acteristics demand model.[17] In principle, one would think that this would eliminate the concerns raised by Petrin (2002) about ε_{ijt} driving the welfare results. Instead, in their Monte Carlo results they find that using the pure characteristics model matters for the estimated elasticities (and mean utilities) but not the welfare numbers. They conclude that, consistent with the results in Nevo (2003, 2011), "the fact that the contraction fits the shares exactly means that the extra gain from the logit errors is offset by lower δ's, and this roughly counteracts the problems generated for welfare measurement by the model with tastes for products." Ackerberg and Rysman (2005) propose to use historical data to estimate how δ should change as a function of the number of products, in order to capture the change needed to rationalize the post-introduction data (if it were observed).

4 Demand Estimation

We have seen above the important role played by heterogeneity in consumer preferences for generating realistic patterns of product substitution and price elasticities. This flexibility comes at a cost: the more general model is more difficult to estimate. In this section we discuss the empirical problem of identifying and estimating the parameters of the market demand system with heterogeneous consumer preferences.

4.1 The Estimation Problem

The parameters to be estimated are defined in equation (3.7). They include $(\alpha_0,\ \beta_0)$ which are the parameters in the mean utility defined by equation (3.6). For reasons that will become clear below, we will refer to these as the "linear parameters." Next, are the coefficients of the demographic variables in equations (3.5) and (3.4) captured by the matrix Γ of dimension $(K+1) \times L$. Finally, the matrix Σ, a $(K+1) \times (K+1)$ diagonal matrix with the diagonal equal to $(\alpha_v, \beta_v^{(1)}, ..., \beta_v^{(K)})$, captures the idiosyncratic "taste for characteristics." Jointly, Γ and Σ will often be called the "non-linear parameters." The full parameter vector to be estimated is $\theta = (\alpha_0, \beta_0, \Gamma, \Sigma)$.

In estimation it is typical to treat the distribution of the idiosyncratic "taste for characteristics" $\nu_{it} = (\nu_{it}^{(0)}, \ldots, \nu_{it}^{(K)})$ as independent of the distribution of demographics D_{it} , In

step, of predicting the counterfactual market by, say, observing the market shares post introduction, the Logit model can get the correct answer in that example.

[17] As we explained above, ε_{ijt} help rationalize observed choices. Indeed, once we drop them the model can in principle have difficulty rationalizing certain patterns of behavior. See Athey and Imbens (2007) for a discussion of the potential problems with the pure characteristics model and an alternative model.

other words,

$$F\left(D_{it},\nu_{it}\right) = F_D(D_{it})F_\nu(\nu_{it}).$$

A further restriction often used in practice is to treat each $\nu_{it}^{(k)}$ as independent across $k = 0,\ldots,K$ and distributed standard normal. This is a strong assumption and not necessary for identification and estimation, but is usually assumed in applied work. At the end of this section we discuss some papers that relax this assumption.

The data used to estimate θ will generally have three types of variables. First, quantities of the J products purchased in market t, which are an aggregation of choices made by individual consumers.[18] A market t is implicitly defined by a set of consumers facing the same prices, $\boldsymbol{p}_t$, characteristics, $\boldsymbol{x}_t$ and demand shocks $\boldsymbol{\xi}_t$. Aggregate quantities can be converted to market shares by making an assumption about the market size I_t, namely the number of consumer who made choices, including the choice of the outside good. We will then define observed market shares as $s_{jt} = q_{jt}/I_t$. In many applications we think that I_t is sufficiently large so that we can ignore the sampling errors in these shares.[19]

Second, we will observe prices, p_{jt}, and ("observed") product characteristics, x_{jt}, of the J products in market t.[20] We do not observe all the characteristics consumers observe, but we assume that those that we do observe vary by market and/or product but are common to all consumers in a market.[21]

Third, the data may contain information on consumer demographics D_{ilt}. In micro data, the actual D_{ilt} will be observable. In other data sets the researcher may have access instead to the distribution of demographics, $F_t(D)$ (or have samples from it). At times the researcher will have data that is more aggregated than the consumer-level, but also less aggregated than the market level. For example, the average age of consumers who purchase product j, or the shares by income.

[18] In some of the illustrations below we will assume that we observe individual choices, which is often referred to as "micro data". Furthermore, some of the extensions discussed in Section 6 use micro data, but we do not offer a complete treatment of estimation with micro data.

[19] At the end of this section, we will revisit the role of sampling error in market shares.

[20] In some cases the number of products will vary across markets t. For simplicity of notation we focus on $J_t = J$.

[21] In rare cases, one might have characteristics data that varies by consumer. For example, different consumers in the same market might pay different prices when transaction-level data are available. A big issue with such data is that the price paid by a consumer for a purchased product might be observed but data for products not purchased will rarely be observed and need to be imputed.

4.2 What Variation in the Data Can Identify the Parameters?

In the subsection we provide intuition for what variation in the data allows us to identify the parameters θ.[22] Our goal is to use the logic of identification to motivate different estimation strategies.

We start by assuming that we have micro data, on individual choices, from a single market and we shut down part of the model by setting $\Sigma = 0$. We show that in this case there is an intuitive two-step procedure for estimating the remaining parameters. This procedure cannot be exactly replicated with market-level data but it provides us a road map for estimation in the more general case. It also provides direction on the type of variation that can identify Γ.

Next, we reintroduce the the random taste shocks, ν_{it} into the model (i.e., we do not restrict Σ to be zero). We show that in order to identify Σ we need different variation than what we use to identify Γ. To illustrate this point we assume we have market-level data from a single market and show the identifying power of moment conditions interacting ξ with IVs.

These simple cases, with data from a single market, are used for illustrative purposes. We use the insight from each of the cases to develop a set of practices that allow the full variation in the data to be leveraged to inform the estimation of θ. In the next section we will bring together these pieces in a general estimation framework that encompasses many of the procedures used in the literature.

4.2.1 Intuition from Individual-level Data

To gain intuition for the more general estimation procedure we start with a simple version of the model and assume that we have individual-level data from a single model. We have two goals in this discussion. First, we propose a simple way to estimate the parameters that will serve as a road map for the more general estimation problem. Second, the discussion will allow us to examine the sources of variation needed to estimate the different parameters.

Assume that we have data $\{y_{ij}, D_i\}_{i=1,...I}$, where $y_{ij} = 1$, for $j = 0, 1, ...J$ if consumer i chooses product j and $\sum_{j=0}^{J} y_{ij} = 1$. All the consumers are from a single market, in the sense that they face the same prices and product characteristics, both observed, $\boldsymbol{x}$ and unobserved $\boldsymbol{\xi}$. It might seem impossible to estimate demand in this setting. We only

[22] For a more formal treatment of identification see Berry and Haile (2021) in this Handbook.

observe a single snapshot of the market. How could we ever recover how quantities vary with changes in prices if prices do not vary? The answer relies on exploiting variation across households and across products to estimate the choice model, and then using the choice model to compute substitution as we saw in the previous section.

For exposition purposes we shut off certain parts of the model defined by (3.3). Specifically, we will assume that $\Sigma = 0$, namely that heterogeneity will only be driven by observed demographics. Also, we will assume, for this subsection, the price p_j is one of the observed characteristics x_j simply to ease the exposition. Given these assumptions the conditional indirect utility from product j (dropping the subscript t since we have a single market) is given by

$$u_{ij} = \underbrace{x_j\beta_0 + \xi_j}_{\delta_j} + \sum_{k,l} \beta_d^{(l,k)} D_{il} x_{jk} + \varepsilon_{ij}. \tag{4.1}$$

Note, that if we did not have ξ_j then we could estimate the parameters of the model, (β^0, Γ), by maximizing the likelihood of observing the choices in the sample as a function of $\boldsymbol{x}$ and D_i. The presence of ξ means that we need to modify the estimation somewhat. Specifically, we can estimate the parameters of the model in two steps. In the first step we include a product-specific intercept, that will capture δ, and absorb both $x_j\beta_0$ and ξ_j. In this step we estimate $\tilde{\theta} = (\delta_1, \ldots, \delta_J, \Gamma)$ using maximum likelihood. This allows us to "control" for the presence of ξ.

In the second step, we estimate β_0 by "projecting" the estimated $\hat{\delta}$'s on the x's. If we assume that $E(\xi_j|x_j) = 0$ we can use (weighted) least squares for this second stage. Alternatively, if we are concerned that a subset of the x's is correlated with ξ we can base the second stage on

$$E(\xi_j|Z_j) = 0, \tag{4.2}$$

where Z are a vector of exogenous variables, which we will discuss further below.

Let us examine the two-step procedure sketched above in further depth. If we parameterize the model according to $\tilde{\theta} = (\delta_1, \ldots, \delta_J, \Gamma)$, then it is fairy straightforward to show the first-order conditions implies the maximum likelihood estimates of the product-specific intercepts δ_j are found by setting the observed market shares, or average choice probabilities, equal to the ones predicted by the model. Namely, $\hat{s}_j = \hat{\sigma}(\hat{\delta}_1, \ldots, \hat{\delta}_J)$ for a fixed value of Γ. Under quite general conditions (Berry et al., 2013) this relation can be inverted to yield

$$\hat{\delta}_j = \hat{\sigma}_j^{-1} \left(\hat{s}_1, \ldots, \hat{s}_J \right). \tag{4.3}$$

As $I \to \infty$ the limit of this expression will be

$$\delta_j = \sigma_j^{-1} \left(s_1, \ldots, s_J \right) \quad j = 1, \ldots, J,$$

which will play a key role in the estimation with aggregate data, which we discuss below. It can be seen as the limit of (4.3), which comes from the first-order conditions of MLE.

Turning to the first-order conditions with respect to Γ, one can show that maximum likelihood estimates of Γ are the ones that equate the observed and predicted covariance between the demographic variables of those consumers that chose product j and the characteristics of the product. In the limit, Γ is identified as the solution to the system of the $L(K+1)$ equations

$$E_{Population} \left[x^k D^l \right] = E_{Model} \left[x^k D^l ; \Gamma \right]. \tag{4.4}$$

That is, Γ sets the model's prediction about the covariance between each demographic variable and the product characteristic of the chosen alternative equal to the population counterpart. The MLE moment conditions are simply sample analogues to these limiting moment conditions. As we will discuss below, the intuition gained from (4.4) can be useful even if the researcher does not have consumer choice data.

Suppose we go back to the more general model and allow for unobserved heterogeneity in tastes for characteristics at the individual level. That is, we are back to equation (3.3) where we have both sets of parameters Γ and Σ, and data from a single market. To estimate this model we could consider the same two-step approach as above. The first-order/moment conditions for (δ, Γ) derived above continue to hold. We also have first-order conditions with respect to Σ, which look almost identical to the conditions with respect to Γ. The difference is that it is not clear what is the counterpart of the data covariance that the model is matching, since ν is unobserved. To put it slightly differently, it is not clear what variation or moments in the data identifies these parameters.[23]

The answer for how to identify Σ when we observe data from a single market lies in using the variation in the second stage moments defined in equation (4.2). In the next subsection, we show how this moment helps identify Σ.

[23] If we observe multiple markets, we can use variation in the choice sets, if it exists, to identify Σ.

4.2.2 The Informational Content of $E[\xi \mid \boldsymbol{Z}] = 0$

In this subsection we show how the moment condition given in equation (4.2) provides identifying power for Σ. For exposition purposes, we simplify the model further and assume here that we do not have any demographic variables, D_i, and therefore the choice data is only used to compute aggregate market shares. We therefore can simply assume that we observe market-level data. We assume that we have data from one market and that the indirect utility is given by

$$u_{ij} = \delta_j + \sum_k \beta_\nu^{(k)} \nu_{ik} x_{jk} + \varepsilon_{ij} \tag{4.5}$$

The question becomes what variation in the data pins down the parameter vector (δ, Σ), where as previously defined Σ is a diagonal matrix with a diagonal equal to $\beta_\nu^{(1)}, ..., \beta_\nu^{(K)}$

As in the previous subsection, in order to get an expression of the aggregate choice probability that is a function of data and parameters, we can parameterize the δ's as product-specific intercepts and estimate them as parameters. As before, the estimated $\hat{\delta}$'s will set the predicted shares equal to the observed shares, for any given Σ. In other words, for every Σ there exists $\delta(\Sigma)$ that perfectly explains the observed market shares $\hat{s}$. There is no information remaining in the (aggregate) choice data alone that would distinguish one set of implied mean utilities $\delta(\tilde{\Sigma})$ from another assignment $\delta(\hat{\Sigma})$. The identification of the true δ requires using more of the structure of the model and adding an additional assumption. We recall that $\delta_j = x_j\beta_0 + \xi_j$.

Berry et al. (1995) propose adding the moment restriction, $E[\xi_j \mid \boldsymbol{Z}] = 0$. Below we will discuss variables that might satisfy this condition. This will require that estimates of β_0 and Σ not just fit the aggregate market shares as given by the MLE moments, but also fit the sample analog of this moment condition.

To gain intuition for how this works, let us work with a common exogeneity restriction that $\boldsymbol{Z} = \boldsymbol{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_J)$, i.e., $\boldsymbol{Z}$ is a stack of all the product characteristics in the market.[24] Our goal here is not to justify the assumption but rather to understand its empirical usefulness. We can view $\boldsymbol{x}$ as representing the market structure, namely, a configuration of the number of products and their product positions. The moment restriction (4.2) thus states that the unobserved component ξ_j of mean utility is mean-independent of market

[24] Observe that this same restriction on product characteristics being determined outside of the model was used in our discussion of Bresnahan (1987) in Section 2.

structure. In particular, the empirical bite of the assumption is that the ξ_j must be uncorrelated with the proximity of competition.

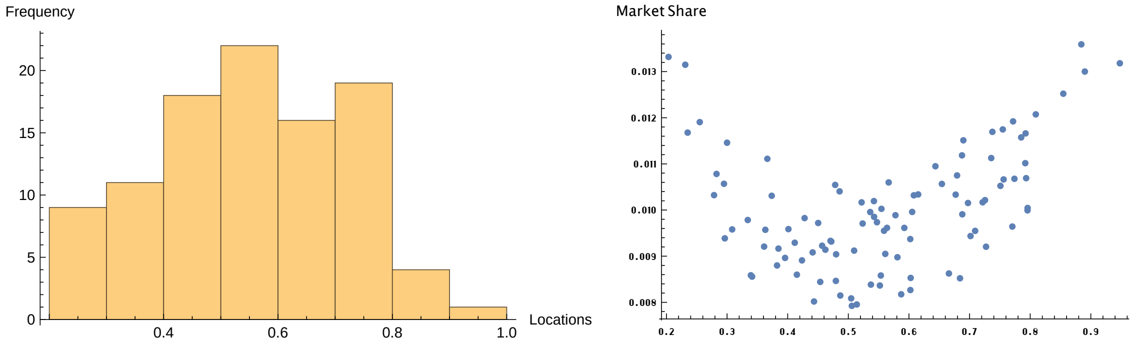
To see this point better, consider a one-dimensional (Hotelling-like) variation of our demand model. The utility to consumer i for product j in this model is

$$u_{ij} = \theta \cdot d\left(t_i, x_j\right) + \xi_j + \varepsilon_{ij} \quad j = 0, \ldots, J$$

where θ is the travel cost and d is the distance between the location $t_i \in [0,1]$ of consumer i and the location $x_j \in [0,1]$ of product j. ξ_j is the mean quality of product j in the population of consumers and ε_{ij} are i.i.d. idiosyncratic taste shocks around this mean quality drawn from a type-1 extreme value distribution.

If $\theta > 0$, i.e., travel costs are positive, a product j will draw demand in higher proportion from other products k that have characteristic x_k close to x_j. The larger the travel cost, the more this "local competition" effect will dominate the substitution patterns from the simple Logit component $\xi_j + \epsilon_{ij}$ of the model. Suppose we observe data $\{s_j, x_j\}_{j=1}^{J}$ and we want to infer the magnitude of travel costs θ from the data. To visualize the empirical problem, we plot in Figure 4.1 data generated by drawing 100 products where each product has a location $x_j \in [0,1]$ and quality $\xi_j \in \mathbb{R}$.[25]

Figure 4.1: Distribution of product locations and market shares


(a) Product locations

(b) Scatter plot of market shares

Notes: The data for these simulations was generated by drawing 100 products from the Hotelling-like model described in the text, where each product was drawn independently from a Beta distribution with both shape parameters equal to 4. In panel (a) we display the histogram of product locations on the line and in panel (b) we show the scatter plot of market shares and product locations in the resulting data.

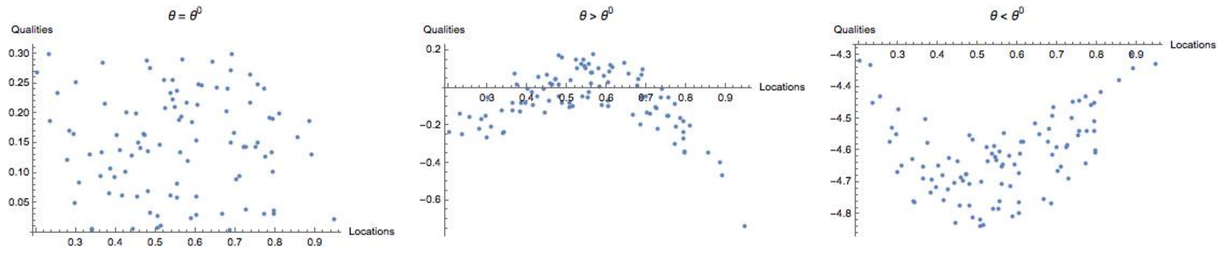
[25] Each product location was drawn independently from a Beta distribution with both shape parameters equal to 4.

In panel (a) we display the histogram of product locations on the line. The histogram shows that there is relatively more “bunching” of products in this market near the center of the line, and relative isolation of products near end points. We then compute market shares for each product $j = 1, \dots, 100$ based on a true $\theta^0 = 2$ where $\theta^0 \in \Theta$ denotes the true value of travel costs. In panel (b) we show the scatter plot of market shares and product locations in the resulting data. We can see a distinct pattern arise - in the more crowded part of product space (near the middle of the line in this simulation) market shares tend to be relatively smaller. How can we explain this pattern? Why do market shares fall in the middle of the line where products locations are also more tightly concentrated? Intuitively, there are at least two conflicting hypotheses that can explain the correlation of markets shares and locations. One hypothesis is that travel costs are relatively large, and therefore products mostly compete locally. In this case, products that are located in a more crowded part of the line have lower market shares because these products face more competition for the same consumers.

An alternative hypothesis is that travel costs are zero ($\theta = 0$), and products j that are located in a more crowded center part of the line have lower market shares because these products have systematically lower qualities ξ_j.

Market share data alone cannot sort out these alternative explanations, because for any value of θ the implied quality will adjust to match the observed market share. However, assuming (4.2) will do the job. We visually show this in Figure 4.2. The figure shows a scatter plot of locations x_j and implied quality $\xi_j(\theta)$ under three scenarios for travel costs: in panel (a) $\theta = \theta^0$, in panel (b) $\theta > \theta^0$, and in panel (c) $\theta < \theta^0$.[26]

Figure 4.2: Scatter plot of product locations and quality assignment


(a) $\theta = \theta^0$

(b) $\theta > \theta^0$

(c) $\theta < \theta^0$

Notes: The figure shows a scatter plot of locations x_j and implied quality $\xi_j(\theta)$ under three scenarios for travel costs: in panel (a) $\theta = \theta^0 = 2$, in panel (b) $\theta = 4$, and in panel (c) $\theta = 0$.

[26] We use $\theta = 0$ (the Logit case) and $\theta = 4$ (a doubling of the true value $\theta^0 = 2$) as our two departures from the true value.

Panel (a) of Figure 4.2 shows that at the true parameter value there is no correlation between quality and location. Panels (b) and (c) show that when θ is different from the truth the data exhibit correlation. In panel (b) the travel cost is overstated and therefore local competition is overstated. Hence to explain the observed market shares the implied quality $\xi_j(\theta)$ has to be systematically higher in the crowded part of product space. In panel (c) the pattern is reversed.

Thus, by using the sample analog of the moment restriction (4.2) in estimation we are shutting off the opportunity for the model to explain the data through a systematic correlation between ξ_j and the local market structure across products j. As we will see below, local market structure measures serve as IVs to estimate the non-linear parameters of the model. In addition to helping us pin down Σ and the implied $\delta(\Sigma)$, the moment conditions implied by (4.2) will help us pin down β_0.

The key lesson from the above example is that as long we have an assumption like (4.2), the parameters of the model are in principle identified from a cross section of products within a market. In reality, we might need variation across markets to have more power. The key empirical challenge, which we will discuss below is how to choose the IVs that are informative. We discuss this as well as computation and other empirical details for the more general case in the next section.

4.3 The General Estimation Procedure

We now generalize the setting described in the previous section to allow for multiple markets $t = 1, \dots, T$. We continue to use a conditional moment restriction $E\left[\xi_{jt} \mid \boldsymbol{Z}_t\right] = 0$, where $\boldsymbol{Z}_t$ is a vector of exogenous variables, as the basis for estimation of heterogeneous preferences in the general Mixed Logit demand model. In Section 4.2.2 we saw the identifying power of this moment condition, when we focused only on within-market variation across products. We now discuss how to use this moment condition in practice in the general setting that has both within- and across-market variation. As we discuss below, having multiple markets potentially provides additional variation in choices set and characteristics of products, and therefore helpful in estimation.

We note that the conditional moment implies a large set of potential IVs $z_{jt} = A_j(\boldsymbol{Z}_t) \in \mathbb{R}$ for which the unconditional moment restriction holds

$$E[z_{jt}\xi_{jt}] = 0. \tag{4.6}$$

We refer to A_j as the IV function, and z_{jt} as IVs. We provide some guidance on the choice of IVs below. For a given choice of IVs, estimation proceeds on the basis of empirical analogues of the population moments (4.6) for each IV n. Let

$$m^n(\theta) = \hat{E}[z^n_{jt}\xi_{jt}(\theta)] \tag{4.7}$$

be the n^{th} moment (where $\hat{E}[\cdot]$ is the expectation taken with respect to the sample distribution), and θ are the parameters of the model. Stacking the moments as $\boldsymbol{m}(\theta) = (m^n(\theta))^N_{n=1})$, the standard GMM estimator for the problem as formulated by Berry (1994) and Berry et al. (1995) is

$$\hat{\theta} = \arg\min_{\theta} \boldsymbol{m}(\theta)' W \boldsymbol{m}(\theta) \tag{4.8}$$

for a positive definite $N \times N$ weighting matrix W. Inference can be based on the standard tools for GMM.[27]

A key practical difficulty in using this moment condition is computing ξ_{jt}, and therefore the moment, as a function of data and parameters. This is where the logic discussed in Section 4.2.1 enters. Specifically we will solve a non-linear system of equations like (4.3) for each market t in the data - first the mean utility vector $\boldsymbol{\delta_t}$ must be inverted from market shares in each market, and the econometric error $\xi_{jt}(\theta)$ then computed. Berry (1994) and Berry et al. (1995) propose a contraction mapping algorithm that globally converges and can be used for computation. In Section 4.3.4 we discuss this and other algorithms in greater detail.

4.3.1 Instrumental Variables

The above discussion assumed that we have exogenous variables $\boldsymbol{Z}_t$ that can be used as the basis for constructing IVs such that (4.2) holds. We now discuss various variables that have been used in the literature for these purposes. Before we do so we review the dual role of IVs. This is best done with a series of examples. We start with the Logit model, which was described in Section 3.2. For the Logit model $\sigma_j^{-1}(\boldsymbol{s}_t, \boldsymbol{x}_t, \boldsymbol{p}_t) = \ln(s_{jt}/s_{0t})$, and the estimating equation is

$$\ln\left(\frac{s_{jt}}{s_{0t}}\right) = \delta_{jt} \equiv x_{jt}\beta_0 + \alpha_0 p_{jt} + \xi_{jt}. \tag{4.9}$$

[27] See Freyberger (2015) and Berry et al. (2004) for more detail.

This equation can be estimated using linear methods. In this case, an IV is needed if one of the right-hand side variable, say p_{jt}, is correlated with the error term. The intuition for what constitutes a good IV is just the "standard" logic in linear models. We will provide examples below.

This particular role of the IV remains even when estimate models that allow for more heterogeneity in preferences. However, now there is an additional role that links directly to our discussion in Section 4.2.2. To see this consider the Nested Logit model. As Berry (1994) shows, the inversion is now given by $\sigma_j^{-1}(\boldsymbol{s}_t, \boldsymbol{x}_t, \boldsymbol{p}_t) = \ln(s_{jt}/s_{0t}) - \rho \ln(s_{jt}/s_{G(j)t})$, and the estimating equation is

$$\ln\left(\frac{s_{jt}}{s_{0t}}\right) = x_{jt}\beta_0 + \alpha_0 p_{jt} + \xi_{jt} + \rho \ln\left(\frac{s_{jt}}{s_{G(j)t}}\right) \tag{4.10}$$

where $s_{G(j)t}$ is the share of nest G, and ρ is the nesting parameter. This model, like the Logit model, can also be estimated using linear methods. Relative to equation (4.9), which describes the estimation equation of the Logit model, in the Nested Logit model we have an additional term: the within nest share. This last term is a function of the share of product j, s_{jt}, and therefore will be correlated with ξ_{jt}. In other words, even if we believe that $E(\xi_{jt}|x_{jt}, p_{jt}) = 0$ we still cannot consistently estimate this equation using OLS: we need an IV.[28]

This formulation also suggests where we could find IVs. We want variables that generate variation in the nest share but that also satisfy (4.2). Natural candidates are variables that impact the other products in the nest, and therefore the nest share, but are uncorrelated with ξ_{jt}. For example, if looking at a given market over time we might want to use the entry and exit of products into the nest.

As we move to models that allow for even richer patterns of heterogeneity we generally do not have an analytic expression for the inversion. However, the intuition from above continues to hold. Gandhi and Houde (2020) and Gandhi et al. (2021) show that $\sigma_j^{-1}(\boldsymbol{s}_t, \boldsymbol{x}_t, \boldsymbol{p}_t) = \ln(s_{jt}/s_{0t}) - f_j(\boldsymbol{s}_t, \boldsymbol{x}_t, \boldsymbol{p}_t)$, where $f_j(\cdot)$ is an unknown function that in the BLP algorithm (discussed in Section 4.3.4) is computed numerically. In Section 4.4.2 we discuss ways to approximate $f_j(\cdot)$.

We now turn to a discussion of what exogeneity restrictions and IVs researchers have used in practice.

[28] As we noted on page 21, the Nested Logit model can be viewed as a special case of the Mixed Logit model. In the above formulation of the model ρ is the equivalent of Σ in our general formulation.

BLP Instruments By far, the most popular IVs are $\boldsymbol{Z}_t = \boldsymbol{x}_t$, namely the characteristics of all products in the market.[29] Typically price, and maybe advertising, will be excluded from this set. The identifying power of this exogeneity restriction is based on the same logic that we saw in Sections 2 and 4.2.2. They are informative because they can be used to measure the proximity of competition, just as we saw in Figure 2.1, and therefore should be correlated with price and other endogenous variables. They will also be correlated with terms like the within nest share, in the Nested Logit model, or $f_j(\boldsymbol{s_t}, \boldsymbol{x}_t, \boldsymbol{p}_t)$ in the more general model.

The question still remains which of the many possible functions $A_j(\cdot)$ of $\boldsymbol{x}_t$ should we use to construct IVs. Different suggestions have been made in the literature. Berry et al. (1995) propose to use: (1) own product characteristics, $\boldsymbol{x}_{jt}$, (2) sum of characteristics of the other products produced by the same firm (for multi-product firms), $\sum_{j' \neq j, j' \in \mathcal{J}_{f(j)}} \boldsymbol{x}_{j't}$, and (3) sum of the characteristics of competitor products $\sum_{j' \notin \mathcal{J}_{f(j)}} \boldsymbol{x}_{j't}$. The logic for this particular set is as follows. The own product characteristics are instruments "for themselves." The other two sets try to capture the logic that the price of product j (and $f_j(\cdot)$) will depend on characteristics of other products, and that the dependence differs if these are own products or competitors products.[30]

Gandhi and Houde (2020) propose to refine how we use the information in (4.6) in order to improve empirical performance and avoid weak IV challenges that can arise in practice. Specifically, they use the economic structure of the model to motivate a class of IVs they term "differentiation Instruments", which intuitively capture the relative isolation of each product in characteristics space. Defining $\boldsymbol{d}_{jkt} = \boldsymbol{x}_{jt} - \boldsymbol{x}_{kt}$ as the vector of characteristic differences between product j and product k in market t, they construct two two distinct sets of differentiation IVs that are useful for applied work.

The first set consists of (1) own product characteristics, $\boldsymbol{x}_{jt}$, (2) the distance squared between product j and other products along dimension k, $\sum_{j' \neq j} \left(d^k_{jj't}\right)^2, \forall k$ and (3) the interaction between the distance in dimensions k and l, $\sum_{j' \neq j} d^k_{jj't} \times d^l_{jj't}, \forall k \neq l$. The sum of square of characteristic differences captures a continuous measure of product isolation proportional to the Euclidean distance of product j along each dimension k. The interaction terms capture the covariance between two dimensions of differentiation.

[29] One of the reasons for the popularity of these IVs is that they typically do not require any additional data: they are part of the data needed to estimate the model to start with.

[30] There is some disagreement among researchers about whether the term "BLP Instruments" refers narrowly to the these specific functional forms or more broadly to the idea $E[\xi_{jt} \mid \boldsymbol{x}_t] = 0$.

A second set consists of: (1) own product characteristics, $\boldsymbol{x}_{jt}$, (2) the number of products within a certain "band" of j, $\sum_{j'\neq j} 1\left(|d^k_{jj't}| < \kappa^k\right), \forall k$, and (3) the interactions between the number in dimensions k and l, $\sum_{j'\neq j} 1\left(|d^k_{jj't}| < \kappa^k\right) \times 1\left(|d^l_{jj't}| < \kappa^l\right), \forall k \neq l$. These IVs try to capture the economics behind models of localized competition. The second element measures the number of "close-by" products along each dimension of differentiation. The interaction of the indicator function with $\boldsymbol{d}_{jj't}$ captures the correlation in characteristics between firms that are direct competitors. When characteristics are discrete, the indicator variables can be replaced by $1(d^k_{jj't} = 0)$; which can be thought of as a product-segment indicator. Moreover, additional neighborhoods can be constructed to impose additional restrictions on the model (e.g. $0 < |d_{jj't}| \leq \kappa_1$, $\kappa_1 < |d_{jj't}| \leq \kappa_2$, etc.)

All these various permutations of IVs are motivated by the search for "powerful" IVs, assuming the IVs are valid, namely that $E(\xi_{jt}|\boldsymbol{x}_t) = 0$. It is not difficult to come up with economic models where this validity is violated. For example, if characteristics are chosen by the firms after they observe (some components of) $\boldsymbol{\xi}_t$ then this assumption will be violated. A typical defense of the assumption is that even if the characteristics are chosen, they are chosen in advance and before $\boldsymbol{\xi}_t$ is observed. For example, in the case of cars elements of design are chosen many years in advance.

This is only a partial defense since firms might be forward-looking and could anticipate in part the realization of $\boldsymbol{\xi}_t$. This can be dealt with, if we observe panel data, by relying on the ideas of the dynamic panel literature (Arellano and Bond, 1991; Blundell and Bond, 1998). For example, Sweeting (2013) assumes that $\xi_{jt} = \rho\xi_{jt-1} + u_{jt}$ where u_{jt} is unanticipated at time $t-1$. He then bases the estimation on the conditional moment

$$E\left[\xi_{jt} - \rho\xi_{jt-1} \mid \boldsymbol{x}_{t-1}\right] = 0. \tag{4.11}$$

Hausman Instruments The "textbook" IVs for prices when estimating demand are cost variables. In most IO applications cost is not observed. Furthermore, even if we observe (marginal) cost, or some proxies for it, rarely will it vary by product.[31] Hausman et al. (1994) and Hausman (1996) propose using prices in other markets as IVs, often called "Hausman IVs". Nevo (2001, 2000a) builds on this idea to estimate a discrete choice model. To see how these IVs operate, consider estimation of equation (4.9) and assume that p_{jt} is correlated with ξ_{jt}. The idea is to use prices in other markets, namely, $p_{jt'},\ t' \neq t$,

[31] Villas-Boas (2007) uses cost IVs by gathering information on input prices and interacting these prices with product dummy variables. This is trying to capture the idea different products use a different mix of inputs and therefore will have a different relationship between prices and input prices.

as IVs. Depending on the structure of the data we could use all $t' \neq t$, or markets in the same time period, same region, or otherwise matched. These IVs are potentially valid if, conditional on $\boldsymbol{x}_t$ and $\boldsymbol{x}'_t$, pricing is independent across markets and ξ_{jt} and $\xi_{jt'}$ are independent. These IVs are trying to exploit common cost shocks across markets for identification.

There are two main problems with these IVs. First, it is not difficult to come up with arguments why they are not valid. For example, if there is an (unobserved) promotional or advertising campaign across markets then the independence assumption would be violated. Second, it is less obvious how the prices of the own brand in other markets will help, for example, in estimation of equation (4.10): it is not clear that the these IVs will be correlated with the within nest share.[32] In principle, one could use $p_{j't'}, j' \neq j,$ and $t' \neq t$, i.e., the price of other products in other markets, to proxy for cost shocks to other products. However, we are unaware of a published paper that uses this approach.

Waldfogel Instruments In some cases researchers have used attributes of other markets, such as demographics, in a slightly different way than Hausman instruments. For example, Town and Liu (2003) estimate the welfare associated with the Medicare HMO program known as Medicare+Choice. To do so they estimate a Nested Logit model at the county-level and use the fact that each plan is typically offered in several counties. One of the IVs they use is the mean number of competitors in the other counties where the plan is offered. Similarly, Fan (2013), when estimating demand for newspapers using county-level data, exploits the fact that newspapers sell in multiple counties and uses demographics in other counties as IVs.

Like Hausman IVs, these IVs use information in other markets, but the logic is different: Hausman IVs rely on common cost shocks, while these IVs rely on consumption or preference externalities. If the product is offered in multiple counties the price and characteristics of the product will be impacted by the attributes, say demographics, in the other counties. So, for example, if a product is offered in counties A and B its price should be a function of demographics in both counties. For this reason these IVs are often referred to as ”Waldfogel IVs” (Waldfogel, 2003). These IVs are valid if, conditional on the variables included in the model, ξ_{jt} is not correlated across counties, just like the requirement for the Hausman IVs. For the same reason this assumption was suspect there,

[32] As Berry and Haile (2014) show, separate IVs are needed for prices relative to market shares - the model has two distinct sets of endogenous variables (as we saw in the Nested Logit example).

it could be suspect here as well. Furthermore, the set of counties covered by a plan is not exogenous and could be an indication that the counties are similar in some ways.

4.3.2 Additional Sources of Variation

We now briefly discuss additional sources of variation that can aid in estimation and identification.

Multiple markets Some of the above IVs discussed above could in principle be constructed with data from a single market, e.g., BLP IVs (as we saw in Section 4.2.2. However data from multiple markets can significantly aid in identification and estimation.

In the case of BLP IVs a main advantage of having multiple markets is the potential for variation in the number of products and their characteristics. For example, consider the estimation of equation (4.10) using data across markets. The within nest share might vary because different products are available in different markets. This has been found to be a powerful way to estimate this model, especially in cases where entry and exit of products are arguably exogenous. This idea generalizes. The intuition given in Section 4.2.2 that focused on the informational content of equation (4.2) with data from a single market, can be extended to multiple markets, especially as competitive conditions vary across markets.

An additional way data from multiple markets can be used is through demographic data. As we saw in Section 4.2.1 having consumer level data can aid in identification and estimation. Data from multiple markets, with variation in demographics can achieve similar results. For example, suppose that we observe markets with different distribution of ages. This allows us to correlate the outcomes and demographics. We can do that by imposing "micro moments" as we discuss below. Or by using the full distribution of demographics to compute the shares in equation (3.7). See Nevo (2000b) and the computational discussion below for details.

Furthermore, having data from multiple markets allows us to control for unobserved product characteristics that do not vary across markets . Finally, Hausman IVs and Waldfogel IVs require multiple markets in order to be computed. In sum, data from several markets is very helpful in estimation of the model, and in general the more the better.

Micro moments and Second Choice Data As we saw in Section 4.2.1 having information on demographics and consumer choices can be very useful in estimation. The intuition gained from equation (4.4) can be useful even if the researcher does not have consumer choice data (i.e., data that includes both the choices of individuals and their demographics). These moments can be computed from other data sources and added to the estimation. For example, suppose a researcher is estimating demand for cars and has information on the average family size conditional on owning a minivan. We can match the model's prediction of this choice behavior in the population with the sample analogues. This mimics the logic of the moments in equation (4.4), "as if" we had micro data.

Petrin (2002) follows this approach and finds that the micro moments impact the estimates of consumer heterogeneity and have important implications for his estimate of the welfare gains from the introduction of minivans. Follow-up research has found similar results. It is therefore advisable to add micro moments whenever possible.[33]

Another source of data that is powerful, albeit much harder to obtain, is second choice data. Berry et al. (2004) use survey data on second choices, i.e., what the consumer would choose if the actual chosen alternative were not in the choice set. Such data provide a direct empirical insight into substitution patterns among products and a useful source of identification and estimation.

Supply-Side Moments Another way to help identify the parameters is to add supply-side moments. Assume that the marginal cost is given by

$$mc_{jt} = \mathrm{w}_{jt}\boldsymbol{\gamma} + \omega_{jt},$$

where $\mathrm{w}_{\boldsymbol{jt}}$ is a vector of observed characteristics of product j, ω_{jt} is an unobserved component, and $\boldsymbol{\gamma}$ is a vector of parameters to be estimated. If we further assume the Nash-Bertrand pricing model discussed in Section 2.1.1 and combine the cost with the pricing equation (2.2) we get

$$\boldsymbol{p}_t = \mathrm{w}_{\boldsymbol{t}}\boldsymbol{\gamma} + \Omega^{-1}\boldsymbol{q}(\boldsymbol{p_t}) + \boldsymbol{\omega}_t. \tag{4.12}$$

Using this equation we can form supply-side moments by assuming $E(\omega_{jt}|\boldsymbol{Z}_t) = 0$, where $\boldsymbol{Z}_t$ is a vector of IVs that includes products characteristics and cost shifters ($\boldsymbol{z}_{jt} = [\boldsymbol{x}_t, \mathrm{w}_{jt}]$ and $\boldsymbol{Z}_t = [z_{1t}, z_{2t}, ...z_{Jt}]$.) Note, that this equation is informative about both the supply parameters, γ and the demand parameters, which impact Ω. It also suggests that the cost

[33] See Grieco et al. (2021) for an efficient estimator when micro data is available.

shifters could be used as additional demand-side IVs. We can combine the demand and supply-side moments and estimate the parameters using GMM.

4.3.3 Efficiency

The conditional moment $E\left[\xi_{jt} \mid \boldsymbol{Z}_t\right] = 0$ implies a large set of potential unconditional moments $E[z_{jt}\xi_{jt}] = 0$, where $z_{jt} = A_j(\boldsymbol{Z}_t)$ are IVs, which are the basis for estimation. The choice of IVs is closely tied to the efficiency of the GMM estimator. We can use the semi-parametric efficiency bound (Chamberlain, 1987) to guide strategies for constructing IVs and efficient estimation. There are two basic approaches.

1. We can allow the dimension of the IVs $A_j(\boldsymbol{Z}_t)$ to grow with the sample size and capture the informational content in $E[\xi_{jt} \mid \boldsymbol{Z}_t] = 0$. The asymptotic efficiency bound is reached by applying the optimal weight matrix W^* to the GMM problem with a suitably rich set of "low-order" basis functions as IVs (see, for example, Donald et al. (2003)). The Differentiation IVs above can serve as such a basis class.

2. We can compute the "optimal IVs" from Chamberlain (1987). This approach is especially likely to be productive, if IVs $A_j(\boldsymbol{Z}_t)$ do not approximate well the full informational content in $E[\xi_{jt} \mid \boldsymbol{Z}_t] = 0$. Theoretically, the optimal IVs are given by

   $$z^*_{njt} = E\left[\frac{\partial \xi_j(\theta^0)}{\partial \theta}\Big|\boldsymbol{Z}_t\right] \qquad n = 1, \ldots, dim(\theta) \tag{4.13}$$

   where θ^0 is the true value of the parameters. This creates a just-identified problem, e.g., as many IVs as parameters and the weighting matrix is the identity matrix. These are obviously not feasible but can be heuristically approximated. For example, Reynaert and Verboven (2014)) explore the heuristic

   $$E\left[\frac{\partial \xi_j(\theta^0)}{\partial \theta}\Big|\boldsymbol{Z}_t\right] \approx \frac{\partial \xi_{jt}(\boldsymbol{s}_t, \hat{\boldsymbol{p}}_t, \boldsymbol{x}_t; \hat{\theta})}{\partial \hat{\theta}}\bigg|_{\xi_{jt}=0, \forall j,t} = A_j(\boldsymbol{x}_t). \tag{4.14}$$

   Since the IV vector depends on θ, users must first obtain an estimate of the parameters, denoted by $\hat{\theta}$.

The two approaches are ultimately complementary as discussed in Gandhi and Houde (2020) and Conlon and Gortmaker (2020). The performance of the optimal IV approximation depends critically on good first-stage estimates. In particular if weak IVs are used in the first stage, the approximation will not work well in practice.

4.3.4 Computational Algorithms

Several computational algorithms have been suggested in the literature to solve the estimation problem we discussed above. We focus on three of them.[34]

Nested Fixed Point Berry et al. (1995) provide a method to compute the estimator they propose. We now describe the basic steps, for more details see Nevo (2000b) and for updated best practices (and code) see Conlon and Gortmaker (2020). The method consists of the following steps.

In a preliminary step, we draw R random draws from $F_\nu(\nu)$, which is the (standardized) parametric distribution assumed for ν (almost always a standard normal), and $\hat{F}_D$, which is either an estimated parametric distribution (e.g., log normal for income estimated outside the model), or an empirical distribution (for example, from Census data). These draws are held constant throughout the computation. Denote them as $\hat{F} = \left\{\hat{\nu}_{it}, \hat{D}_{it}\right\}_{i=1}^{R}$. In most data sets, we will observe quantities and not market shares. Quantities are converted to market shares by making an assumption on the total market size, namely all the consumers who purchase and those who decided not to purchase. Let I_t denote this quantity, then $s_{jt} = q_{jt}/I_t$. With these preliminaries in hand the algorithm proceeds as follows.

1. Step 1: for a given value of the "non-linear" parameters, Γ and Σ, and vector of mean utilities $\boldsymbol{\delta}_t$ compute market shares predicted by the model. The easiest way to do this is via simulation using the draws from the preliminary step,[35] namely:

$$\tilde{\sigma}(\boldsymbol{\delta}_t; \Gamma, \Sigma, \boldsymbol{x}_t, \boldsymbol{p}_t, \hat{F}) = \frac{1}{R}\sum_{i=1}^{R} \frac{exp\left\{\delta_{jt} + (x_{jt}, p_{jt}) \cdot (\Gamma D_{it} + \Sigma \nu_{it})\right\}}{1 + \sum_{k=1}^{J} exp\left\{\delta_{kt} + (x_{kt}, p_{kt}) \cdot (\Gamma D_{it} + \Sigma \nu_{it})\right\}}.$$

2. Step 2: For for a given value of the "non-linear" parameters, Γ and Σ, compute the vector of mean utilities that equates the shares predicted in Step 1 to those observed in the data. This can be computed by starting with a guess for $\boldsymbol{\delta}_t$ (say the values

[34] See Hong et al. (2021) for an additional method that we do not discuss here.

[35] See Nevo (2000b) and Conlon and Gortmaker (2020) for alternative, more efficient ways to compute these shares.

from the Logit model $\ln(s_{jt}/s_{0t})$) and computing the contraction mapping proposed by Berry (1994):

$$\boldsymbol{\delta}_t^{r+1} = \boldsymbol{\delta}_t^{r} + \ln(\boldsymbol{s}_t) - \ln\tilde{\sigma}(\boldsymbol{\delta}_t^{r};,\Gamma,\Sigma,\boldsymbol{x}_t,\boldsymbol{p}_t\hat{F}).$$

Stop when $\left\|\boldsymbol{\delta}_t^{r} - \boldsymbol{\delta}_t^{r-1}\right\| < \tau$, where τ is a pre set tolerance level (say 10^{-12}).[36]

3. Step 3: Use the result of step 2, $\delta_{jt}(\Gamma,\Sigma)$, to compute $\xi_{jt} = \delta_{jt}(\Gamma,\Sigma) - x_{jt}t\beta_0 - \alpha_0 p_{jt}$, interact it with the IVs to form the GMM objective function and compute (4.8) using a non-linear search routine.

This algorithm is relatively easy to program, although to improve computational speed various bells and whistles are needed. See Nevo (2000b) and Conlon and Gortmaker (2020) for details and code.

Mathematical Programming with Equilibrium Constraints (MPEC): Dubé et al. (2012) advocate the use of an MPEC algorithm instead of the above Nested fixed point. The basic idea is to maximize the same GMM objective function as above subject to the constraints that the predicted shares equal the observed shares. However, demand shocks, $\boldsymbol{\xi}$ are treated as parameters. Formally,

$$\begin{aligned} \min_{\theta,\xi} \quad & \boldsymbol{\xi}'\boldsymbol{z}\boldsymbol{W}\boldsymbol{z}'\boldsymbol{\xi} \\ \text{subject to} \quad & \tilde{\sigma}(\boldsymbol{\delta}(\boldsymbol{\xi});\boldsymbol{x},\boldsymbol{p},\hat{F},\theta) = \boldsymbol{s} \end{aligned}$$

Note that both θ and ξ in this problem and therefore the search is over a much higher dimension search than (4.8). $\boldsymbol{\xi}$ is a now a vector of parameters, and unlike before it is not a function of θ. The advantage of this approach is that it avoids the need to perform the inversion at each and every iteration of the search. This inversion can be a significant computational cost, especially when performed for values of the parameters far from θ^0. The resulting programming problem can be quite large, but there are off-the-shelf programs (e.g., Knitro) that can solve it effectively. Dubé et al. (2012) report significant speed improvements over the nested fixed point. Note, that this is purely a computational algorithm: the result should be identical to the result of the BLP algorithm.

This approach is more complicated to program, and to get the computational benefits one needs to analytically provide various derivatives. Once programmed properly

[36] With a tolerance level that is not strict enough the algorithm can become unstable as shown by Knittel and Metaxoglou (2014).

it seems to perform well, but some have found it slow in very large problems (many markets and many products) and not worth the extra programming time.

Approximate BLP (ABLP): Lee and Seo (2015) propose an alternative estimator with significant computational advantages that they call Approximate BLP. The basic idea is to approximate the share equation $\sigma(\cdot)$ using a first-order Taylor approximation. This allows them to substitute an analytic inversion, for the numerical inversion, in Step 2 of the original BLP algorithm. Like the MPEC algorithm the inversion is exact only at the solution, but unlike MPEC the optimization is over a lower-dimensional parameter space.

They compute a first-order Taylor approximation to $\tilde{\sigma}(\boldsymbol{\xi}_t; \boldsymbol{x}_t, \boldsymbol{p}_t, \hat{F}, \theta) \equiv \tilde{\sigma}(\boldsymbol{\delta}(\boldsymbol{\xi}_t); \boldsymbol{x}_t, \boldsymbol{p}_t, \hat{F}, \theta)$ around $\boldsymbol{\xi}_t^0$ given by

$$\ln \tilde{\sigma}(\boldsymbol{\xi}_t; \boldsymbol{x}_t, \boldsymbol{p}_t, \hat{F}, \theta) \approx ln\tilde{\sigma}^A(\boldsymbol{\xi}_t; \boldsymbol{x}_t, \boldsymbol{p}, \hat{F}, \theta) \equiv \ln \tilde{\sigma}(\boldsymbol{\xi}_t^0; \boldsymbol{x}_t, \boldsymbol{p}_t, \hat{F}, \theta) + \frac{\partial \ln \tilde{\sigma}(\boldsymbol{\xi}_t^0; \boldsymbol{x}_t, \boldsymbol{p}_t, \hat{F}, \theta)}{\partial \ln \boldsymbol{\xi}_t'}(\boldsymbol{\xi}_t - \boldsymbol{\xi}_t^0).$$

They equate this approximation to the observed shares, and invert this relation to get

$$\boldsymbol{\xi}_t = \Phi_t(\boldsymbol{\xi}_t^0, \theta) \equiv \boldsymbol{\xi}_t^0 + \left[\frac{\partial \ln \tilde{\sigma}(\boldsymbol{\xi}_t^0; \boldsymbol{x}_t, \boldsymbol{p}_t, \hat{F}, \theta)}{\partial \ln \boldsymbol{\xi}_t'}\right]^{-1} (\ln s_t - \ln \tilde{\sigma}(\boldsymbol{\xi}_t^0; \boldsymbol{x}_t, \boldsymbol{p}_t, \hat{F}, \theta)).$$

This analytic inversion (of the approximation) allows them to skip the numerical inversion in Step 2 of BLP.

With the aid of the approximation they can estimate the parameters θ by

$$\min_\theta \Phi(\boldsymbol{\xi}^0, \theta)' \boldsymbol{z} \boldsymbol{W} \boldsymbol{z}' \Phi(\boldsymbol{\xi}^0, \theta)$$

They nest this idea into the following procedure. Guess $\boldsymbol{\xi}_t^0$ and set $r = 1$

1. Step 1: Compute a GMM estimate

$$\theta^r = \arg\min_\theta \Phi(\boldsymbol{\xi}^{r-1}, \theta)' \boldsymbol{z} \boldsymbol{W} \boldsymbol{z}' \Phi(\boldsymbol{\xi}^{r-1}, \theta)$$

2. Step 2: Update $\boldsymbol{\xi}$

$$\boldsymbol{\xi}^r = \Phi(\boldsymbol{\xi}_t^{r-1}, \theta^r)$$

and $r = r + 1$

3. Repeat Steps 1 and 2 until convergence.

Lee and Seo show this estimator is equivalent to the BLP estimator in large samples. The advantage of this approach is that like MPEC it avoids inversion at each stage, but has low-dimensional parameters search.

4.4 Extensions

4.4.1 Error in Market Shares

We typically assume that aggregate quantities are based on a large number of underlying choices and therefore measured without error. This assumption can be problematic when products have a small market share: with a large number of products even with samples generated from thousands of consumers market shares can be measured with error. This is especially problematic if the data include a large number of products with a market share of zero. Ad hoc fixes, sometimes used in practice, such as dropping zeros from the data or replacing them with small positive numbers, are subject to biases which can be quite large. Mathematically, the root of the problem is that the slope of the inverse demand function approaches infinity as the share approaches zero.

Solutions to this problem can be split into two groups, depending on how we view the root for the problem. Which approach is more appropriate for a specific data set depends on how the zeroes are believed to be generated in that data set. The first group views the root of the problem as the wedge between choice probabilities, which come from the theoretical demand model, and market shares, which are the empirical estimates based on the realized choices of consumers in the data. Although the choice probabilities are strictly positive in the underlying model, observed shares may be zero due to sampling error. This is more likely if the underlying choice probability is small.

Gandhi et al. (2021) take this approach. At a high level, they construct lower and upper bounds for the inverse demand function by adding a bit of noise to the observed shares. If they observe a set of products whose empirical shares are unlikely to be zero, then they can point identify the parameters. If there are no such products (for example, because the number of consumers is small), their bound construction leads to a set of moment inequalities that partially identify the parameters. They apply this approach to scanner data and find that the new approach yields demand estimates that can be more than twice as elastic as standard estimates that select out the zeroes.

Dubé et al. (2021) tackle the zeroes problem from a different angle. They assume that $s_{jt} = 0$ if and only if product j is not in the set of products that the consumers in market t consider, or in other words, the choice set. They then offer a specific model of the selection into the choice set in order to estimate the model. Their results rely on carefully placed separability and exclusion restrictions.

4.4.2 Non-parametric and Flexible Estimation

Up to this point we have assumed a parametric functional form for utility, given in (3.3), and specific distribution of heterogeneity. The identification results in Berry and Haile (2014) hold for more general models. Therefore, an obvious question is whether more flexible models, that imply flexible substitution patterns, can be estimated.

In order to appreciate the problem of flexible estimation, it is useful to recall the integral (3.7) that defines aggregate demand in the Mixed Logit model. Up until now in our discussion we have treated the distribution of consumer "types" in the population $F(D_{it}, \nu_{it})$ as a known distribution: we assume we have data to estimate the distribution of D_i, and we assumed a parametric distribution for ν_i. A more flexible model is to keep the type-1 extreme value distribution assumption on ε_{ijt}, but allow for a flexible mixing distribution. The joint distribution $F(D_{it}, \nu_{it})$ is an unrestricted distribution that is estimated from the data. In this case (3.7) can be treated as an integral equation for identifying F.[37] [38] A further generalization maintains the linear utility form in equation (3.3) but treats $(\alpha_i, \beta_i, \varepsilon_{it})$ as distributed according to a general distribution $F(\alpha_i, \beta_i, \varepsilon_{it})$. In this case the integral equation becomes

$$\begin{aligned} &s_{jt} = s_j\left(\boldsymbol{x}_t, \boldsymbol{p}_t, \boldsymbol{\xi}_t\right) = \\ &\int \mathbf{1}\left(u(x_{jt}, p_{jt}, \xi_{jt}; \alpha_i, \beta_i, \varepsilon_{ijt}) > u(x_{kt}, p_{kt}, \xi_{kt}; \alpha_i, \beta_i, \varepsilon_{ikt}), \forall k \neq j\right) \, dF(\alpha_i, \beta_i, \varepsilon_{it}; \theta). \end{aligned} \tag{4.15}$$

The question in either (3.7) or (4.15) is whether the distribution F can be estimated in a flexible way, e.g., either non-parametrically or with a flexible parameterization θ. Un-

[37] The joint distribution $F(D_{it}, \nu_{it})$ can be constrained so that the marginal F_D equals the actual distribution of demographics in the market, which may be observed/known.

[38] With micro data the problem is different. For example, Dubois et al. (2020) utilize a long panel of consumer choices to estimate a Logit model where the coefficients on price and characteristics are estimated separately for each consumer, which avoids making distributional assumptions on the random coefficients.

fortunately the presence of ξ_{jt} in the integral equation (4.15) complicates the application of standard estimators for flexible heterogeneity in discrete choice models.[39]

One approach to the problem is to change the focus away from estimation of preferences to estimation of demand by estimating the demand function σ (or more precisely, $\sigma^{-1}(\cdot)$) directly in a flexible way. The basic idea is to approach estimation of the inverse demand function $\sigma^{-1}(\boldsymbol{s}_t, \boldsymbol{p}_t, \boldsymbol{x}_t)$ directly rather than estimating a model of preferences first as a means to constructing demand. These approaches all have to address the dimensionality problem of $\sigma^{-1}(\cdot)$ that arises without an explicit preference structure - the model is now expressed in product space and as we discussed in Section 3.1, the number of parameters to estimate can be very large.

Compiani (2019) is an example of this approach. He proposes to directly non-parametrically estimate the inverse demand function $\sigma^{-1}(\cdot)$ through a sieve approximation. This has the advantage of requiring fewer assumptions than aggregating demand from a random coefficient choice model - in principle it only requires invertibility of demand that is guaranteed by the connected substitutes condition in Berry et al. (2013). Thus, the class of demand models that are consistent with the estimator is broader than Mixed Logit models, including models with some degree of product complementarity as well as models that allow for behavioral economic effects at the level of the consumer. The cost is that there is a curse of dimensionality encountered in a product space model, which was the motivation for using explicit choice models in the literature as discussed above. Compiani (2019) shows that Bernstein polynomials can allow for some parsimony to be added to the problem through linear constraints on the parameters that are motivated by theoretical restrictions on demand implied by choice models, such as monotonicity and symmetry. The specification nevertheless requires significant data for moderately sized markets.

Another example is Salanié and Wolak (2019) who use the idea of "artificial regressions" from Davidson and MacKinnon (1989) that takes a first-order expansion of the residual function used in non-linear IV estimation, e.g., a first-order expansion of $\xi_{jt}(\theta)$ around an initial value for $\theta = \theta'$. The approach retains the full parsimony of the underlying Mixed Logit model, and is "fast" in the sense of being a single IV regression, however it is approximate in that it only iterates the regression a single time (multiple iterations would equate to a Gauss-Newton optimization of the GMM objective function) and also

[39] See e.g., Fox and Gandhi (2016), Fox et al. (2011), and Fox et al. (2016) for a discussion of identification and estimation of discrete choice models with flexible heterogeneity where product-market unobservables ξ_{jt} are not present.

requires a starting value to run. It is a fast way however to generate starting values that can be used for non-linear estimation.[40]

A third example is Fosgerau et al. (2020) who generalize the analytic structure of the Nested Logit functional form for inverse demand to a broader class of consumer demand models - what they call "inverse product differentiation logit" model. The approach is also based on linear IV estimation, and the generalized term they add to the model allows for complementarity among products as well as being consistent with an underlying representative agent model of consumer demand. However it is a preference model expressed in "product space" as opposed to a characteristics space model, and requires the researcher to specify ex-ante dimensions of segmentation in the market where products can be categorized.

A challenge with the above approaches is that while they estimate the demand function $\sigma(\cdot)$ in a flexible way under different conditions, they do not estimate the distribution F of consumer heterogeneity that is a central to many applications. Gandhi et al. (2021) take a different approach that allows researchers to flexibly recover both demand $\sigma^{-1}(\cdot)$ and the distribution of heterogeneity F. They proceed in two steps. In the first step, like the above papers, they estimate the inverse demand model. Their main specification, does not directly tie the inverse demand to preference model parameters, unlike Salanié and Wolak (2019) and Fosgerau et al. (2020). However, unlike Compiani (2019) they use the structure of Mixed Logit demand systems to capture the parsimony attained in characteristic space models in a product space specification of demand via exchangeability properties. Building on the discussion introduced in Section 4.3.1 they estimate a specification

$$\ln(s_{jt}/s_{0t}) = \alpha p_{jt} + x_{jt}\beta + f_j(\boldsymbol{s}_t, \boldsymbol{x}_t, \boldsymbol{p}_t) + \xi_{jt}. \tag{4.16}$$

It is convenient to rewrite $f_j(\cdot)$ as a function of $\{(s_{kt}, \boldsymbol{d}_{jkt})\}_{k\neq j}$, where $\boldsymbol{d}_{jkt} = \boldsymbol{x}_{jt} - \boldsymbol{x}_{kt}$, is the vector of distance in characteristics space (and price) to other products, as defined on page 36. We can think of this as a re-normalization that focuses on firm j and measures the distance of competitors from it. The key result is that in the Mixed Logit model $f_j(\cdot)$ is symmetric, or exchangeable, in its arguments (i.e., it depends on the states, but not their order) and does not depend on j.

[40] Lee and Seo (2015) approximate $\sigma(\cdot)$ using Newton expansions for the purposes of proposing an alternative estimator to nested fixed point as discussed above.

A consequence of exchangeability, also discussed further in Gandhi and Houde (2020), is that $f_j(\cdot)$ can WLOG be represented as

$$f_j(\boldsymbol{s}_t, \boldsymbol{x}_t, \boldsymbol{p}_t) = g(EDF\left(\{(s_{kt}, \boldsymbol{d}_{jkt})\}_{k \neq j}\right)),$$

where EDF denotes the empirical distribution function taken over the products in market t (specifically, a distribution over all products $k \neq j$ in market t). Based on this representation, they propose approximating f by using the first- and second-order set of empirical moments to approximate the EDF above. In principle, one could use higher moments as well. These moments are similar in spirit to the "within-group share" term in the Nested Logit model (4.10). These terms are endogenous, which couples closely with the IVs proposed in Gandhi and Houde (2020) that is based on a similar theoretical structure. They also use a flexible functional form for $g(\cdot)$, which they take to be a generalized additive model in each one of the moments used to approximate the EDF. There are other flexible approximations that may also work in practice. Finally, using the Implicit Function Theorem they show how these first stage estimates can be used to recover own- and cross-price elasticities.

To recover the distribution of heterogeneity, they recover ξ_{jt} as the residual from the demand equation (4.16). In the second step of their procedure they plug this residual into equation (4.15). This controls for the effect of the ξ_{jt} in the integral equation and allows them to estimate the distribution of heterogeneity, F, in a flexible parametric or non-parametric way using standard mixtures techniques. A key benefit of the approach even relative to the parametric Mixed Logit estimation of F discussed earlier is that it avoids the numerical complexity of demand inversion. This confers several benefits for the speed, reliability, and robustness of the estimator.

5 Supply

Having discussed ways to specify and estimate demand for differentiated products, we now turn to the supply side. Our focus is on pricing, but in principle the analysis below can apply to any continuous characteristic that can be flexibly adjusted.[41] We have two goals in this section. First, we aim to show a few applications of the demand models

[41] For example, Fan (2013) looks at characteristics choice by newspaper, and how they change after a merger.