.

(iii) Let $A := \{|S_{km} - \mu_{km}| \le c_\mu/2\}$. Split $\mathrm{E}\, r_{km}^2 = \mathrm{E}[r_{km}^2 \mathbb{1}_A] + \mathrm{E}[r_{km}^2 \mathbb{1}_{A^c}]$.

Step 1: the part on A. As in Step 2 of (ii) with $t = c_\mu/2$, $|r_{km}| \le (2/c_\mu)|S_{km} - \mu_{km}|$ on A. By (C2)(d) with exponent 2, $\mathrm{E}(S_{km} - \mu_{km})^2 \le C_2 J_k^{-1}$. So $\mathrm{E}[r_{km}^2 \mathbb{1}_A] \le 4 c_\mu^{-2} C_2 J_k^{-1}$.

Step 2: a fourth-moment bound. Since $\log x \le x$, $\log S_{km} \le S_{km}$. Since $\log$ is concave, $\log S_{km} \ge J_k^{-1} \sum_{j \in k} \delta_{jm}/\sigma_k$. With (i), this gives

$$|r_{km}| \le S_{km} + \frac{1}{J_k} \sum_{j \in k} \frac{|\delta_{jm}|}{\underline{\sigma}} + \bar{C}_\mu, \qquad \bar{C}_\mu := \max\{|\log c_\mu|, |\log C_\mu|\}.$$

By Jensen, $S_{km}^4 \le J_k^{-1} \sum_{j \in k} e^{4|\delta_{jm}|/\underline{\sigma}}$, whose mean is bounded by (F2). By Jensen and (F1), the fourth moment of the average of the $|\delta_{jm}|$ is bounded. So $C_r := \sup_{k,m,J} \mathrm{E}\, r_{km}^4 < \infty$.

Step 3: the part off A. By Cauchy–Schwarz, $\mathrm{E}[r_{km}^2 \mathbb{1}_{A^c}] \le C_r^{1/2} \mathbb{P}(A^c)^{1/2}$. By (42) with $q = 1$ and $t = c_\mu/2$, $\mathbb{P}(A^c) \le 4 c_\mu^{-2} C_2 J_k^{-1}$. So $\mathrm{E}[r_{km}^2 \mathbb{1}_{A^c}] \le 2 c_\mu^{-1} (C_2 C_r)^{1/2} J_k^{-1/2}$.

Steps 1 and 3 give $\mathrm{E}\, r_{km}^2 \le C J_k^{-1/2}$, since $J_k \ge 1$. □

Lemma 7 (Centres). *Under (C1), (C2)(c),(e),(f), for all k, m and J,*

$$|\bar{\lambda}_{km} + \log J| \le C_\lambda, \tag{43}$$

for some finite C_λ.

Proof. Write $J_\ell = \pi_{\ell J} J$ and multiply the numerator and the denominator inside the logarithms of (40) by $J^{-\sigma_k}$:

$$\bar{\lambda}_{km} + \log J = \log \frac{(\pi_{kJ}\mu_{km})^{\sigma_k - 1}}{\sum_\ell J^{\sigma_\ell - \sigma_k} (\pi_{\ell J}\mu_{\ell m})^{\sigma_\ell} + J^{-\sigma_k}}.$$

By (C1), $\pi_{\ell J} \in [\underline{\pi}, 1]$, and by Lemma 6(i), $\mu_{\ell m} \in [c_\mu, C_\mu]$. So $|\log(\pi_{\ell J}\mu_{\ell m})| \le C_\pi + \bar{C}_\mu$ for every ℓ, m, with $C_\pi := |\log \underline{\pi}|$ and $\bar{C}_\mu := \max\{|\log c_\mu|, |\log C_\mu|\}$. By (C2)(f), there is C_D with $J^{\sigma_\ell - \sigma_k} \le e^{C_D}$ for all J. Split the log into the log of the numerator and the log of the denominator:

$$\begin{aligned} |\bar{\lambda}_{km} + \log J| &\le \underbrace{|\sigma_k - 1| \left|\log(\pi_{kJ}\mu_{km})\right|}_{\le C_\pi + \bar{C}_\mu} + \underbrace{\left|\log\Big(\sum_\ell J^{\sigma_\ell - \sigma_k}(\pi_{\ell J}\mu_{\ell m})^{\sigma_\ell} + J^{-\sigma_k}\Big)\right|}_{\le C_D + C_\pi + \bar{C}_\mu + \log(K+1)} \\ &\le 2(C_\pi + \bar{C}_\mu) + C_D + \log(K+1) =: C_\lambda. \end{aligned}$$

□

Lemma 8 (Errors). *For all k and m,*

$$|e_{km}| \le |r_{km}| + \max_{\ell \le K} |r_{\ell m}| \le 2 \max_{\ell \le K} |r_{\ell m}|. \tag{44}$$

Proof. By (40),

$$e_{km} = \lambda_{km} - \bar{\lambda}_{km} = (\sigma_k - 1)\big(\log S_{km} - \log \mu_{km}\big) - \big(\log D_m - \log D_m^0\big) = (\sigma_k - 1)\, r_{km} - \log \frac{D_m}{D_m^0}.$$

The first term is at most $|r_{km}|$ in absolute value, since $|\sigma_k - 1| < 1$. For the second, $S_{\ell m} = \mu_{\ell m} e^{r_{\ell m}}$ gives $(J_\ell S_{\ell m})^{\sigma_\ell} = (J_\ell \mu_{\ell m})^{\sigma_\ell} e^{\sigma_\ell r_{\ell m}}$, so

$$\frac{D_m}{D_m^0} = \sum_\ell \omega_{\ell m}\, e^{\sigma_\ell r_{\ell m}} + \omega_{0m}, \qquad \omega_{\ell m} := \frac{(J_\ell \mu_{\ell m})^{\sigma_\ell}}{D_m^0}, \quad \omega_{0m} := \frac{1}{D_m^0},$$

where the weights $\omega_{0m}, \omega_{1m}, \ldots, \omega_{Km}$ are non-negative and sum to one. A weighted average lies between the smallest and the largest of its terms, so

$$\min\{1, \min_{\ell} e^{\sigma_\ell r_{\ell m}}\} \leq \frac{D_m}{D_m^0} \leq \max\{1, \max_{\ell} e^{\sigma_\ell r_{\ell m}}\},$$

and taking logs, $\left|\log(D_m/D_m^0)\right| \leq \max_\ell \sigma_\ell |r_{\ell m}| \leq \max_\ell |r_{\ell m}|$. The triangle inequality gives (44). □

Proposition 2 (Verification of (G1)–(G9) for the nested logit)**.** *Consider (41), with centres $\bar{\lambda}_{km}$, group-market error $e_{jm} = e_{k(j)m}$ and product-market error v_{jm} as in (40).*

(i) Under (C1), (C2)(c),(e),(f), (G1) holds.

(ii) Under (C2)(a), (G2) holds.

(iii) Under (C2)(a),(b),(c),(e), (G3) and (G4) hold.

(iv) Under (C2)(c), (G5) holds.

(v) Under (C1), (C2)(c),(d),(e), (G6) holds.

(vi) Under (C2)(c),(e),(f), (G7) holds if and only if (C5$_x$) holds, with possibly different constants.

(vii) Under (C1), (C2)(a),(c), (C3) and (C6$_x$), (G8) holds.

(viii) Under (C3), (G9) holds.

Proof. (i) By (C2)(e), $\|\beta_k\| = \|\beta\|/\sigma_k \leq \|\beta\|/\underline{\sigma}$ for all k. By Lemma 7, $|\bar{\lambda}_{km} + \log J| \leq C_\lambda$ for all k, m, M and J. The centres are non-random by (40). So (G1) holds with $\Theta := \{\vartheta : \|\vartheta\| \leq \|\beta\|/\underline{\sigma}\}$, $\mathcal{A} := [-C_\lambda, C_\lambda]$ and $a_{m,J} := -\log J$.

(ii) w_{jm} is $\mathcal{X}_m$-measurable and $v_{jm} = \xi_{jm}/\sigma_{k(j)}$ is a function of market m's primitives. So the market blocks $((w_{jm}, v_{jm})_{j \leq J})_m$ are independent across m by (C2)(a), which is (G2).

(iii) By (C2)(a), $(\xi_{jm}, \mathcal{X}_m)$ is independent of $\mathcal{X}_n$ for $n \neq m$. So $\mathrm{E}[\xi_{jm} \mid \mathcal{X}] = \mathrm{E}[\xi_{jm} \mid \mathcal{X}_m] = 0$ by (C2)(b), and $\mathrm{E}[v_{jm}] = 0$. Since w_{jm} is $\mathcal{X}_m$-measurable, $\mathrm{E}[w_{jm}v_{jm}] = \mathrm{E}[w_{jm}\,\mathrm{E}[v_{jm} \mid \mathcal{X}_m]] = 0$. By (C2)(c),(e), $\sup \|v_{jm}\|_{\psi_2} \leq \sup \|\xi_{jm}\|_{\psi_2}/\underline{\sigma} < \infty$. For all j and i, $|\mathrm{E}[v_{jm}v_{im}]| = \sigma_{k(j)}^{-1}\sigma_{k(i)}^{-1}\left|\mathrm{E}\left[\mathrm{E}[\xi_{jm}\xi_{im} \mid \mathcal{X}_m]\right]\right| \leq \underline{\sigma}^{-2}\left\|\mathrm{E}[\xi_{jm}\xi_{im} \mid \mathcal{X}_m]\right\|_2$, by (C2)(e) and Cauchy–Schwarz. Average over markets and sum over (j, i). By (C2)(b), the bound in (G4) holds with $L := L_\xi/\underline{\sigma}^2$.

(iv) $w_{jm} = (p_{jm}, \mathbf{x}'_{jm})'$. By (C2)(c), $\sup \|w_{jm}\|_{\psi_2} < \infty$, which is (G5).

(v) By Lemma 8, $e_{km}^2 \le 4 \max_\ell r_{\ell m}^2 \le 4 \sum_\ell r_{\ell m}^2$ for every k. So

$$\max_{j \le J} \frac{1}{M} \sum_m e_{jm}^2 \le \frac{4}{M} \sum_m \sum_\ell r_{\ell m}^2.$$

Taking expectations and applying Lemma 6(iii),

$$\mathrm{E}\Big[\max_{j \le J} \frac{1}{M} \sum_m e_{jm}^2 \Big] \le 4KC \big(\min_k J_k \big)^{-1/2}.$$

Under (C1), $\min_k J_k \ge \underline{\pi} J \to \infty$, so the right side tends to zero. Markov's inequality gives (G6).

(vi) Fix $k \ne \ell$ and a product j. Write $b_m := \bar{\lambda}_{\ell m} - \bar{\lambda}_{km}$, $X_m := w'_{jm}(\beta_\ell - \beta_k)$ and $B := M^{-1} \sum_m b_m^2$.

Step 1: b_m is the bracket in (C5$_x$). By (40), $\bar{\lambda}_{km} = (\sigma_k - 1) \log(J_k \mu_{km}) - \log D_m^0$. The last term does not depend on k, so it cancels in b_m. Since $J_k \mu_{km} = \mathrm{E}[IV_m^k]$,

$$b_m = (1 - \sigma_k) \log \mathrm{E}\left[IV_m^k\right] - (1 - \sigma_\ell) \log \mathrm{E}\left[IV_m^\ell\right].$$

So (C5$_x$) says that $B \ge c$.

Step 2: the slope part is small. By (C2)(e),(f), $\|\beta_\ell - \beta_k\| = \|\beta\| \, |\sigma_k - \sigma_\ell| / (\sigma_k \sigma_\ell) \le C / \log J$. By (C2)(c) and (F1), $\mathrm{E}\,\|w_{jm}\|^2 \le C$. So $|\,\mathrm{E}\, X_m| \le C / \log J$ and $\mathrm{E}\, X_m^2 \le C / \log^2 J$.

Step 3: (C5$_x$) implies (G7). Since b_m is non-random, $\mathrm{E}(X_m + b_m)^2 = b_m^2 + 2 b_m \,\mathrm{E}\, X_m + \mathrm{E}\, X_m^2 \ge b_m^2 - 2|b_m| \,|\,\mathrm{E}\, X_m|$. Average over markets and apply Cauchy–Schwarz:

$$\frac{1}{M} \sum_m \mathrm{E}(X_m + b_m)^2 \ge B - \frac{2C\sqrt{B}}{\log J} = \sqrt{B}\Big(\sqrt{B} - \frac{2C}{\log J} \Big) \ge \frac{c}{2},$$

once $2C / \log J \le \sqrt{c}/2$. This is (G7) with $c/2$.

Step 4: (G7) implies (C5$_x$). Since $(x+y)^2 \le 2x^2 + 2y^2$, $\frac{1}{M} \sum_m \mathrm{E}(X_m + b_m)^2 \le 2B + 2C / \log^2 J$. So (G7) with constant c gives $B \ge c/2 - C / \log^2 J \ge c/4$ for J large.

(vii) For a set A of products write $n_A := |A|/J$ and

$$\Lambda(A) := \frac{1}{M} \sum_m \frac{1}{|A|} \sum_{j \in A} \big(w_{jm} - \bar{w}_{Am} \big) \big(w_{jm} - \bar{w}_{Am} \big)',$$

so that (C6$_x$) says $\lambda_{\min}(\mathrm{E}\,\Lambda(A)) \geq c_R$ for large subsets of a nest. Write $\bar{W} := \max_{j \leq J} M^{-1}\|w_j\|^2$.

Step 1: the design eigenvalue and $\Lambda(A)$. Fix k, ℓ, $\tilde{k}$ and let $A := \{j : k(j) = k,\ \tilde{k}(j) = \ell\}$, with $n_A > 0$. The matrix in (G8) is

$$\begin{pmatrix} P_A & Q_A \\ Q_A' & n_A I_M \end{pmatrix}, \qquad P_A := \frac{1}{J}\sum_{j \in A} \frac{1}{M} \sum_m w_{jm} w_{jm}',$$

where Q_A is the $d \times M$ matrix with columns $(J\sqrt{M})^{-1} \sum_{j \in A} w_{jm}$. Its Schur complement is $P_A - Q_A Q_A'/n_A = n_A \Lambda(A)$. Take $(x, y) \in \mathbb{R}^d \times \mathbb{R}^M$ with $\|x\|^2 + \|y\|^2 = 1$ and let $z := y + Q_A' x / n_A$. The quadratic form equals

$$x' P_A x + 2x' Q_A y + n_A \|y\|^2 = n_A\, x' \Lambda(A) x + n_A \|z\|^2.$$

By Cauchy–Schwarz, $\|Q_A\|^2 \leq \|Q_A\|_F^2 \leq n_A^2 \bar{W}$. So $\|y\| \leq \|z\| + \sqrt{\bar{W}}\|x\|$, $\|y\|^2 \leq 2\|z\|^2 + 2\bar{W}\|x\|^2$, and $1 \leq (1 + 2\bar{W})\|x\|^2 + 2\|z\|^2$. The form is at least $n_A \lambda_{\min}(\Lambda(A))\|x\|^2 + n_A\|z\|^2$. This is at least $n_A \min\{\lambda_{\min}(\Lambda(A))/(1 + 2\bar{W}), 1/2\}$ times $(1 + 2\bar{W})\|x\|^2 + 2\|z\|^2 \geq 1$. Hence

$$\hat{\rho}_{k\ell}(\tilde{k}) \ \geq\ n_A \min\Big\{ \frac{\lambda_{\min}(\Lambda(A))}{1 + 2\bar{W}},\ \frac{1}{2} \Big\}. \tag{45}$$

Step 2: a large piece. Fix k and $\tilde{k}$. The K sets $\{j : k(j) = k,\ \tilde{k}(j) = \ell\}$, $\ell \leq K$, split nest k, which has $J_k \geq \underline{\pi} J$ products by (C1). So at least one of them has $|A| \geq \underline{\pi} J/K$, that is, $n_A \geq \underline{\pi}/K$.

Step 3: uniform concentration over sets. For $i, j \leq J$ let $D_{ij} := M^{-1} \sum_m (w_{im} - w_{jm})(w_{im} - w_{jm})'$. For every nonempty A, the identity

$$\frac{1}{|A|} \sum_{j \in A} \big(w_{jm} - \bar{w}_{Am}\big)\big(w_{jm} - \bar{w}_{Am}\big)' = \frac{1}{2|A|^2} \sum_{i \in A} \sum_{j \in A} (w_{im} - w_{jm})(w_{im} - w_{jm})'$$

gives $\Lambda(A) = (2|A|^2)^{-1} \sum_{i \in A} \sum_{j \in A} D_{ij}$. An average is at most a maximum, so

$$\sup_A \big\| \Lambda(A) - \mathrm{E}\,\Lambda(A) \big\| \leq \frac{1}{2} \max_{i,j \leq J} \big\| D_{ij} - \mathrm{E}\, D_{ij} \big\|,$$

where the supremum is over all nonempty sets of products. Each entry of $D_{ij} - \mathrm{E}\, D_{ij}$ is an average over markets of terms that are independent by (C2)(a), have mean zero, and are sub-

exponential with norm bounded uniformly by (C2)(c). Bernstein's inequality and a union bound over the J^2 pairs and the d^2 entries give, for every $t \in (0,1]$,

$$\mathbb{P}\Big(\max_{i,j \le J} \big\| D_{ij} - \mathrm{E}\, D_{ij} \big\| > t \Big) \le 2d^2 J^2 e^{-c\, M t^2},$$

for a constant $c > 0$. This tends to zero by (C3).

Step 4: a bound on $\bar{W}$. As in the proof of Lemma 5(i), Bernstein's inequality gives, for every $\delta > 0$ and uniformly in j, $\mathbb{P}(M^{-1}\|w_j\|^2 \ge \bar{M}) = o(M^{-\delta})$. A union bound over j and (C3) give $\mathbb{P}(\bar{W} \ge \bar{M}) \to 0$.

Step 5: conclusion. Consider the event $\{\max_{i,j} \|D_{ij} - \mathrm{E}\, D_{ij}\| \le c_R\} \cap \{\bar{W} < \bar{M}\}$, whose probability tends to one by Steps 3 and 4. On it, fix k and $\tilde{k}$, and take the piece A of Step 2. By (C6$_x$), Step 3 and Weyl's inequality, $\lambda_{\min}(\Lambda(A)) \ge c_R - c_R/2 = c_R/2$. By (45) and $n_A \ge \underline{\pi}/K$,

$$\max_{\ell \le K} \hat{\rho}_{k\ell}(\tilde{k}) \;\ge\; \rho := \frac{\underline{\pi}}{K} \min\Big\{ \frac{c_R}{2(1+2\bar{M})},\, \frac{1}{2} \Big\}.$$

This holds for all k and $\tilde{k}$ at once, which is (G8).

(viii) The first half of (C3), $J = O(M^{q_1})$, is (G9) with $q = q_1$. □

From (C) to (G). Proposition 2 verifies the general conditions for (41). The map is:

General condition	Implied by	Where
(G1) bounded slopes and centres, up to the drift $-\log J$	(C1), (C2)(c),(e),(f)	Proposition 2(i)
(G2) independent markets	(C2)(a)	Proposition 2(ii)
(G3), (G4) orthogonality, tails, bound L	(C2)(a),(b),(c),(e)	Proposition 2(iii)
(G5) tails of w	(C2)(c)	Proposition 2(iv)
(G6) perturbation $e_{jm} = e_{k(j)m}$ vanishes	(C1), (C2)(c),(d),(e)	Proposition 2(v)
(G7) separation	(C2)(c),(e),(f), (C5$_x$)	Proposition 2(vi)
(G8) design	(C1), (C2)(a),(c), (C3), (C6$_x$)	Proposition 2(vii)
(G9) $J = O(M^q)$	(C3), with $q = q_1$	Proposition 2(viii)

Theorem B.2 (Misclassification, nested-logit first step)**.** *Under (C1)–(C3), (C5$_x$) and (C6$_x$), the classifier (31) applied to (41) satisfies, for every $\varsigma > 0$,*

$$\mathbb{P}\big(\hat{k} \ne \tau \circ k \text{ for every permutation } \tau\big) = o(1) + o\big(JM^{-\varsigma}\big).$$

In particular, the probability that the estimated partition differs from the true one tends to zero.

Proof. Proposition 2 verifies (G1)–(G9). Apply Theorem B.1. □

Sufficient conditions for (C5$_x$). By Step 1 of the proof of Proposition 2(vi), with $J_k = \pi_{kJ} J$, the bracket in (C5$_x$) is

$$\bar{\lambda}_{\ell m} - \bar{\lambda}_{km} = (\sigma_\ell - \sigma_k) \log J + (\sigma_\ell - 1) \log \left(\pi_{\ell J} \mu_{\ell m}\right) - (\sigma_k - 1) \log \left(\pi_{kJ} \mu_{km}\right). \tag{46}$$

The first term is bounded under (C2)(f). So nests are separated by three sources: the drift constant $(\sigma_\ell - \sigma_k) \log J$, the nest sizes, through $\log \pi_{kJ}$, and the composition terms $\log \mu_{km}$, which differ when the distribution of (w, ξ) differs across nests. Under (C2)(f) the drift constant need not vanish, so it can offset the other two. The centres of two nests can then coincide even though $\pi_{kJ}\mu_{km} \neq \pi_{\ell J}\mu_{\ell m}$. The next lemma gives two conditions that rule this out.

Lemma 9 (Sufficient conditions for separation)**.** *Assume (C1), (C2)(c),(e),(f). For $k \neq \ell$ let $L_{k\ell m} := \log \left(\pi_{kJ}\mu_{km}/(\pi_{\ell J}\mu_{\ell m})\right)$ and $\bar{L}_{k\ell} := M^{-1} \sum_m L_{k\ell m}$. Suppose there is $c_L > 0$ such that each pair $k \neq \ell$ satisfies (i) or (ii):*

(i) $(\sigma_k - \sigma_\ell) \log J \to 0$, and $M^{-1} \sum_m L_{k\ell m}^2 \geq c_L$ for all J, M large;

(ii) $M^{-1} \sum_m \left(L_{k\ell m} - \bar{L}_{k\ell}\right)^2 \geq c_L$ for all J, M large.

Then (C5$_x$) holds.

Proof. Fix $k \neq \ell$ and let b_m be the bracket in (C5$_x$). Since $-(1 - \sigma_\ell) = -(1 - \sigma_k) - (\sigma_k - \sigma_\ell)$, (46) gives

$$b_m = (\sigma_\ell - \sigma_k) \log J + (1 - \sigma_k) L_{k\ell m} + \epsilon_m, \qquad \epsilon_m := -(\sigma_k - \sigma_\ell) \log \left(\pi_{\ell J} \mu_{\ell m}\right).$$

By (C1) and Lemma 6(i), $|\log(\pi_{\ell J}\mu_{\ell m})| \leq C$. By (C2)(f), $|\sigma_k - \sigma_\ell| \leq C/\log J$. So $|\epsilon_m| \leq C/\log J$. By (C2)(e), $1 - \sigma_k \geq 1 - \bar{\sigma}$.

(i) The term $(\sigma_\ell - \sigma_k) \log J$ tends to zero. By the triangle inequality,

$$M^{-1/2} \|b\| \geq (1 - \bar{\sigma}) \sqrt{c_L} - \left|(\sigma_\ell - \sigma_k) \log J\right| - \frac{C}{\log J},$$

which is at least $(1-\bar{\sigma})\sqrt{c_L}/2$ for J large.

(ii) The term $(\sigma_\ell - \sigma_k)\log J$ does not vary across markets. So, with $\bar{b}$ and $\bar{\epsilon}$ the averages over markets, $b_m - \bar{b} = (1-\sigma_k)(L_{k\ell m} - \bar{L}_{k\ell}) + (\epsilon_m - \bar{\epsilon})$. Since $M^{-1/2}\|b\| \geq M^{-1/2}\|b - \bar{b}\|$ and $M^{-1/2}\|\epsilon - \bar{\epsilon}\| \leq M^{-1/2}\|\epsilon\|$,

$$M^{-1/2}\|b\| \geq (1-\bar{\sigma})\sqrt{c_L} - \frac{C}{\log J},$$

which is at least $(1-\bar{\sigma})\sqrt{c_L}/2$ for J large.

In both cases, $M^{-1}\sum_m b_m^2 \geq (1-\bar{\sigma})^2 c_L/4$. Since K is fixed, this holds for all pairs at once, which is (C5$_x$) with $c := (1-\bar{\sigma})^2 c_L/4$. □

Condition (i) strengthens (C2)(f) to $(\sigma_k - \sigma_\ell)\log J \to 0$ and asks the nests to differ in size or composition. Condition (ii) keeps (C2)(f) and asks the relative attractiveness of the two nests to change across markets. Under (C2)(f), a shift in utility common to all products of a market changes $L_{k\ell m}$ only by $O(1/\log J)$, so it helps with neither condition.

A sufficient condition for (C6$_x$). (C6$_x$) is stated over all large sets of products of a nest. A pairwise condition is easier to read.

Lemma 10 (Sufficient condition for design)**.** *Assume (C1) and (C2)(c). Suppose there is $c_P > 0$ such that, for all $i \neq j$ with $k(i) = k(j)$ and all J, M large,*

$$\lambda_{\min}\Big(\frac{1}{M}\sum_{m=1}^{M} \operatorname{var}\big(w_{im} - w_{jm}\big)\Big) \geq c_P.$$

Then (C6$_x$) holds with $c_R := c_P/4$.

Proof. Fix a nest k and a set $A \subset \{j : k(j) = k\}$ with $|A| \geq \underline{\pi}J/K$. Then $|A| \geq 2$ for J large. By the identity in Step 3 of the proof of Proposition 2(vii), the matrix in (C6$_x$) equals

$$\frac{1}{2|A|^2}\sum_{i\in A}\sum_{j\in A}\frac{1}{M}\sum_m \mathrm{E}\left[(w_{im} - w_{jm})(w_{im} - w_{jm})'\right].$$

For each pair, $\mathrm{E}[XX'] \geq \operatorname{var}(X)$ in the positive semidefinite order. The terms with $i = j$ are zero. Each of the $|A|(|A|-1)$ terms with $i \neq j$ has smallest eigenvalue at least c_P, and

the smallest eigenvalue is superadditive over sums of symmetric matrices. So the smallest eigenvalue of the matrix in (C6$_x$) is at least $c_P(|A|-1)/(2|A|) \geq c_P/4$. □

The pairwise condition asks that no two products of a nest have covariates that move together across markets in any direction. It fails when characteristics are fixed across markets, $\mathbf{x}_{jm} = \mathbf{x}_j$, since then $\text{var}(\mathbf{x}_{im} - \mathbf{x}_{jm}) = 0$. In that case (C6$_x$) can still hold through the dispersion of the $\mathbf{x}_j$ across the products of a nest, which the pairwise condition does not use.

B.4 Endogenous prices: one-stage classifier

Prices may now depend on the demand shocks. There is an instrument z_{jm} in $\mathcal{X}_m$, with $\mathrm{E}[\xi_{jm} \mid \mathcal{X}_m] = 0$ by (C2)(b). The price projection is (8), with price groups equal to the nests. This section studies a one-stage classifier that substitutes the projection into the share equation and classifies on the exogenous variables. Section 4.1 uses the two-stage classifier instead; Appendix B.5 treats it. Write $\tilde{w}_{jm} := (\mathbf{x}'_{jm}, z'_{jm})'$ for the exogenous variables. Inserting (8) in δ_{jm}, and the result in (41),

$$\log \mathbb{P}_{jm} = \tilde{w}'_{jm}\tilde{\beta}_{k(j)} + \bar{\bar{\lambda}}_{k(j)m} + e_{k(j)m} + \tilde{v}_{jm}, \tag{47}$$

where

$$\tilde{\beta}_k := \frac{1}{\sigma_k}\begin{pmatrix} \beta_x + \beta_p\gamma_x^k \\ \beta_p\gamma_z^k \end{pmatrix}, \qquad \bar{\bar{\lambda}}_{km} := \bar{\lambda}_{km} + \frac{\beta_p}{\sigma_k}\kappa_{km}, \qquad \tilde{v}_{jm} := \frac{\xi_{jm} + \beta_p\omega_{jm}}{\sigma_{k(j)}},$$

and $\bar{\lambda}_{km}$ and e_{km} are as in (40). So (47) is the general model (30) with covariates $\tilde{w}_{jm}$, slopes $\tilde{\beta}_k$, centres $\bar{\bar{\lambda}}_{km}$, perturbation $e_{k(j)m}$ and product-market error $\tilde{v}_{jm}$. The one-stage classifier is (31) applied to (47) with covariates $\tilde{w}_{jm}$, that is,

$$\big(\hat{k}(1), \ldots, \hat{k}(J)\big) = \underset{k(1),\ldots,k(J)}{\text{argmin}} \left\{ \min_{\boldsymbol{\theta},\boldsymbol{\alpha}} \sum_{j=1}^{J}\sum_{m=1}^{M} \big(\log(J\mathbb{P}_{jm}) - \theta'_{k(j),z}z_{jm} - \theta'_{k(j),x}\mathbf{x}_{jm} - \alpha_{k(j)m}\big)^2 \right\}. \tag{48}$$

Conditions. Conditions (C4) are those of Section 5.2. The one-stage classifier replaces (C5)(a),(b) by a single separation condition, and uses the design condition (C6)(a).

(C5$_p$) *Separation, one-stage classifier.* Let $b_m^{k\ell}$ be the bracket in (C5)(a). There is $c > 0$ such that, for all $k \neq \ell$, all j and all J, M large,

$$\frac{1}{M}\sum_{m=1}^{M} \mathrm{E}\left[\left(b_m^{k\ell} + \frac{\beta_p}{\sigma_k}\big((\gamma^\ell - \gamma^k)'\tilde{w}_{jm} + \kappa_{\ell m} - \kappa_{km}\big)\right)^2\right] \geq c.$$

Under (C2)(f), the contrast in (G7) for (47) equals the term inside the square in (C5$_p$), up to $O(1/\log J)$. Its price term is β_p/σ_k times the gap between the prices that the projections of nests ℓ and k predict for product j. With a common price equation the gap is zero, and (C5$_p$) is (C5)(a). With nest-specific price equations, the gap is an additional source of separation, and it can also offset $b_m^{k\ell}$. In (C6)(a) the covariates are $(\mathbf{x}, z)$. Prices do not enter the design, and the instrument does. Lemma 10 applies with $\tilde{w}_{jm}$ in place of w_{jm}.

Proposition 3 (Verification of (G1)–(G9) with endogenous prices)**.** *Consider the reduced form (47).*

(i) Under (C1), (C2)(c),(e), (C4)(c) and (C2)(f), (G1) holds.

(ii) Under (C2)(a), (G2) holds.

(iii) Under (C2)(a),(b),(c),(e), (C4), (G3) and (G4) hold.

(iv) Under (C2)(c) and (C4)(b), (G5) holds.

(v) Under (C1), (C2)(c),(d),(e), (G6) holds.

(vi) Under (C2)(c),(e), (C4)(b),(c) and (C2)(f), (G7) holds if and only if (C5$_p$) holds, with possibly different constants.

(vii) Under (C1), (C2)(a),(c), (C3), (C4)(b) and (C6)(a), (G8) holds.

(viii) Under (C3), (G9) holds.

Proof. (i) Let $\bar{\gamma} := \sup_{k,J} \|\gamma^k\|$ and $\bar{\kappa} := \sup_{k,m,J} |\kappa_{km}|$, which are finite by (C4)(c). By (C2)(e), $\|\tilde{\beta}_k\| \leq \|\beta\|(1+\bar{\gamma})/\underline{\sigma}$. By Lemma 7, $|\bar{\tilde{\lambda}}_{km} + \log J| \leq C_\lambda + \|\beta\|\bar{\kappa}/\underline{\sigma}$. The slopes and centres are non-random. So (G1) holds with these bounds and $a_{m,J} = -\log J$.

(ii) $\tilde{w}_{jm}$ is $\mathcal{X}_m$-measurable. $\tilde{v}_{jm}$ is a function of market m's primitives, since $\omega_{jm} = p_{jm} - \gamma^{k\prime}\tilde{w}_{jm} - \kappa_{km}$. So the market blocks $((\tilde{w}_{jm}, \tilde{v}_{jm})_{j\leq J})_m$ are independent across m by (C2)(a), which is (G2).

(iii) As in Proposition 2(iii), $\mathrm{E}[\xi_{jm} \mid \mathcal{X}] = 0$, and $\tilde{w}_{jm}$ is $\mathcal{X}_m$-measurable. So $\mathrm{E}[\xi_{jm}] = 0$ and $\mathrm{E}[\tilde{w}_{jm}\xi_{jm}] = 0$. By (C4)(a), $\mathrm{E}[\omega_{jm}] = 0$ and $\mathrm{E}[\tilde{w}_{jm}\omega_{jm}] = 0$. So $\mathrm{E}[\tilde{v}_{jm}] = 0$ and $\mathrm{E}[\tilde{w}_{jm}\tilde{v}_{jm}] = 0$. By (C2)(c), (C4)(b),(c), ω_{jm} and ξ_{jm} are subgaussian, uniformly in j, m and J. With (C2)(e), so is $\tilde{v}_{jm}$. For the bound L, average over markets. Since $\sigma_{k(j)}$ and $\sigma_{k(i)}$ do not depend on m,

$$\sigma_{k(j)}\sigma_{k(i)}\frac{1}{M}\sum_m \mathrm{E}[\tilde{v}_{jm}\tilde{v}_{im}] = \frac{1}{M}\sum_m \mathrm{E}[\xi_{jm}\xi_{im}] + \beta_p\frac{1}{M}\sum_m \mathrm{E}[\xi_{jm}\omega_{im}] + \beta_p\frac{1}{M}\sum_m \mathrm{E}[\omega_{jm}\xi_{im}] + \beta_p^2\frac{1}{M}\sum_m \mathrm{E}[\omega_{jm}\omega_{im}].$$

Take absolute values and sum over (j, i). As in Proposition 2(iii), the first term contributes at most L_ξ to the double sum in (G4), by (C2)(b). The second and third terms give the same sum after exchanging j and i. Move the absolute values inside the averages over markets. By (C4)(a), each contributes at most $|\beta_p|L_\omega$, and the last term at most $\beta_p^2 L_\omega$. So the bound in (G4) holds with $L := \big(L_\xi + (2|\beta_p| + \beta_p^2)L_\omega\big)/\underline{\sigma}^2$.

(iv) $\tilde{w}_{jm} = (\mathbf{x}'_{jm}, z'_{jm})'$. By (C2)(c) and (C4)(b), $\sup \|\tilde{w}_{jm}\|_{\psi_2} < \infty$, which is (G5).

(v) The perturbation e_{km} is that of (41). The proof of Proposition 2(v) uses the price only through δ_{jm} and (C2)(c),(d), so it applies unchanged.

(vi) Fix $k \neq \ell$ and a product j. Let $X_m := \tilde{w}'_{jm}(\tilde{\beta}_\ell - \tilde{\beta}_k) + \bar{\tilde{\lambda}}_{\ell m} - \bar{\tilde{\lambda}}_{km}$ be the contrast in (G7), and let Y_m be the term inside the square in (C5$_p$). For arrays of random variables, write $\|X\| := (M^{-1}\sum_m \mathrm{E}\, X_m^2)^{1/2}$, which satisfies the triangle inequality. (G7) says $\|X\|^2 \geq c$, and (C5$_p$) says $\|Y\|^2 \geq c$. By Step 1 of the proof of Proposition 2(vi), $\bar{\lambda}_{\ell m} - \bar{\lambda}_{km} = b_m^{k\ell}$. Collecting terms,

$$X_m - Y_m = \Big(\frac{1}{\sigma_\ell} - \frac{1}{\sigma_k}\Big)\Big(\beta_x'\mathbf{x}_{jm} + \beta_p\big(\gamma^{\ell\prime}\tilde{w}_{jm} + \kappa_{\ell m}\big)\Big).$$

By (C2)(e),(f), $|1/\sigma_\ell - 1/\sigma_k| \leq C/\log J$. By (C2)(c),(e), (C4)(b),(c), the second factor has a second moment bounded by C. So $\|X - Y\| \leq C/\log J$, and the triangle inequality gives $\big|\|X\| - \|Y\|\big| \leq C/\log J$. If (C5$_p$) holds with c, then $\|X\| \geq \sqrt{c} - C/\log J \geq \sqrt{c}/2$ for J large, which is (G7) with $c/4$. The converse is the same argument with X and Y exchanged.

(vii) The proof of Proposition 2(vii) uses the covariates only through (C1), the independence of their market blocks, a uniform bound on their subgaussian norms, (C3) and (C6$_x$). For $\tilde{w}_{jm}$, these follow from (C1), (C2)(a),(c) with (C4)(b), (C3) and (C6)(a).

(viii) As in Proposition 2(viii). □

Theorem B.3 (Misclassification, one-stage classifier)**.** *Under (C1)–(C4), (C5$_p$) and (C6)(a), the classifier (31) applied to (47), with covariates $\tilde{w}_{jm}$, satisfies the conclusion of Theorem B.2.*

Proof. Proposition 3 verifies (G1)–(G9). Apply Theorem B.1. □

Relative to the exogenous case, (G6) and (G9) are verified as before, (C4) are added, and (C5$_x$) and (C6$_x$) become (C5$_p$) and (C6)(a).

B.5 Two-stage classifier, any price grouping, and proof of Theorem 5.1

Price groups. Let $g(\cdot)$ be a deterministic partition of the products into G price groups, with G fixed. The partition is unknown and may differ from $k(\cdot)$. With $g = g(j)$, the price projection is

$$p_{jm} = (\gamma^g)'\tilde{w}_{jm} + \kappa_{gm} + \omega_{jm}, \tag{49}$$

with $\gamma^g := \left((\gamma^g_x)', (\gamma^g_z)'\right)'$. Appendix B.4 is the case $g = k$, and a single price group is the case $G = 1$. Write $\tilde{p}_{jm} := (\gamma^{g(j)})'\tilde{w}_{jm} + \kappa_{g(j)m}$ for the predicted price and $\breve{w}_{jm} := (\tilde{p}_{jm}, \mathbf{x}'_{jm})'$.

When $g(\cdot)$ differs from $k(\cdot)$, the slopes and centres of the reduced form (47) depend on the pair $(k(j), g(j))$. So (47) no longer has K groups. The two-stage classifier (9)–(10) therefore proceeds as follows.

1. *Price stage.* Run (31) with G groups, outcome p_{jm} and covariates $\tilde{w}_{jm}$. It returns $\hat{g}$, $\hat{\gamma}^g$ and $\hat{\kappa}_{gm}$. Let $\hat{p}_{jm} := (\hat{\gamma}^{\hat{g}(j)})'\tilde{w}_{jm} + \hat{\kappa}_{\hat{g}(j)m}$.

2. *Demand stage.* Run (31) with K groups, outcome $\log \mathbb{P}_{jm}$ and covariates $\hat{w}_{jm} := (\hat{p}_{jm}, \mathbf{x}'_{jm})'$. It returns $\hat{k}$.

The price stage is the general model (30) with groups $g(\cdot)$, outcome p_{jm}, covariates $\tilde{w}_{jm}$, slopes γ^g, centres κ_{gm} with $a_{m,J} = 0$ in (G1), no perturbation and error ω_{jm}. For the demand stage, insert $p_{jm} = \tilde{p}_{jm} + \omega_{jm}$ in (41):

$$\log \mathbb{P}_{jm} = \breve{w}'_{jm}\beta_k + \bar{\lambda}_{km} + e_{km} + \tilde{v}_{jm}, \tag{50}$$

with $\beta_k = \beta/\sigma_k$ and $\tilde{v}_{jm}$ as in Appendix B.4. This is the general model (30) with covariates $\breve{w}_{jm}$, observed with error $\Delta_{jm} := \hat{w}_{jm} - \breve{w}_{jm} = (\hat{p}_{jm} - \tilde{p}_{jm}, 0')'$. Its slopes are the structural slopes β/σ_k.

Conditions.

(P) *Price groups.* $g(\cdot)$ is deterministic, every price group holds at least a share $\underline{\pi}$ of the products for all J, and (C4) hold with $g(j)$ in place of $k(j)$.

(C5$_\gamma$) *Price separation.* With $g = k$ this is (C5)(b) of Section 5.2. There is $c > 0$ such that, for all price groups $g \neq g'$, all j and all J, M large,

$$\frac{1}{M}\sum_{m=1}^{M} \mathrm{E}\left[\left((\gamma^{g'} - \gamma^{g})'\tilde{w}_{jm} + \kappa_{g'm} - \kappa_{gm}\right)^2\right] \geq c.$$

(C6$_z$) *Price design.* (C6$_x$) holds with the price groups in place of the nests and $\tilde{w}_{jm}$ in place of w_{jm}. With $g = k$ this is (C6)(a) of Section 5.2.

(C6$_p$) *Demand design.* (C6$_x$) holds with $\check{w}_{jm}$ in place of w_{jm}. Since $\tilde{p}_{jm} = p_{jm} - \omega_{jm}$, this is (C6)(b) of Section 5.2.

Comments.

- (C5$_\gamma$) is (G7) for the price stage. It asks price groups to differ in their price equations. No drift term arises, because the price equation has no nest parameters.
- In the demand stage the slopes are β/σ_k, which differ by $O(1/\log J)$ under (C2)(f). So (C5)(a) alone separates the nests. The predicted-price gap of (C5$_p$) plays no role, because $\hat{p}_{jm}$ carries the price equation of each product.
- (C6$_p$) asks the predicted price and x to vary across the products of a large set of a nest within markets, in every direction. Within a price group, $\tilde{p}_{jm} = (\gamma^g)'\tilde{w}_{jm} + \kappa_{gm}$. So the instrument has to move the predicted price, $\gamma_z^g \neq 0$, and z must not be collinear with x within markets. This is the relevance of the instrument in the first step.
- When $g = k$ or $G = 1$, both Appendix B.4 and this section apply. They rest on different separation conditions: (C5$_p$) for the one-stage classifier, and (C5)(a) together with (C5$_\gamma$) for the two-stage classifier.

Proposition 4 (Two-stage classification with any price grouping)**.** *Assume (C1)–(C3), (P), (C5)(a), (C5$_\gamma$), (C6$_z$) and (C6$_p$).*

(i) *The price stage recovers the price groups: for every* $\varsigma > 0$, $\mathbb{P}\big(\hat{g} \neq \tau \circ g \text{ for every permutation } \tau\big) = o(1) + o(JM^{-\varsigma})$.

(ii) $\max_{j \leq J} M^{-1/2} \|\hat{p}_j - \tilde{p}_j\| \to_p 0$.

(iii) *The demand stage recovers the nests: for every* $\varsigma > 0$, $\mathbb{P}\big(\hat{k} \neq \tau \circ k \text{ for every permutation } \tau\big) = o(1) + o(JM^{-\varsigma})$.

Proof. (i) We verify (G1)–(G9) for the price stage. (G1): γ^g and κ_{gm} are non-random and bounded by (P). (G2): $\tilde{w}_{jm}$ is $\mathcal{X}_m$-measurable and ω_{jm} is a function of market m's primitives, so (C2)(a) applies. (G3) and (G4): $\mathrm{E}[\omega_{jm}] = 0$ and $\mathrm{E}[\tilde{w}_{jm}\omega_{jm}] = 0$ by (P). ω_{jm} is subgaussian as in the proof of Proposition 3(iii), and the bound in (G4) holds with $L = L_\omega$ by (C4)(a) under (P). (G5): (C2)(c) and (P). (G6): there is no perturbation. (G7): (C5$_\gamma$). (G8): the proof of Proposition 2(vii) applies with the price groups in place of the nests, the share bound of (P) in place of (C1), $\tilde{w}_{jm}$ in place of w_{jm}, and (C6$_z$). (G9): (C3). Theorem B.1 gives (i).

(ii) Relabel the price groups by the permutation of Proposition 1 for the price stage. By the proof of Theorem B.1, $\hat{g} = g$ with probability tending to one. On that event,

$$\hat{p}_{jm} - \tilde{p}_{jm} = \big(\hat{\gamma}^{g(j)} - \gamma^{g(j)}\big)' \tilde{w}_{jm} + \hat{\kappa}_{g(j)m} - \kappa_{g(j)m}.$$

So, for every j,

$$M^{-1/2}\|\hat{p}_j - \tilde{p}_j\| \leq \max_g \big\|\hat{\gamma}^g - \gamma^g\big\| \max_i M^{-1/2}\|\tilde{w}_i\| + \max_g \left(\frac{1}{M}\sum_m \big(\hat{\kappa}_{gm} - \kappa_{gm}\big)^2\right)^{1/2}.$$

By Proposition 1, the first factor and the last term are $o_p(1)$. By Step 4 of the proof of Proposition 2(vii), applied to $\tilde{w}_{jm}$, $\max_i M^{-1/2}\|\tilde{w}_i\| = O_p(1)$. This gives (ii).

(iii) We verify (G1)–(G9) for (50). The slopes, centres and perturbation are those of Appendix B.3. So (G1) holds as in Proposition 2(i). $\tilde{p}_{jm}$ is a non-random linear function of $\tilde{w}_{jm}$. So $\breve{w}_{jm}$ is $\mathcal{X}_m$-measurable, and (G2) holds as in Proposition 3(ii). For (G3) and (G4), with $g = g(j)$, $\mathrm{E}[\tilde{p}_{jm}\omega_{jm}] = (\gamma^g)' \mathrm{E}[\tilde{w}_{jm}\omega_{jm}] + \kappa_{gm}\,\mathrm{E}[\omega_{jm}] = 0$ by (P). The rest is as in Proposition 3(iii). (G5): $\tilde{p}_{jm}$ is subgaussian by (C2)(c) and (P). (G6) holds as in Proposition 2(v). (G7): the proof of Proposition 2(vi) uses the covariates only through a uniform bound on their second moments, which (G5) gives for $\breve{w}_{jm}$. So (G7) holds by (C5)(a). (G8): the proof of Proposition 2(vii) applies with $\breve{w}_{jm}$ in place of w_{jm} and (C6$_p$). (G9): (C3). By (ii),

$\max_j M^{-1/2}\|\Delta_j\| \to_p 0$. The demand stage is (31) applied to the outcome $\log \mathbb{P}_{jm}$ and the covariates $\hat{w}_{jm} = \check{w}_{jm} + \Delta_{jm}$. Lemma 4 gives (iii). □

Proof of Theorem 5.1. Take $g = k$ and $G = K$ in Proposition 4. Then (P) is (C4) together with (C1), and (C5$_\gamma$), (C6$_z$) and (C6$_p$) reduce to (C5)(b), (C6)(a),(b) of Section 5.2. The price stage (9) and the demand stage (10) are the two stages of the proposition. Parts (i) and (iii) give the two claims. □

B.6 The demeaned second-step estimator

In this section, $\hat{\theta}$, $\bar{g}$ and $\nabla\hat{G}$ refer to the demeaned estimator defined below, not to the levels estimator (12) of the main text.

An alternative to the levels estimator of Section 4.2 takes deviations from the product mean across markets, which removes the drift $-\log J_{k(j)}$ of the within-nest share. For any variable a_{jm}, let $\bar{a}_j \equiv M^{-1}\sum_m a_{jm}$ and $\nabla a_{jm} \equiv a_{jm} - \bar{a}_j$. Demeaning (4) gives

$$\nabla\big(\log \mathbb{P}_{jm} - \log \mathbb{P}_{0m}\big) = \beta_p \nabla p_{jm} + \beta_x' \nabla \mathbf{x}_{jm} + (1-\sigma_{k(j)})\nabla \log \mathbb{P}_{j|k(j),m} + \nabla\xi_{jm}, \tag{51}$$

where $\nabla \log \mathbb{P}_{j|k(j),m}$ is bounded in probability as J grows. Demeaning also removes product-level unobservables that do not vary across markets, so this estimator also covers a model with product fixed effects (Section 6).

Demeaned estimator. The demeaned GMM estimator solves

$$\hat{\theta} = \arg\min_\theta \ \bar{g}(\theta)'\hat{W}\bar{g}(\theta), \qquad \bar{g}(\theta) \equiv \frac{1}{JM}\sum_{j=1}^{J}\sum_{m=1}^{M} \nabla\tilde{z}_{jm}\big(\nabla(\log \mathbb{P}_{jm} - \log \mathbb{P}_{0m}) - \nabla R_{jm}'\theta\big), \tag{52}$$

where $\hat{W}$ is a positive definite weight matrix. Since $\bar{g}(\theta)$ is linear in θ, the estimator has the closed form

$$\hat{\theta} = \big(\nabla\hat{G}'\hat{W}\nabla\hat{G}\big)^{-1}\nabla\hat{G}'\hat{W}\,\frac{1}{JM}\sum_{j=1}^{J}\sum_{m=1}^{M}\nabla\tilde{z}_{jm}\nabla\big(\log \mathbb{P}_{jm} - \log \mathbb{P}_{0m}\big), \tag{53}$$

where $\nabla\hat{G} \equiv (JM)^{-1}\sum_{j=1}^{J}\sum_{m=1}^{M}\nabla\tilde{z}_{jm}\nabla R_{jm}'$ is the sample Jacobian.

Asymptotic distribution. We state the result for the demeaned oracle estimator, computed with the true nests. Its asymptotic properties depend on the moment function at the truth, $\bar{g}(\theta_0)$, and on the Jacobian $\nabla\hat{G}$. The within-nest share depends on the other products of the nest through the nest average $S_{k(j)m}$ of (13), which fluctuates around its centring $\mu_{k(j)m}$. The population Jacobian replaces the nest average by its centring. With $k = k(j)$, let

$$F_{jm} \equiv \big(p_{jm},\ \mathbf{x}'_{jm},\ \big(\sigma_k^{-1}\delta_{jm} - \log\mu_{km} - \log(J_k/J)\big)\, e_k'\big)',$$

as in (C7). The term $\log(J_k/J)$ does not vary across markets and drops out after demeaning, and define the population Jacobian and the variance of the moment function at the truth,

$$\Gamma_{JM} \equiv \frac{1}{JM}\sum_{j=1}^{J}\sum_{m=1}^{M} \mathrm{E}\left[\nabla\tilde{z}_{jm}\,\nabla F'_{jm}\right], \qquad \Omega_{JM} \equiv \frac{1}{JM}\sum_{m=1}^{M}\sum_{j=1}^{J}\sum_{i=1}^{J} \mathrm{E}\left[(\tilde{z}_{jm}-\mathrm{E}\,\bar{\tilde{z}}_j)(\tilde{z}_{im}-\mathrm{E}\,\bar{\tilde{z}}_i)'\,\xi_{jm}\xi_{im}\right]. \tag{54}$$

We assume that $\hat{W} - W_{JM} \to_p 0$ for non-random matrices W_{JM} with eigenvalues in $[c_W, C_W]$, $0 < c_W \le C_W < \infty$, and add one condition.

(C7^D) *Rank and variance, demeaned estimator.* There is $c_\Gamma > 0$ with $\lambda_{\min}(\Gamma'_{JM}\Gamma_{JM}) \ge c_\Gamma$ for all J, M large, and $\liminf_{J,M}\lambda_{\min}(\Omega_{JM}) > 0$. In addition, $\sup_{j,m,J}\mathrm{E}\,\|z_{jm}\|^8 < \infty$ and $\sup_{j,m,J}\mathrm{E}\,|h_{jm}|^8 < \infty$.

(C7^D) requires the population Jacobian to have full column rank, the variance of the moment function to be non-degenerate, and the instruments to have eight moments.

Theorem B.4 (Demeaned oracle estimator: consistency and asymptotic normality)**.** *Under (C1)–(C3) and (C7^D), $\nabla\hat{G}'\hat{W}\nabla\hat{G}$ is invertible with probability tending to one, and:*

(i) $\hat{\theta}(k) - \theta_0 = \big(\Gamma'_{JM}W_{JM}\Gamma_{JM}\big)^{-1}\Gamma'_{JM}W_{JM}\,\bar{g}(\theta_0) + o_p\big((JM)^{-1/2}\big) = O_p\big((JM)^{-1/2}\big)$;

(ii) $V_{JM}^{-1/2}\sqrt{JM}\,\big(\hat{\theta}(k) - \theta_0\big) \to_d N(0, I)$, *where*

$$V_{JM} \equiv (\Gamma'_{JM}W_{JM}\Gamma_{JM})^{-1}\Gamma'_{JM}W_{JM}\,\Omega_{JM}\,W_{JM}\Gamma_{JM}(\Gamma'_{JM}W_{JM}\Gamma_{JM})^{-1}.$$

With $W_{JM} = \Omega_{JM}^{-1}$, $V_{JM} = (\Gamma'_{JM}\Omega_{JM}^{-1}\Gamma_{JM})^{-1}$.

The result is stated with finite-sample population matrices, so no limit of the Jacobian or of the variance is assumed. Only $\mathrm{E}[\xi_{jm} \mid \mathcal{X}_m] = 0$ and the covariance bound in (C2)(b)

enter the rate. The fourth-moment bound in (C2)(b) enters once, for the Lyapunov condition in the central limit theorem. The proof is below. The main step is to show that the sample Jacobian converges to Γ_{JM} although the regressor contains the nest averages S_{km}, which are random and shared by all products in a nest.

This section studies the demeaned GMM estimator (53) when the nesting structure $k(\cdot)$ is known. Its asymptotic properties depend on two objects: the moment function at the truth, $\bar{g}(\theta_0)$, and the Jacobian, $\nabla\hat{G}$. The central limit theorem for $\bar{g}(\theta_0)$ follows standard arguments. The Jacobian requires more work. The within-nest share $\log \mathbb{P}_{j|k(j),m}$ contains the shocks ξ_{jm}, so it needs an instrument. This is the usual endogeneity of a regressor in an IV problem. The within-nest share also depends on the other products of the nest through S_{km}. S_{km} is an average over the nest, and it fluctuates around its non-random centring μ_{km}. We therefore write $\nabla\hat{G}$ as the same matrix with S_{km} replaced by μ_{km}, plus an error. The first term has expectation Γ_{JM} of (54) and the error vanishes. Throughout this section we use (C1)–(C3) of Section 5.2 and (C7^D), with F_{jm}, Γ_{JM}, Ω_{JM} and W_{JM} as defined above. (C2)(c) is used together with the moment bounds on z_{jm} and h_{jm} in (C7^D).

Within-nest shares. With S_{km} the nest average (13), the within-nest share satisfies the exact identity

$$\log \mathbb{P}_{j|k(j),m} = \frac{\delta_{jm}}{\sigma_{k(j)}} - \log J_{k(j)} - \log S_{k(j)m}. \tag{55}$$

The term $\log J_{k(j)}$ does not vary across markets and is removed by the within-product demeaning. The term $\log S_{k(j)m}$ depends on the demand shocks of all products in the nest and fluctuates around $\log \mu_{k(j)m}$.

Labelling. For a partition $\tilde{k}$ of the products into K groups, let $c(\tilde{k})$ number the groups of $\tilde{k}$ in the order of their smallest product index, as in Section 5.4. Then $c(\tau \circ \tilde{k}) = c(\tilde{k})$ for every permutation τ, and the labels $c(k)$ of the true partition are non-random. The estimators $\hat{\theta}(\hat{k})$ and $\hat{\theta}(k)$ of the main text are $\hat{\theta}(c(\hat{k}))$ and $\hat{\theta}(c(k))$ in the notation of this section. All results below for the oracle are stated for $\hat{\theta}(c(k))$, with the nest-specific components of θ_0, Γ_{JM} and V_{JM} labelled by $c(k)$.

Uniformity of the inclusive values over markets. Lemma 6(ii) bounds the tail of r_{km} for each nest-market pair. Under (C1) and (C3), a union bound over the KM nest-market pairs

gives, for every $\eta \in (0,1]$,

$$\mathbb{P}\Big(\max_{k\le K,\, m\le M} |r_{km}| > \eta\Big) \le KM \sup_{k,m} \mathbb{P}(|r_{km}| > \eta) \le C_q\, \eta^{-2q} KM (\underline{\pi} J)^{-q} \to 0 \tag{56}$$

for $q > q_2$, since $J_k \ge \underline{\pi} J$ by (C1) and $M = O(J^{q_2})$ by (C3). So $\max_{k,m\le M} |r_{km}| = o_p(1)$. This is the only place where (C1) enters Theorem B.4.

Weighted sums of demand shocks.

Lemma 11. *Under (C2)(a),(b),(c), let A_{jm} be $\mathcal{X}$-measurable arrays, vectors or matrices, with $\mathrm{E}\,\|A_{jm}\|^4$ bounded uniformly in j, m, J and M. Then:*

(i) $\mathrm{E}[\xi_{jm} \mid \mathcal{X}] = 0$, $\mathrm{E}[\xi_{jm}\xi_{im} \mid \mathcal{X}] = \mathrm{E}[\xi_{jm}\xi_{im} \mid \mathcal{X}_m]$, *and* $\mathrm{E}[\xi_{jm}\xi_{in} \mid \mathcal{X}] = 0$ *whenever* $m \ne n$;

(ii) $\frac{1}{JM}\sum_{j=1}^{J}\sum_{m=1}^{M} A_{jm}\xi_{jm} = O_p\big((JM)^{-1/2}\big)$;

(iii) if each A_{jm} is a vector and $\mathcal{X}_m$-measurable, $\mathrm{E}\,\|A_{jm}\|^8$ is bounded uniformly in j, m, J and M, and

$$\liminf_{J,M} \lambda_{\min}(\Omega^A_{JM}) > 0, \qquad \Omega^A_{JM} := \frac{1}{JM}\sum_{m=1}^{M}\sum_{j=1}^{J}\sum_{i=1}^{J} \mathrm{E}\left[A_{jm}A'_{im}\xi_{jm}\xi_{im}\right],$$

then $(\Omega^A_{JM})^{-1/2}(JM)^{-1/2}\sum_{j=1}^{J}\sum_{m=1}^{M} A_{jm}\xi_{jm} \to_d N(0, I)$.

Proof. (i) By (C2)(a), the block $((\xi_{jm})_{j\le J}, \mathcal{X}_m)$ is independent of $(\mathcal{X}_n)_{n\ne m}$. So $\mathrm{E}[\xi_{jm} \mid \mathcal{X}] = \mathrm{E}[\xi_{jm} \mid \mathcal{X}_m] = 0$ by (C2)(b), and $\mathrm{E}[\xi_{jm}\xi_{im} \mid \mathcal{X}] = \mathrm{E}[\xi_{jm}\xi_{im} \mid \mathcal{X}_m]$. Across markets, for $n \ne m$, $\mathrm{E}[\xi_{jm}\xi_{in} \mid \mathcal{X}] = \mathrm{E}[\xi_{jm} \mid \mathcal{X}]\ \mathrm{E}[\xi_{in} \mid \mathcal{X}] = 0$ by (C2)(a).

(ii) Let $U := \frac{1}{JM}\sum_{j=1}^{J}\sum_{m=1}^{M} A_{jm}\xi_{jm}$ and write $\langle\cdot,\cdot\rangle$ for the trace inner product. Since the A_{jm} are $\mathcal{X}$-measurable, (i) removes the cross-market terms:

$$\mathrm{E}\left[\|U\|^2 \mid \mathcal{X}\right] = \frac{1}{(JM)^2}\sum_{m=1}^{M}\sum_{j=1}^{J}\sum_{i=1}^{J}\langle A_{jm}, A_{im}\rangle\ \mathrm{E}[\xi_{jm}\xi_{im} \mid \mathcal{X}_m].$$

Take expectations. Let $C_A := \sup \mathrm{E}\,\|A_{jm}\|^4$. Hölder's inequality with exponents 4, 4 and 2 gives $\big|\,\mathrm{E}\left[\langle A_{jm}, A_{im}\rangle\ \mathrm{E}[\xi_{jm}\xi_{im} \mid \mathcal{X}_m]\right]\big| \le C_A^{1/2}\big\|\,\mathrm{E}[\xi_{jm}\xi_{im} \mid \mathcal{X}_m]\big\|_2$. By (C2)(b),

$$\mathrm{E}\,\|U\|^2 \le \frac{C_A^{1/2}}{(JM)^2}\sum_{m=1}^{M}\sum_{j=1}^{J}\sum_{i=1}^{J}\big\|\,\mathrm{E}[\xi_{jm}\xi_{im} \mid \mathcal{X}_m]\big\|_2 \le \frac{C_A^{1/2}L_\xi}{JM}.$$

Chebyshev gives $U = O_p((JM)^{-1/2})$.

(iii) Write $(JM)^{-1/2}\sum_{j=1}^{J}\sum_{m=1}^{M} A_{jm}\xi_{jm} = M^{-1/2}\sum_m T_m$ with $T_m := J^{-1/2}\sum_j A_{jm}\xi_{jm}$. T_m is a function of market m only, so $T_1, \dots, T_M$ are independent by (C2)(a). $\mathrm{E}\, T_m = 0$ by (i), and $M^{-1}\sum_m \mathrm{E}\, T_m T_m' = \Omega_{JM}^A$ by the definition of Ω_{JM}^A. For Lyapunov, fix a unit vector λ and let $a_{jm} := \lambda' A_{jm}$, which is $\mathcal{X}_m$-measurable. Then

$$\mathrm{E}(\lambda' T_m)^4 = \frac{1}{J^2}\sum_{j,i,l,n=1}^{J} \mathrm{E}\left[a_{jm}a_{im}a_{lm}a_{nm}\ \mathrm{E}[\xi_{jm}\xi_{im}\xi_{lm}\xi_{nm} \mid \mathcal{X}_m]\right].$$

Let $C_8 := \sup \mathrm{E}\,\|A_{jm}\|^8$. By Cauchy–Schwarz and Hölder's inequality with four factors in L^8, each term is at most $C_8^{1/2}\big\|\,\mathrm{E}[\xi_{jm}\xi_{im}\xi_{lm}\xi_{nm} \mid \mathcal{X}_m]\big\|_2$. By (C2)(b), $\mathrm{E}(\lambda' T_m)^4 \le C_8^{1/2} L_\xi$. Hence $M^{-2}\sum_m \mathrm{E}(\lambda' T_m)^4 \le C_8^{1/2} L_\xi / M \to 0$. This is the Lyapunov condition for $M^{-1/2}\sum_m \lambda' T_m$, whose variance $\lambda'\Omega_{JM}^A\lambda$ is bounded away from zero. The Lindeberg–Feller theorem and the Cramér–Wold device give the claim. □

The moment function at the truth. By (53) and (51),

$$\begin{aligned}\hat\theta &= \theta_0 + \left(\nabla\hat G'\hat W\nabla\hat G\right)^{-1}\nabla\hat G'\hat W\,\bar g(\theta_0),\\ \bar g(\theta_0) &= \frac{1}{JM}\sum_{j=1}^{J}\sum_{m=1}^{M}\nabla\tilde z_{jm}\nabla\xi_{jm} = \frac{1}{JM}\sum_{j=1}^{J}\sum_{m=1}^{M}\nabla\tilde z_{jm}\xi_{jm}.\end{aligned} \tag{57}$$

Lemma 12 (Moment at the truth: rate and CLT)**.** *Under (C2)(a),(b),(c):*

(i) $\bar g(\theta_0) = O_p\big((JM)^{-1/2}\big)$;

(ii) if $\liminf_{J,M}\lambda_{\min}(\Omega_{JM}) > 0$, *with* Ω_{JM} *as in* (54), *then* $\Omega_{JM}^{-1/2}\sqrt{JM}\,\bar g(\theta_0) \to_d N(0, I)$.

Proof. Step 1: (i). Use the second expression in (57). $\nabla\tilde z_{jm}$ is $\mathcal{X}$-measurable and has four moments by (C2)(c), so Lemma 11(ii) applies.

Step 2: (ii), the demeaning remainder. Write $u_j := \bar{\tilde z}_j - \mathrm{E}\,\bar{\tilde z}_j$, so that $\nabla\tilde z_{jm} = (\tilde z_{jm} - \mathrm{E}\,\bar{\tilde z}_j) - u_j$. By (57),

$$\sqrt{JM}\,\bar g(\theta_0) = \frac{1}{\sqrt{M}}\sum_m T_m - \sqrt{\frac{M}{J}}\sum_j u_j\,\bar\xi_j, \qquad T_m := \frac{1}{\sqrt{J}}\sum_j(\tilde z_{jm} - \mathrm{E}\,\bar{\tilde z}_j)\,\xi_{jm}.$$

For all j and i, $\mathrm{E}[u_j' u_i\,\bar\xi_j\bar\xi_i] = \mathrm{E}\left[u_j' u_i\ \mathrm{E}[\bar\xi_j\bar\xi_i \mid \mathcal{X}]\right]$, and $\mathrm{E}[\bar\xi_j\bar\xi_i \mid \mathcal{X}] = M^{-2}\sum_m \mathrm{E}[\xi_{jm}\xi_{im} \mid \mathcal{X}_m]$

by Lemma 11(i). u_j is a mean of M independent mean-zero terms, (C2)(a), with fourth moments, (C2)(c), so Rosenthal's inequality (Ros) gives $\mathrm{E}\,\|u_j\|^4 \le CM^{-2}$. Hölder's inequality with exponents 4, 4 and 2 gives

$$\left|\,\mathrm{E}[u_j' u_i\,\bar{\xi}_j\bar{\xi}_i]\right| \le \frac{C}{M}\cdot\frac{1}{M^2}\sum_{m=1}^{M}\left\|\,\mathrm{E}[\xi_{jm}\xi_{im}\mid \mathcal{X}_m]\right\|_2.$$

Sum over (j,i) and apply (C2)(b). Therefore

$$\mathrm{E}\left\|\sqrt{\frac{M}{J}}\sum_j u_j\bar{\xi}_j\right\|^2 = \frac{M}{J}\sum_{j=1}^{J}\sum_{i=1}^{J}\mathrm{E}\left[u_j' u_i\,\bar{\xi}_j\bar{\xi}_i\right] \le \frac{M}{J}\cdot\frac{C}{M^3}\cdot MJL_\xi = \frac{CL_\xi}{M}\to 0.$$

Step 3: (ii), the CLT. $A_{jm} := \tilde{z}_{jm} - \mathrm{E}\,\bar{\tilde{z}}_j$ is $\mathcal{X}_m$-measurable, has eight moments by (C2)(c) and (C7^D), and gives $\Omega^A_{JM} = \Omega_{JM}$. Lemma 11(iii) gives $\Omega_{JM}^{-1/2}M^{-1/2}\sum_m T_m \to_d N(0,I)$. Combine with Step 2, using that $\Omega_{JM}^{-1/2}$ is bounded. □

Decomposition of the regressors. By (55) and $\log S_{km} = \log\mu_{km} + r_{km}$, with $k = k(j)$ and $E_K := [\,0_{K\times(1+d_x)}\;\; I_K\,]$, so that $E_K' e_k = (0,0,e_k')'$,

$$R_{jm} = F_{jm} + R^r_{jm} - \log J\; E_K' e_k, \tag{58}$$

with F_{jm} as defined above and

$$R^r_{jm} := -r_{km}\,E_K' e_k.$$

The term $\log J\,E_K' e_k$ does not vary across markets, so demeaning removes it:

$$\nabla R_{jm} = \nabla F_{jm} + \nabla R^r_{jm}, \qquad \nabla\hat{G} = \nabla\hat{G}^\mu + \nabla\hat{G}^r, \tag{59}$$

with $\nabla\hat{G}^\bullet := (JM)^{-1}\sum_{j=1}^J\sum_{m=1}^M \nabla\tilde{z}_{jm}\,\nabla R^{\bullet\prime}_{jm}$.

Lemma 13 (Jacobian). *Under (C1)–(C3),* $\sup_{J,M}\|\Gamma_{JM}\| < \infty$ *and* $\nabla\hat{G} - \Gamma_{JM} \to_p 0$.

Proof. Throughout, $\nabla\tilde{z}_{jm}$ is $\mathcal{X}$-measurable and $\sup\mathrm{E}\,\|\nabla\tilde{z}_{jm}\|^4 < \infty$ by (C2)(c).

The r_{km} component. $\nabla R^r_{jm} = -\nabla r_{k(j)m}\,E_K' e_{k(j)}$, so

$$\|\nabla\hat{G}^r\| \le \max_{k\le K,\,m\le M}|\nabla r_{km}|\cdot\frac{1}{JM}\sum_{j=1}^J\sum_{m=1}^M\|\nabla\tilde{z}_{jm}\|.$$

The first factor is at most $2\max_{k,m\le M}|r_{km}| = o_p(1)$ by (56), which uses Lemma 6(ii), (C1) and (C3). The second has expectation bounded by $\sup \mathrm{E}\,\|\nabla\tilde{z}_{jm}\| < \infty$, so it is $O_p(1)$ by Markov. Hence $\nabla\hat{G}^r = o_p(1)$.

The μ component. Let $a_{jm} := \tilde{z}_{jm}$ and $b_{jm} := F_{jm}$, so that $\nabla F_{jm} = \nabla b_{jm}$ and $\mathrm{E}\,\nabla\hat{G}^\mu = \Gamma_{JM}$ by (54).

Step 1 (moments). $\sup \mathrm{E}\,\|a_{jm}\|^4 < \infty$ by (C2)(c). b_{jm} is linear in $(p_{jm}, \mathbf{x}_{jm}, \xi_{jm})$ with coefficients bounded by (C2)(e), plus $\log\mu_{km}$ and $\log(J_k/J)$, which are bounded by Lemma 6(i) and (C1). So $\sup \mathrm{E}\,\|b_{jm}\|^4 < \infty$ by (C2)(c). By (29),

$$\nabla\hat{G}^\mu = \frac{1}{M}\sum_m Q_m - \frac{1}{J}\sum_j \bar{a}_j\bar{b}_j', \qquad Q_m := \frac{1}{J}\sum_j a_{jm}b_{jm}'.$$

By Jensen and Cauchy–Schwarz, $\mathrm{E}\,\|Q_m\|^2 \le J^{-1}\sum_j(\mathrm{E}\,\|a_{jm}\|^4)^{1/2}(\mathrm{E}\,\|b_{jm}\|^4)^{1/2} \le C$, with C free of J and M. Likewise $\mathrm{E}\,\|\bar{a}_j\bar{b}_j'\| \le C$. Hence $\mathrm{E}\,\|\nabla\hat{G}^\mu\| \le C$, and $\|\Gamma_{JM}\| \le C$.

Step 2 (first term). Q_m is a function of market m only, so the Q_m are independent across m by (C2)(a). Hence $\mathrm{E}\,\|M^{-1}\sum_m(Q_m - \mathrm{E}\,Q_m)\|^2 = M^{-2}\sum_m \mathrm{E}\,\|Q_m - \mathrm{E}\,Q_m\|^2 \le C/M$.

Step 3 (second term). Write $\bar{a}_j = \mathrm{E}\,\bar{a}_j + u_j$ and $\bar{b}_j = \mathrm{E}\,\bar{b}_j + v_j$, so $\|\mathrm{E}\,\bar{a}_j\|, \|\mathrm{E}\,\bar{b}_j\| \le C$. By independence across markets, $\mathrm{E}\,\|u_j\|^2 = M^{-2}\sum_m \mathrm{E}\,\|a_{jm} - \mathrm{E}\,a_{jm}\|^2 \le C/M$, and likewise $\mathrm{E}\,\|v_j\|^2 \le C/M$. Then

$$\bar{a}_j\bar{b}_j' - \mathrm{E}\;\bar{a}_j\bar{b}_j' = (\mathrm{E}\,\bar{a}_j)v_j' + u_j(\mathrm{E}\,\bar{b}_j)' + \big(u_jv_j' - \mathrm{E}\,u_jv_j'\big).$$

Averaging over j and using the triangle inequality and Cauchy–Schwarz, the L^1 norms of the three averages are at most $CM^{-1/2}$, $CM^{-1/2}$, and $2C/M$.

Step 4 (conclude). By Steps 2 and 3, $\mathrm{E}\,\|\nabla\hat{G}^\mu - \Gamma_{JM}\| \le CM^{-1/2} \to 0$ by (C3), and Markov gives $\nabla\hat{G}^\mu - \Gamma_{JM} \to_p 0$. Nothing about the dependence of (p, z, h) across products within a market is used. The law of large numbers runs over independent markets. The within-product demeaning only adds the second term, which is small because product means over M independent markets concentrate.

Summing the two components and using (59) gives $\nabla\hat{G} - \Gamma_{JM} \to_p 0$. □

No rate for $\nabla\hat{G}$ is needed: the $\sqrt{JM}$ rate of $\hat{\theta} - \theta_0$ comes from $\bar{g}(\theta_0)$ alone, by Slutsky.

Proof of Theorem B.4. Write $\hat{\theta}$ for the oracle estimator $\hat{\theta}(c(k))$.

Proof of Theorem B.4. Step 1: (i). By Lemma 13 and (C7^D), $\nabla\hat{G}'\hat{W}\nabla\hat{G} - \Gamma'_{JM}W_{JM}\Gamma_{JM} \to_p 0$. By (C7^D) and $\lambda_{\min}(W_{JM}) \geq c_W$, $\lambda_{\min}(\Gamma'_{JM}W_{JM}\Gamma_{JM}) \geq c_W c_\Gamma > 0$ for large J, M, so $\nabla\hat{G}'\hat{W}\nabla\hat{G}$ is invertible with probability tending to one and

$$(\nabla\hat{G}'\hat{W}\nabla\hat{G})^{-1}\nabla\hat{G}'\hat{W} = (\Gamma'_{JM}W_{JM}\Gamma_{JM})^{-1}\Gamma'_{JM}W_{JM} + o_p(1),$$

with both sides $O_p(1)$. Insert in (53) and use Lemma 12(i).
Step 2: (ii). Let $B_{JM} := (\Gamma'_{JM}W_{JM}\Gamma_{JM})^{-1}\Gamma'_{JM}W_{JM}$, which is bounded by (C7^D) and Lemma 13. By (i), $\sqrt{JM}(\hat{\theta} - \theta_0) = B_{JM}\sqrt{JM}\,\bar{g}(\theta_0) + o_p(1)$. Let $Z_{JM} := \Omega_{JM}^{-1/2}\sqrt{JM}\,\bar{g}(\theta_0)$, so that $Z_{JM} \to_d N(0, I)$ by Lemma 12(ii). Then

$$V_{JM}^{-1/2}\sqrt{JM}(\hat{\theta} - \theta_0) = H_{JM}Z_{JM} + o_p(1), \qquad H_{JM} := V_{JM}^{-1/2}B_{JM}\Omega_{JM}^{1/2}.$$

The $o_p(1)$ term stays $o_p(1)$ because $\lambda_{\min}(V_{JM})$ is bounded away from zero, by (C7^D) and Lemma 13. $H_{JM}H'_{JM} = I$ by the definition of V_{JM}, so H_{JM} is bounded. Along any subsequence, H_{JM} has a further subsequence converging to some H with $HH' = I$, and along it $H_{JM}Z_{JM} \to_d N(0, HH') = N(0, I)$. The limit does not depend on the subsequence, so $H_{JM}Z_{JM} \to_d N(0, I)$. □

Only $\mathrm{E}[\xi_{jm} \mid \mathcal{X}_m] = 0$ and the covariance bound in (C2)(b) enter (i). The fourth-moment bound in (C2)(b) enters (ii) once, for the Lyapunov bound in Lemma 11(iii).

B.7 The second step in levels

This section proves Theorem 5.2 and Corollary 1. Berry (1994) uses the Berry inversion in levels, (4), as the basis of an IV estimator with J fixed and M large. This section applies the same GMM estimator to (4) with $J \to \infty$, without demeaning. The regressor $\log \mathbb{P}_{j|k,m}$ then contains the deterministic level $-\log J_k = -\log J - \log \pi_{kJ}$, by (55), and the Jacobian has entries of order $\log J$; the nest-specific part $\log \pi_{kJ}$ is bounded by (C1). A rescaling of the parameter by $\log J$ in the nest directions restores a Jacobian of order one; the nest parameters are then estimated at rate $\log J\sqrt{JM}$ and β at rate $\sqrt{JM}$.

Estimator. Write $y^0_{jm} := \log \mathbb{P}_{jm} - \log \mathbb{P}_{0m}$. The regressor R_{jm} of Section 4.2 is in levels, and (4) reads $y^0_{jm} = R'_{jm}\theta_0 + \xi_{jm}$. With the same instruments $\tilde{z}_{jm}$ and weight matrix $\hat{W}$,

$$\hat{\theta}^L := \left(\hat{G}'\hat{W}\hat{G}\right)^{-1}\hat{G}'\hat{W}\,\frac{1}{JM}\sum_{j=1}^{J}\sum_{m=1}^{M}\tilde{z}_{jm}y^0_{jm} = \theta_0 + \left(\hat{G}'\hat{W}\hat{G}\right)^{-1}\hat{G}'\hat{W}\,\bar{g}^L(\theta_0),$$

$$\hat{G} := \frac{1}{JM}\sum_{j=1}^{J}\sum_{m=1}^{M}\tilde{z}_{jm}R'_{jm}, \qquad \bar{g}^L(\theta_0) := \frac{1}{JM}\sum_{j=1}^{J}\sum_{m=1}^{M}\tilde{z}_{jm}\,\xi_{jm}.$$

So $\hat{\theta}^L$ is the oracle estimator $\hat{\theta}(k)$ of (12), and $\bar{g}^L(\theta_0)$ is $\bar{g}(\theta_0)$ of (11).

Decomposition and scaling. By (58), $R_{jm} = F_{jm} + R^r_{jm} - \log J\, E'_K e_{k(j)}$, with F_{jm} as in (C7). Without demeaning, the $\log J$ term stays in the regressor. Define the scaling matrix, the nest means of the instruments, and the scaled population Jacobian:

$$D_J := \begin{pmatrix} I_{1+d_x} & 0 \\ 0 & (\log J)^{-1} I_K \end{pmatrix}, \qquad \Pi_{JM} := \frac{1}{JM}\sum_{j=1}^{J}\sum_{m=1}^{M}\mathrm{E}\left[\tilde{z}_{jm}e'_{k(j)}\right],$$

$$\Gamma^L_{JM} := \frac{1}{JM}\sum_{j=1}^{J}\sum_{m=1}^{M}\mathrm{E}\left[\tilde{z}_{jm}F'_{jm}\right]D_J - \Pi_{JM}E_K.$$

Since $\log J\, E_K D_J = E_K$, the $\log J$ term becomes $-\tilde{z}_{jm}e'_{k(j)}E_K$ after scaling.

The nest columns of Γ^L_{JM} are $-\Pi_{JM} + O(1/\log J)$. So, for large J, the nest means of the instruments identify the nest coefficients. They play the role of nest constants. This is the sense in which the level $-\log J$ of the regressor identifies σ_k. It requires the model to have no free nest intercepts.

Lemma 14 (Jacobian, levels). *Under (C1)–(C3),* $\sup_{J,M}\|\Gamma^L_{JM}\| < \infty$ *and* $\hat{G}D_J - \Gamma^L_{JM} \to_p 0$.

Proof. Let $\hat{G}^r := (JM)^{-1}\sum_{j=1}^{J}\sum_{m=1}^{M}\tilde{z}_{jm}R^{r\prime}_{jm}$. Then $\hat{G}^r \to_p 0$ by the argument for the r_{km} component in the proof of Lemma 13, with $\tilde{z}_{jm}$ in place of $\nabla\tilde{z}_{jm}$. By the decomposition above and $\log J\, E_K D_J = E_K$,

$$\hat{G}D_J = \frac{1}{M}\sum_{m}Q^L_m + \hat{G}^r D_J, \qquad Q^L_m := \frac{1}{J}\sum_{j}\tilde{z}_{jm}\left(F'_{jm}D_J - e'_{k(j)}E_K\right).$$

Q^L_m is a function of market m only, so the Q^L_m are independent across m by (C2)(a). By

definition, $M^{-1}\sum_m \mathrm{E}\, Q_m^L = \Gamma_{JM}^L$. As in Step 1 of the μ component in the proof of Lemma 13, $\mathrm{E}\,\|Q_m^L\|^2 \le C$, since $\|D_J\| \le 1$ and the term $\log(J_k/J)$ in F_{jm} is bounded by (C1). So $\|\Gamma_{JM}^L\| \le C$ and $\mathrm{E}\,\|M^{-1}\sum_m (Q_m^L - \mathrm{E}\, Q_m^L)\|^2 \le C/M \to 0$ by (C3). With $\hat{G}^r D_J \to_p 0$, this gives the claim. □

Proof of Theorem 5.2. Condition (C7), Γ_{JM}^L and Ω_{JM}^L are defined in Section 5.2, and V_{JM}^L in Theorem 5.2.

Proof. Step 1: (i). Since D_J is invertible, $(\hat{G}'\hat{W}\hat{G})^{-1}\hat{G}'\hat{W} = D_J\big(D_J'\hat{G}'\hat{W}\hat{G}D_J\big)^{-1}D_J'\hat{G}'\hat{W}$, so

$$D_J^{-1}(\hat{\theta}^L - \theta_0) = \big(D_J'\hat{G}'\hat{W}\hat{G}D_J\big)^{-1}D_J'\hat{G}'\hat{W}\,\bar{g}^L(\theta_0).$$

By Lemma 14 and (C7), $D_J'\hat{G}'\hat{W}\hat{G}D_J - \Gamma_{JM}^{L\prime}W_{JM}\Gamma_{JM}^L \to_p 0$ and $\lambda_{\min}(\Gamma_{JM}^{L\prime}W_{JM}\Gamma_{JM}^L) \ge c_W c_\Gamma > 0$ for large J, M. So the inverse exists with probability tending to one and

$$\big(D_J'\hat{G}'\hat{W}\hat{G}D_J\big)^{-1}D_J'\hat{G}'\hat{W} = \big(\Gamma_{JM}^{L\prime}W_{JM}\Gamma_{JM}^L\big)^{-1}\Gamma_{JM}^{L\prime}W_{JM} + o_p(1),$$

both sides $O_p(1)$. $\bar{g}^L(\theta_0) = O_p((JM)^{-1/2})$ by Lemma 11(ii) with $A_{jm} = \tilde{z}_{jm}$. The rates for $\hat{\beta}^L$ and $\hat{\sigma}_k^L$ read off the two blocks of D_J.

Step 2: (ii). By (i), $\sqrt{JM}D_J^{-1}(\hat{\theta}^L - \theta_0) = (\Gamma_{JM}^{L\prime}W_{JM}\Gamma_{JM}^L)^{-1}\Gamma_{JM}^{L\prime}W_{JM}\sqrt{JM}\,\bar{g}^L(\theta_0) + o_p(1)$, and V_{JM}^L is bounded with $\lambda_{\min}$ bounded away from zero by (C7) and Lemma 14. Lemma 11(iii) with $A_{jm} = \tilde{z}_{jm}$, which is $\mathcal{X}_m$-measurable and has eight moments by (C2)(c) and (C7), gives $(\Omega_{JM}^L)^{-1/2}\sqrt{JM}\,\bar{g}^L(\theta_0) \to_d N(0, I)$. There is no demeaning remainder. The rest is Step 2 of the proof of Theorem B.4, with Γ_{JM}^L and Ω_{JM}^L in place of Γ_{JM} and Ω_{JM}. □

Rates. All K nest parameters are estimated at $\log J\sqrt{JM}$. The regressor $\log \mathbb{P}_{j|k,m}e_k$ has the deterministic level $-\log J_k = -\log J + O(1)$ in every nest, and the outcome has the matching level $(\sigma_k - 1)\log J_k$. After demeaning, the own-product variation of order one is the only source of identification. Here the $\log J$ level dominates it.

Rank condition. For large J, the rank part of (C7) is a condition on a matrix with two blocks. Its first $1 + d_x$ columns are those of $(JM)^{-1}\sum_{j=1}^J\sum_{m=1}^M \mathrm{E}[\tilde{z}_{jm}F_{jm}']$. Its last K columns are $-\Pi_{JM}$, the nest means of the instruments. With $\tilde{z} = (x', z', he_k')'$, the he_k block of Π_{JM} is

diagonal with entries $\pi_{kJ}\ \mathrm{E}[h_{jm} \mid k(j) = k]$. It has full rank whenever the nest means of h are bounded away from zero. For the exogenous share projection $h_{jm} = \mathbf{x}_{jm} - \log(J_k^{-1} \sum_{i \in k} e^{\mathbf{x}_{im}})$ with Gaussian $\mathbf{x}$, this mean is $-\operatorname{var}(x)/2 + o(1)$.

Nest intercepts. The identification uses the absence of free nest intercepts. A nest intercept would absorb the nest columns $-\Pi_{JM} E_K$. The nest parameters would then be identified only from cross-market variation, as in the demeaned estimator.

Comparison with Theorem B.4. The levels estimator converges faster for σ_k: $\log J \sqrt{JM}$ against $\sqrt{JM}$. It requires no free nest intercepts, so it does not allow product fixed effects. The demeaned estimator allows them.

The argument of Corollary 1(i) applies to any statistic computed from the data and a partition, including the demeaned estimator. So the feasible demeaned estimator, computed with $\hat{k}$ from Theorem 5.1, has the asymptotic distribution of Theorem B.4.

Proof of Corollary 1.

Proof of Corollary 1. (i) Let E be the event that $\hat{k} = \tau \circ k$ for some permutation τ. By Theorem 5.1, $\mathbb{P}(E^c) \to 0$. On E, $\hat{k}$ and k split the products into the same groups. The rule c depends on a partition only through its groups, so $c(\hat{k}) = c(k)$ on E. Hence $\hat{\theta}(c(\hat{k})) = \hat{\theta}(c(k))$ on E, and $\mathbb{P}\big(\hat{\theta}(c(\hat{k})) \neq \hat{\theta}(c(k))\big) \leq \mathbb{P}(E^c) \to 0$. The same argument applies to any statistic computed from the data and a partition, since on E it is computed from the same groups.

(ii) The labels $c(k)$ of the true partition are non-random, so Theorem 5.2 holds for $\hat{\theta}(c(k))$, with the nest-specific components of θ_0, Γ_{JM}^L and V_{JM}^L labelled by $c(k)$. Fix $\varepsilon > 0$. The event $\{\|(V_{JM}^L)^{-1/2} \sqrt{JM} D_J^{-1} (\hat{\theta}(c(\hat{k})) - \hat{\theta}(c(k)))\| > \varepsilon\}$ is contained in E^c, so

$$(V_{JM}^L)^{-1/2} \sqrt{JM}\, D_J^{-1} \big(\hat{\theta}(c(\hat{k})) - \theta_0\big) = (V_{JM}^L)^{-1/2} \sqrt{JM}\, D_J^{-1} \big(\hat{\theta}(c(k)) - \theta_0\big) + o_p(1).$$

The first term converges to $N(0, I)$ by Theorem 5.2(ii), and Slutsky's lemma gives the claim. □

Only $\mathbb{P}(E^c) \to 0$ is used in this proof. The rate in Theorem B.1 and the per-product bound of Lemma 3 are not needed for the equivalence. They are what delivers $\mathbb{P}(E^c) \to 0$ when J grows with M.

C Appendix for the Empirical Applications

C.1 Classification Algorithms

The classification problems in (48), (10), and (9) require optimizing not only over regression coefficients but also over partitions of products. The number of possible partitions grows exponentially with the number of products, making direct optimization computationally demanding and, in some cases, infeasible. To address this issue, we extend the algorithm proposed by Bonhomme and Manresa (2015) to accommodate the different empirical models considered in Section 2.

Case 1: No price partition or identical partitions. We first consider the case in which the partition of prices coincides with the partition of products, $g(j) = k(j)$ for all j, or when there is effectively no price partition, i.e., $|\text{supp}(g)| = 1$. In this setting, we propose the following iterative algorithm to solve eq. (48):

1. Let $(\boldsymbol{\theta}^0, \boldsymbol{\alpha}^0) \equiv (\theta_1^0, \ldots, \theta_K^0, \alpha_{11}^0, \ldots, \alpha_{KM}^0)$ denote an initial set of parameter values.

2. Given $(\boldsymbol{\theta}^s, \boldsymbol{\alpha}^s)$, compute for all $j \in J$:

$$k(j)^{s+1} = \underset{k \in \{1,\ldots,K\}}{\arg\min} \sum_{m=1}^{M} \Big(\log \mathbb{P}_{jm} - (\theta_{k,z}^s z_{jm} + \theta_{k,x}^s \mathbf{x}_{jm} + \alpha_{km}^s) \Big)^2,$$

   which determines the updated grouping structure $\mathcal{K}^{s+1}$.

3. Given the updated grouping structure $\mathcal{K}^{s+1}$, compute:

$$(\boldsymbol{\theta}^{s+1}, \boldsymbol{\alpha}^{s+1}) = \underset{\theta_1,\ldots,\theta_K,\alpha_{11},\ldots,\alpha_{KM}}{\arg\min} \sum_{j=1}^{J} \sum_{m=1}^{M} \Big(\log \mathbb{P}_{jm} - (\theta_{k(j),z}^{s+1} z_{jm} + \theta_{k(j),x}^{s+1} \mathbf{x}_{jm} + \alpha_{k(j)m}^{s+1}) \Big)^2.$$

4. Repeat steps 2 and 3 until convergence of the parameters.

In practice, this problem is not globally convex, as is common in clustering problems. We therefore recommend running the algorithm from multiple starting values and selecting the solution that achieves the lowest objective value.

Case 2: Different partitions. We now consider the case in which $g(j) \neq k(j)$ for some j and $|\text{supp}(g)| > 1$. In this case, we proceed in two stages. First, we estimate the partition of prices.

1. Let $(\boldsymbol{\gamma}^0, \boldsymbol{\kappa}^0) \equiv (\gamma^{1,0}, \ldots, \gamma^{G,0}, \kappa_{11}^0, \ldots, \kappa_{GM}^0)$ denote an initial set of parameter values.

2. Given $(\boldsymbol{\gamma}^s, \boldsymbol{\kappa}^s)$, compute for all $j \in J$:

$$g(j)^{s+1} = \underset{g \in \{1,\ldots,G\}}{\arg\min} \sum_{m=1}^{M} \Big(p_{jm} - (\gamma_z^{g,s} z_{jm} + \gamma_x^{g,s} \mathbf{x}_{jm} + \kappa_{gm}^s) \Big)^2,$$

   which determines the updated grouping structure $\mathcal{G}^{s+1}$.

3. Given the updated grouping structure $\mathcal{G}^{s+1}$, compute

$$(\boldsymbol{\gamma}^{s+1}, \boldsymbol{\kappa}^{s+1}) = \underset{\gamma^1,\ldots,\gamma^G,\kappa_{11},\ldots,\kappa_{GM}}{\arg\min} \sum_{j=1}^{J} \sum_{m=1}^{M} \Big(p_{jm} - (\gamma_z^{g(j),s+1} z_{jm} + \gamma_x^{g(j),s+1} \mathbf{x}_{jm} + \kappa_{g(j)m}^{s+1}) \Big)^2.$$

4. Repeat until convergence of the parameters.

After obtaining the price partition and coefficients, we compute the predicted prices

$$\hat{p}_{jm} = \hat{\gamma}_z^{g(j)} z_{jm} + \hat{\gamma}_x^{g(j)} \mathbf{x}_{jm} + \hat{\kappa}_{g(j)m}.$$

We then estimate the grouping structure and coefficients of the demand equation using the following algorithm:

1. Let $(\boldsymbol{\beta}^0, \boldsymbol{\lambda}^0) \equiv (\beta_1^0, \ldots, \beta_K^0, \lambda_{11}^0, \ldots, \lambda_{KM}^0)$ denote an initial set of parameter values.

2. Given $(\boldsymbol{\beta}^s, \boldsymbol{\lambda}^s)$, compute for all $j \in J$:

$$k(j)^{s+1} = \underset{k \in \{1,\ldots,K\}}{\arg\min} \sum_{m=1}^{M} \Big(\log \mathbb{P}_{jm} - (\beta_{k,p}^s \hat{p}_{jm} + \beta_{k,x}^s \mathbf{x}_{jm} + \lambda_{km}^s) \Big)^2,$$

   which determines the updated grouping structure $\mathcal{K}^{s+1}$.

3. Given the updated grouping structure $\mathcal{K}^{s+1}$, compute

$$(\boldsymbol{\beta}^{s+1}, \boldsymbol{\lambda}^{s+1}) = \underset{\beta_1,\ldots,\beta_K,\lambda_{11},\ldots,\lambda_{KM}}{\arg\min} \sum_{j=1}^{J}\sum_{m=1}^{M} \left(\log \mathbb{P}_{jm} - (\beta^{s+1}_{k(j),p}\hat{p}_{jm} + \beta^{s+1}_{k(j),x}\mathbf{x}_{jm} + \lambda^{s+1}_{k(j)m}) \right)^2.$$

4. Repeat until convergence of the parameters.

As before, because these optimization problems are not globally convex, we recommend running the algorithm from multiple starting values and selecting the solution that attains the lowest objective value.

C.2 Robustness Checks on Group Classification

C.2.1 Results using Hausman instruments for prices

Table A2: Average group characteristics for $k^* = 4$ and Hausman instrument

Characteristics	Unconditional values	Conditional values by group			
		1	2	3	4
Unit size (ml)	354.882	354.882	473.176	354.882	354.882
Unit quantity	6.000	6.000	6.000	6.000	6.000
Unit price	10.209	10.295	9.590	10.474	9.604
Price per ml	0.004	0.004	0.003	0.004	0.004
Share of domestic	0.794	0.767	0.836	0.833	0.879
Share of ale	0.158	0.102	0.226	0.287	0.339
Share of regular beer	0.442	0.437	0.437	0.496	0.358
Share of stout and porter	0.015	0.013	0.016	0.020	0.027
Share of light beer	0.369	0.437	0.297	0.171	0.256
Share of other types	0.016	0.011	0.024	0.026	0.020
Share top-ten brands	0.624	0.748	0.459	0.365	0.357
# of products	2806	271	692	1119	724

Notes: Rows 1–4 report median values of unit size, unit quantity, unit price, and price per ml, presented both unconditionally and conditional on group membership. Rows 5–11 report unconditional and conditional mean shares. Groups are estimated using Hausman instruments for prices instead of the baseline Gandhi-Houde instruments.

Table A3: Estimation Results for Nested Logit with k^* = 4

	OLS	IV
log unit size	0.107 (0.006)***	0.431 (0.018)***
log upc quantity	0.121 (0.005)***	0.586 (0.020)***
log price	-0.084 (0.006)***	-0.581 (0.021)***
σ_1	0.330 (0.002)***	0.962 (0.022)***
σ_2	0.127 (0.001)***	0.662 (0.019)***
σ_3	-0.025 (0.001)***	0.469 (0.018)***
σ_4	-0.337 (0.001)***	0.178 (0.019)***
Mean Own-price Elasticity	1.193 (0.003)	-1.612 (0.004)
Mean Cross-price Elasticity	-0.003 (0.008)	0.004 (0.011)
Market Fixed Effects	✓	✓
IV Type		hausman
Number of Products	2806	
Number of Observations	64497	

Note: The table reports OLS and IV estimates of a nested logit demand model where products are grouped by beer type, yielding $k = 4$ nests. The dependent variable is $\log \mathbb{P}_{jm} - \log \mathbb{P}_{0m}$. Regressors include log unit size (ml), log UPC quantity, log price, and four group-specific nesting parameters (σ_g). The IV specification instruments for log price using Hausman IVs and for the within-group log share using the exogenous share projection. Market fixed effects are included. Mean own-price and cross-price elasticities are computed from the IV estimates and weighted by national unit sales. $^{***}p < 0.01, ^{**}p < 0.05, ^{*}p < 0.1$.

C.2.2 Results for $k^* = 3$ and $k^* = 5$

Table A4: Characteristics of beer across groups ($k = 3$)

Characteristics	Unconditional values	Conditional values by group 1	2	3
Unit size (ml)	354.882	354.882	354.882	354.882
Unit quantity	6.000	6.000	6.000	6.000
Unit price	10.209	10.180	10.232	11.138
Price per ml	0.004	0.004	0.004	0.004
Share of domestic	0.794	0.781	0.832	0.792
Share of ale	0.158	0.107	0.278	0.376
Share of regular beer	0.442	0.451	0.414	0.462
Share of stout and porter	0.015	0.012	0.018	0.039
Share of light beer	0.369	0.418	0.263	0.110
Share of other types	0.016	0.013	0.026	0.013
Share top-ten brands	0.624	0.725	0.372	0.332
# of products	2806	393	1119	1294

Notes: Rows 1–4 report median values of unit size, unit quantity, unit price, and price per ml, presented both unconditionally and conditional on group membership. Rows 5–11 report unconditional and conditional mean shares. Groups are estimated using Gandhi-Houde instruments.

Figure A1: Nested logit ($k^* = 3$)



Logos are trademarks of their respective owners

Notes: This figure shows the top ten most national sold beer brands at the national level for each group. The bubble size of each brand reflects the relative size of the market share in each particular group.

Table A5: Estimation Results for Nested Logit with $k = 3$

	OLS	IV
log unit size	0.059 (0.006)***	0.681 (0.040)***
log upc quantity	0.067 (0.004)***	0.824 (0.042)***
log price	-0.067 (0.005)***	-1.038 (0.056)***
σ_1	0.272 (0.001)***	0.968 (0.024)***
σ_2	0.047 (0.001)***	0.649 (0.022)***
σ_3	-0.204 (0.001)***	0.391 (0.022)***
Mean Own-price Elasticity	-0.457 (0.001)	-2.010 (0.002)
Mean Cross-price Elasticity	0.001 (0.003)	0.002 (0.006)
Market Fixed Effects	✓	✓
IV Type		$GH_{\text{units}}, GH_{\text{size}}$
Number of Products	2806	
Number of Observations	64497	

Note: The table reports OLS and IV estimates of a nested logit demand model where products are grouped by beer type (regular beer/lager, light beer, ale, stout/porter, and other), yielding $k = 3$ nests. The dependent variable is $\log \mathbb{P}_{jm} - \log \mathbb{P}_{0m}$. Regressors include log unit size (ml), log UPC quantity, log price, and five group-specific nesting parameters (σ_g). The IV specification instruments for log price using Gandhi-Houde differentiation IVs and for the within-group log share using the exogenous share projection. Market fixed effects are included. Mean own-price and cross-price elasticities are computed from the IV estimates and weighted by national unit sales. $^{***}p < 0.01,^{**}\, p < 0.05,^{*}\, p < 0.1$.

Table A6: Characteristics of beer across groups ($k = 5$)

Characteristics	Unconditional values	Conditional values by group				
		1	2	3	4	5
Unit size (ml)	354.882	473.176	354.882	354.882	354.882	354.882
Unit quantity	6.000	6.000	6.000	6.000	6.000	4.000
Unit price	10.209	8.831	10.194	11.934	10.862	10.150
Price per ml	0.004	0.003	0.004	0.004	0.004	0.005
Share of domestic	0.794	0.838	0.756	0.866	0.778	0.816
Share of ale	0.158	0.188	0.090	0.318	0.337	0.411
Share of regular beer	0.442	0.438	0.446	0.434	0.449	0.364
Share of stout and porter	0.015	0.017	0.012	0.016	0.029	0.061
Share of light beer	0.369	0.333	0.441	0.209	0.172	0.146
Share of other types	0.016	0.024	0.011	0.023	0.013	0.017
Share top-ten brands	0.624	0.537	0.758	0.337	0.378	0.227
# of products	2806	445	197	727	878	559

Notes: Rows 1–4 report median values of unit size, unit quantity, unit price, and price per ml, presented both unconditionally and conditional on group membership. Rows 5–11 report unconditional and conditional mean shares. Groups are estimated using Gandhi-Houde instruments.



Figure A2: Nested logit ($k^* = 5$)

Notes: This figure shows the top ten most national sold beer brands at the national level for each group. The bubble size of each brand reflects the relative size of the market share in each particular group.

Table A7: Estimation Results for Nested Logit with $k = 5$

	OLS	IV
log unit size	0.087 (0.007)***	0.196 (0.027)***
log upc quantity	0.103 (0.005)***	0.241 (0.027)***
log price	-0.037 (0.007)***	-0.084 (0.036)**
σ_1	0.202 (0.002)***	0.698 (0.016)***
σ_2	0.407 (0.002)***	1.000 (0.019)***
σ_3	0.060 (0.001)***	0.516 (0.015)***
σ_4	-0.096 (0.001)***	0.340 (0.015)***
σ_5	-0.442 (0.002)***	0.029 (0.018)
Mean Own-price Elasticity	-0.057 (0.001)	-0.729 (0.004)
Mean Cross-price Elasticity	0.000 (0.002)	0.003 (0.011)
Market Fixed Effects	✓	✓
IV Type		$GH_{\text{units}}, GH_{\text{size}}$
Number of Products	2806	
Number of Observations	64497	

Note: The table reports OLS and IV estimates of a nested logit demand model where products are grouped by beer type (regular beer/lager, light beer, ale, stout/porter, and other), yielding $k = 5$ nests. The dependent variable is $\log \mathbb{P}_{jm} - \log \mathbb{P}_{0m}$. Regressors include log unit size (ml), log UPC quantity, log price, and five group-specific nesting parameters (σ_g). The IV specification instruments for log price using Gandhi-Houde differentiation IVs and for the within-group log share using the exogenous share projection. Market fixed effects are included. Mean own-price and cross-price elasticities are computed from the IV estimates and weighted by national unit sales. $^{***}p < 0.01, ^{**}p < 0.05, ^{*}p < 0.1$.

C.2.3 Classification including Beer Type Indicators

Table A8: Characteristics of beer across groups $k^* = 3$

Characteristics	Unconditional values	Conditional values by group 1	2	3
Unit size (ml)	354.882	473.176	354.882	354.882
Unit quantity	6.000	6.000	6.000	6.000
Unit price	10.209	9.857	10.265	11.020
Price per ml	0.004	0.004	0.004	0.004
Share of domestic	0.794	0.799	0.793	0.777
Share of ale	0.158	0.227	0.126	0.336
Share of regular beer	0.442	0.448	0.437	0.506
Share of stout and porter	0.015	0.016	0.013	0.041
Share of light beer	0.369	0.283	0.411	0.102
Share of other types	0.016	0.026	0.013	0.016
Share top-ten brands	0.624	0.413	0.710	0.349
# of products	2806	1116	420	1270

Notes: This table shows the unconditional and conditional values of product characteristics across groups. Unit size, unit quantity, unit price, and price per ml (rows 1 to 4) are calculated as median. Rows 5 to 11 are unconditional and conditional (mean) shares.

Table A9: Estimation Results for Nested Logit with $k^* = 3$

	OLS	IV
log unit size	0.037 (0.006)***	0.088 (0.024)***
log upc quantity	0.040 (0.005)***	0.122 (0.024)***
log price	-0.040 (0.006)***	-0.146 (0.031)***
σ_1	0.046 (0.001)***	0.303 (0.009)***
σ_2	0.269 (0.001)***	0.578 (0.010)***
σ_3	-0.209 (0.001)***	0.031 (0.009)***
Mean Own-price Elasticity	-0.276 (0.001)	-2.343 (0.006)
Mean Cross-price Elasticity	0.000 (0.002)	0.005 (0.015)
Market Fixed Effects	✓	✓
IV Type		$GH_{\text{units}}, GH_{\text{size}}$
Number of Products	2806	
Number of Observations	64497	

Note: The table reports OLS and IV estimates of a nested logit demand model where products are grouped by beer type (regular beer/lager, light beer, ale, stout/porter, and other), yielding $k = 3$ nests. Importantly, these results extend for beer type indicators. The dependent variable is $\log \mathbb{P}_{jm} - \log \mathbb{P}_{0m}$. Regressors include log unit size (ml), log UPC quantity, log price, and five group-specific nesting parameters (σ_g). The IV specification instruments for log price using Gandhi-Houde differentiation IVs and for the within-group log share using the exogenous share projection. Market fixed effects are included. Mean own-price and cross-price elasticities are computed from the IV estimates and weighted by national unit sales. $^{***}p < 0.01, ^{**}p < 0.05, ^{*}p < 0.1$.

C.2.4 K-means clustering only on characteristics

Table A10: Characteristics of beer across groups $k^* = 4$

Characteristics	Unconditional values	Conditional values by group 1	2	3	4
Unit size (ml)	354.882	473.176	354.882	473.176	739.338
Unit quantity	6.000	6.000	12.000	4.000	1.000
Unit price	10.209	9.544	15.672	8.831	2.693
Price per ml	0.004	0.003	0.003	0.003	0.004
Share of domestic	0.794	0.913	0.748	0.863	0.741
Share of ale	0.158	0.182	0.113	0.279	0.124
Share of regular beer	0.442	0.449	0.395	0.443	0.524
Share of stout and porter	0.015	0.021	0.017	0.015	0.004
Share of light beer	0.369	0.316	0.469	0.246	0.324
Share of other types	0.016	0.032	0.006	0.017	0.024
Share top-ten brands	0.624	0.449	0.726	0.435	0.712
# of products	2806	740	463	1424	179

Notes: This table shows the unconditional and conditional values of product characteristics across groups. Unit size, unit quantity, unit price, and price per ml (rows 1 to 4) are calculated as median. Rows 5 to 11 are unconditional and conditional (mean) shares.

Table A11: Estimation Results for Nested Logit with $k^* = 4$

	OLS	IV
log unit size	-0.019 (0.004)***	0.311 (0.018)***
log upc quantity	0.032 (0.003)***	0.335 (0.017)***
log price	0.004 (0.004)	-0.423 (0.024)***
σ_1	-0.044 (0.001)***	0.113 (0.005)***
σ_2	0.146 (0.001)***	0.330 (0.006)***
σ_3	-0.079 (0.001)***	0.066 (0.005)***
σ_4	0.047 (0.001)***	0.220 (0.006)***
Mean Own-price Elasticity	-0.036 (0.000)	-4.566 (0.010)
Mean Cross-price Elasticity	0.000 (0.000)	0.011 (0.026)
Market Fixed Effects	✓	✓
IV Type		$GH_{\text{units}}, GH_{\text{size}}$
Number of Products	2806	
Number of Observations	64497	

Notes: The table reports OLS and IV estimates of the nested logit demand model with $k^* = 4$ groups, where the first-step classification uses standard k-means clustering on mean observed product characteristics (unit size, quantity, price). The number of groups $k^* = 4$ is selected via the elbow method applied to the within-cluster sum of squares, using multiple random initializations. The dependent variable is $\log \mathbb{P}_{jm} - \log \mathbb{P}_{0m}$. Regressors include log unit size (ml), log UPC quantity, log price, and four group-specific nesting parameters $(\sigma_1, \ldots, \sigma_4)$. The IV specification instruments for log price using Gandhi-Houde differentiation IVs and for $\log \mathbb{P}_{j|g(j),m}$ using the exogenous share projection. Market fixed effects are included. Standard errors are reported in parentheses. Mean own-price and cross-price elasticities are computed from the IV estimates, weighted by total unit state sales. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.