Identification and Estimation in Discrete Choice Demand Models when Endogenous Variables Interact with the Error

Amit Gandhi
University of Wisconsin-Madison

Kyoo il Kim
University of Minnesota, Twin Cities

Amil Petrin
University of Minnesota, Twin Cities and NBER*

December 2009, updated March 2011

Abstract

We develop an estimator for the parameters of a utility function that has interactions between the unobserved demand error and observed factors including price. We show that the Berry (1994)/Berry, Levinsohn, and Pakes (1995) inversion and contraction can still be used to recover the mean utility term that now contains both the demand error and the interactions with the error. However, the instrumental variable (IV) solution is no longer consistent because the price interaction term is correlated with the instrumented price. We show that the standard conditional moment restrictions (CMRs) do not generally suffice for identification. We supplement the standard CMRs with new moments that we call "generalized" control function moments and we show together they are sufficient for identification of all of the demand parameters. A major advantage of our setup is that it requires little more than the existence of the same instruments used in this standard IV setting. We run several monte carlos that show our approach works when the standard IV approaches fail because of non-separability. We also test and reject additive separability in the original Berry, Levinsohn, and Pakes (1995) automobile data, and we show that demand becomes significantly more elastic when the correction is applied.

1 Introduction

Demand estimation is a critical issue in many policy problems and correlation between unobserved demand factors and prices arising from market equilibration can confound estimation. In discrete choice settings the problem is complicated by the fact that the unobserved demand factor enters non-linearly into the demand equation, making standard Instrumental Variables (IV) techniques invalid. A major contribution of Berry (1994) and Berry, Levinsohn, and Pakes (1995) is to show how to invert from market shares the mean utility term. As long as the unobserved demand factor enters mean utility additively, standard IV techniques can be applied to recover the demand parameters subsumed in it.

Restricting the unobserved demand factor to enter utility additively is not always innocuous. Separability rules out several important aspects of economic behavior. For example, separability does not allow unobserved advertising to affect the marginal utility derived from observed characteristics or from the composite commodity index (typically given by residual income), even though this is often the purpose of advertising. Similarly, if the demand error represents unobserved physical characteristics, a separable setup does not allow the marginal utility derived from observed characteristics or the composite commodity index to depend on the level of the unobserved characteristic. Empirically, allowing for the possibility of a non-separable error may be important because the set of product characteristics observed by the practitioner is often limited, leaving a large role for the unobserved demand factor in explaining realized demand.

Our main contribution is to show how to consistently estimate demand parameters while allowing for observed endogenous and exogenous variables to interact with the unobserved factor. We begin by showing when endogenous variables interact with the demand error, the Berry (1994)/Berry, Levinsohn, and Pakes (1995) inversion and contraction can still be used to recover the mean utility term. However, the IV approach is no longer consistent for the parameters embedded in the mean utility term. The instrumented price is correlated with the interaction term between price and the unobserved demand factor, which is now in the estimation equation's error.

We then show in Section 3 that the conditional moment restrictions (CMR) used in the Berry/BLP setup are no longer sufficient for identification. While higher-order moments of the standard CMRs solve the identification problem if only exogenous variables interact with the demand unobservable, they do not help with identification when one (or more) endogenous variables interacts with the demand unobservable. Our non-separable setup thus provides a simple example of the failure of identification using CMRs in settings with non-separable errors (see Blundell and Powell (2003) and Hahn and Ridder (2008)).

Our setup is closest to a model of multiplicative heteroskedasticity with both exogenous and endogenous variables interacting with the error.[1] We achieve identification by coupling the Berry/BLP CMRs with new moment conditions based on insights from Kim and Petrin (2010d), who revisit the early control function literature (see Section 4). We develop a control function that conditions

---

[1]Our approach can be generalized somewhat (see Kim and Petrin (2010b)).

out the correlation between the unobserved demand factor and price. We then construct the new moments conditions based on a specification that includes the control function as an addtional explanatory variable for mean utility. For identification the control function must not have arguments that are perfectly collinear with price and other characteristics entering mean utility. We show the CMR conditions from BLP put shape restrictions on the control function that ensure this collinearity does not occur.

We prove identification for the random coefficients case using a high-level condition from Berry, Linton, and Pakes (2004) (see the Appendix). In Section 5 we provide a proof for the case without random coefficients that more clearly illustrates how combining the two types of moments achieves identification. This proof shows that we require little beyond the standard conditions for identification with valid instruments. Specifically, just as in Berry/BLP, if price is the only endogenous variable then we only require one instrument that shifts price around and is excluded from utility.

We develop a semiparametric sieve estimator for our non-separable demand model and prove consistency in Section 6. In a setting without random coefficients our estimator inverts market shares to recover mean utility and then reduces to three simple steps. With random coefficients for each evaluation of the objective function we use the BLP contraction to solve for the mean utility term and then carry out the same simple steps.[2]

In Section 7 we run three sets of Monte Carlos to illustrate implementation of our estimator and to show the possible impact of interaction terms on estimated demand elasticities. In all of the Monte Carlos both ordinary least squares (OLS) and two-stage least squares (2SLS) are significantly biased while our estimator is consistent.

We then return to the original Berry, Levinsohn, and Pakes (1995) automobile data to investigate whether allowing for interaction terms changes the estimated demand elasticities (see Section 8). In our most general specification where we include interactions terms and random coefficients, we reject at the 5% level that the coefficients on all of the interaction terms are zero, and demand elasticities increase on average by 60% relative to 2SLS.

We are aware of three other approaches that can allow for some form of non-separability with endogenous prices in discrete choice settings.[3] In the case where an observed characteristic exists that is perfectly substitutable (i.e. separable) with the unobserved demand factor, Berry and Haile (2010) show the Berry/BLP CMRs are sufficient for identification. Bajari and Benkard (2005) and Kim and Petrin (2010a) - which are based on Imbens and Newey (2009) - invert out from the pricing function a vector of controls that are exactly one-to-one functions with unobserved factors. The benefit of inverting out the unobserved factors is they are then observed, and one can allow for much more flexible non-separable settings than our setup. The drawback is that they require strong conditions on the demand and supply setting to get existence of the inverse. We provide a more detailed comparison with all three approaches in Section 4.

---

[2]Code is available from the authors for Stata.

[3]Also see a recent nonparametric bounds (partial identification) approach by Chesher, Rosen, and Smolinski (2011).

2   Utility Specification

We use a standard discrete choice model with conditional indirect utility u_{ij} given as a function of observed and unobserved product j and consumer i characteristics. We decompose utility into three components

$$u_{ij} = \delta_j + \mu_{ij} + \epsilon_{ij} \tag{1}$$

where first component, δ_j is a product-specific term common to all consumers, the μ_{ij} term captures heterogeneity in consumer tastes for observed product characteristics and can be a function of demographics, and ϵ_{ij} is a "love of variety" taste term that is assumed to be independent and identically distributed across both products and consumers. Consumer i is assumed to choose the product j out of $J+1$ choices that yields maximal utility, and market shares obtain from aggregating over consumers.

The utility component common to all consumers, δ_j, is usually given as

$$\delta_j = c + \beta' x_j - \alpha p_j + \xi_j,$$

where we normalize the mean utility derived from the outside good be zero ($\delta_0 = 0$), $x_j = (x_{j1}, \ldots, x_{jK})'$ and β are, respectively, the vector of observed (to the econometrician) product characteristics and the population average taste parameters associated with those characteristics, α is the marginal utility of income and p_j denotes the price of good j, and ξ_j is the characteristic observed to consumers and producers but unobserved to the econometrician. It may represent other physical attributes of the product or advertising that is not conditioned upon in the estimation, and it is usually found to be positively correlated with price, biasing elasticities in the positive direction.

μ_{ij} is parameterized as

$$\mu_{ij} = \sum_{k=1}^{K} x_{jk} \left(\sum_{r=1}^{R} \tau_{rk} z_{ir} \right) + \sigma_c \nu_{ic} + \sum_{k=1}^{K} \sigma_k \nu_{ik} x_{jk}$$

where $z_i = (z_{i1}, \ldots, z_{iR})$ is a vector of consumer specific demographics which may include income and $\tau_k = (\tau_{1k}, \ldots, \tau_{Rk})$ with τ_{rk} the taste parameter associated with demographic characteristic r and product characteristic k. $\tau_{rk} z_{ir}$ is then the marginal utility derived from a unit of the kth characteristic for a consumer with demographic z_{ir}. $\nu_i = (\nu_{ic}, \nu_{i1}, \ldots, \nu_{iK})$ are mean-zero standard normal idiosyncratic taste shocks for each consumer-characteristic pair and $\sigma = (\sigma_c, \sigma_1, \ldots, \sigma_K)$ are the standard deviation parameters associated with the taste shocks.

We write the vector of induced tastes for each product for individual i as $\mu_i = (\mu_{i1}, \ldots, \mu_{iJ})$. Letting $f(\mu_i)$ be the induced density and assuming ϵ_{ij} is independent and identically distributed extreme value, the market share of product j is

$$s_j(\delta) = \int \frac{e^{\delta_j + \mu_{ij}}}{\sum\limits_{k=0}^{J} e^{\delta_k + \mu_{ik}}} f(\mu) d\mu$$

Letting $\tau = (\tau_1, \ldots, \tau_K)$, Berry (1994) shows under certain conditions that a unique $\delta(\sigma, \tau) = (\delta_{1,\ldots,}\delta_J)$ exists that exactly matches observed to predicted markets shares,

$$s(\sigma, \tau, \delta(\sigma, \tau)) = s^{Data},$$

and Berry, Levinsohn, and Pakes (1995) provide a contraction mapping that locates it conditional on any values of (σ, τ). Together these results are critical for addressing the endogeneity of price.

2.1 Non-Separable Demand

Our main contribution is to extend this utility framework to a setup where we allow the mean utility term to include interactions between observed and unobserved product attributes

$$\delta_j = c + \beta' x_j - \alpha p_j + \xi_j + \sum_{k=1}^{K} \gamma_k x_{jk} \xi_j + \gamma_p(\bar{y} - p_j)\xi_j. \tag{2}$$

(γ, γ_p) is the new vector of parameters, $\bar{y}$ is representative income, and the interaction terms between the observed variables are included in x_j. Theory readily accommodates this extension (e.g. see McFadden (1981)). The γ_k's allow unobserved advertising or an unobserved product characteristic to impact the marginal utility from observed characteristics. Similarly, γ_p allows the marginal utility of income to depend on the amount of unobserved quality or unobserved advertising. Thus if γ_p is negative consumers become less price sensitive as the demand error increases.

We can continue to use the same result from Berry (1994) to establish the existence and uniqueness of a $\delta(\sigma, \tau) = (\delta_{1,\ldots,}\delta_J)$ that exactly matches observed to predicted markets shares.[4] However, if $\gamma_p \neq 0$ the standard two stage least squares estimator (or GMM estimator) that recovers the parameters contained in δ is inconsistent.

2.2 Standard 2SLS Inconsistent with Non-Separable Demand

Let the instrumented value of p_j be given by $\hat{p}_j$ and rewrite (2) as

$$\delta_j = c + \beta' x_j - \alpha \hat{p}_j + [\xi_j + \sum_{k=1}^{K} \gamma_k x_{jk} \xi_j + \gamma_p(\bar{y} - p_j)\xi_j - \alpha(p_j - \hat{p}_j)] \tag{3}$$

with the new error in brackets. There are several new components to the error but only $(\bar{y} - p_j)\xi_j$ presents an econometric problem. ξ_j is not correlated with the fitted price, $\hat{p}_j$ *asymptotically* and

---

[4]If we allow the interaction term with residual income - $(y_i - p_j)$ instead of $(\bar{y} - p_j)$ - Berry (1994)'s existence and uniqueness result no longer hold. We are working to extend Gandhi (2009)'s inversion result to this setting. This also requires us to develop a new contraction to locate $\delta(\sigma, \tau) = (\delta_{1,\ldots,}\delta_J)$. Once we have done so we can also allow for random coefficients on both ξ and on the interactions between ξ and the observed characteristics and price. This work is well beyond the scope of the current paper.

$\sum_{k=1}^{K} \gamma_k x_{jk} \xi_j$ is also uncorrelated with $\hat{p}_j$ *asymptotically* as long as the instrument(s) include x_j and they are valid. By construction $(p_j - \hat{p}_j)$ is uncorrelated with $\hat{p}_j$.

The problem arises because $\hat{p}_j$ is correlated with $\bar{y} - p_j$, leading to the possibility that $\hat{p}_j$ and $\gamma_p(\bar{y} - p_j)\xi_j$ are correlated conditional on x_j. The sign of the bias depends on the sign of γ_p and the sign of the conditional correlation of $\hat{p}_j$ and $(\bar{y} - p_j)\xi_j$. In the Berry, Levinsohn, and Pakes (1995) automobile data our estimate of γ_p is negative and the standard IV estimate is biased down, which would imply a negative correlation between $\hat{p}_j$ and $(\bar{y} - p_j)\xi_j$ conditional on x_j.

3 Conditional Moment Restrictions Alone Insufficient for Identification

We consider identification using the Berry, Levinsohn, and Pakes (1995) (BLP) conditional moment restrictions (CMR). We collect the model parameters into $\theta = (c, \beta', \alpha, \gamma', \gamma_p)'$ and denote its true value by θ_0. A set of instruments z_j is presumed to exist such that

$$E[\xi_j(\theta_0)|z_j] = 0.$$

We follow BLP and assume z_j includes all observed product characteristics and income. Letting $\xi_j = \xi_j(\theta_0)$, the CMR restriction leads to the moments BLP use for identification, given as

$$E[\xi_j|z_j] = E[\delta_j - (c_0 + \beta_0'x_j - \alpha_0 p_j)|z_j] = 0.$$

x_j and the intercept are included in z_j and thus are valid instruments for themselves. If a valid instrument for price exists then $E[p_j|z_j]$ can replace p_j and all parameters are identified.

Once we generalize the model to the non-separable setting the same CMR leads to the moments

$$E[\xi_j|z_j] = E[\delta_j - (c_0 + \beta_0'x_j - \alpha_0 p_j + \xi_j(\gamma_0'x_j + \gamma_{p0}(\bar{y} - p_j)))|z_j] = 0. \tag{4}$$

x_j and p_j can be treated as in the separable case, and since x_j and $\bar{y}$ are in the conditioning set $E[x_j\xi_j|z_j] = x_j E[\xi_j|z_j] = 0$ and $E[\bar{y}\xi_j|z_j] = \bar{y}E[\xi_j|z_j] = 0$. However, p_j is not generally known given z_j, so $E[p_j\xi_j|z_j] \neq p_j E[\xi_j|z_j]$, and the CMR alone fails to identify any of the parameters.

(4) is an example of simple nonseparable setting that illustrates a more general point regarding non-separable errors and the failure of identification using CMRs (see Blundell and Powell (2003) and Hahn and Ridder (2008)). We have valid conditional moment restrictions and our setting is one where we can explicitly solve for ξ for any candidate value of θ. However, these together are *not* be sufficient for identification. One can see this by solving for ξ_j as a function of the other arguments and expressing the CMR as

$$E[\xi_j|z_j] = E\left[\frac{\delta_j - c_0 - \beta_0'x_j + \alpha_0 p_j}{1 + \gamma_0'x_j + \gamma_{p0}(\bar{y} - p_j)}|z_j\right] = 0.$$

These moment conditions are satisfied for multiple values of the parameters (e.g. any $\gamma_{k0} = \infty$)

and thus do not identify the model parameters.

One approach is to add further restrictions that allow the practitioner to calculate and thus control for $E[p_j \xi_j | z_j]$. However, calculating the value of this expectation with ξ_j unknown is virtually impossible without fully specifying how p_j is determined in equilibrium. Researchers may be reluctant to do so because p_j may be a function of all observed and unobserved characteristics of vehicles in the market, in addition to other cost and demand shifters. An advantage of our solution is that we will add controls to the conditioning set z_j such that price will be known, so we avoid the problem of having to resolve this exact relationship between p_j and ξ_j conditional on z_j.

4   Adding Moments with Control Functions

We add new moment conditions to the CMRs to solve this non-uniqueness problem. We develop a control function that has as arguments new controls and z_j which together condition out the correlation between the demand error ξ_j and price. For identification the control function must not have arguments that are perfectly collinear with (x_j, p_j). The CMR conditions from BLP put shape restrictions on the control function that ensure this collinearity does not occur.

A major advantage of our approach is that our moments require nothing beyond the standard conditions for identification with valid instruments. Specifically, just as in Berry (1994) and Berry, Levinsohn, and Pakes (1995) we require no new instruments beyond those from their setup, and we only require - as they do - that the instruments shift price around while being excluded from the utility function.

Each product j may have its own set of controls that we denote $\mathbf{V}_j$. The control function is the conditional expectation of the error given z_j and $\mathbf{V}_j$, which we write as

$$f(z_j, \mathbf{V}_j) = E[\xi_j | z_j, \mathbf{V}_j].$$

It is well-defined and (almost surely) unique as long as the unconditional expectation $E[\xi_j]$ exists.

$\mathbf{V}_j$ must satisfy the next condition in order to address the endogeneity problem.

Condition 1. *(CF) Any bounded function of (z_j, p_j) is uncorrelated with ξ_j given $f(z_j, \mathbf{V}_j)$.*

While $V_j = p_j$ would trivially satisfy this condition, if we include prices in $\mathbf{V}_j$ we will not be identified because the controls will leave no variation to identify α_0. We look for controls $V_j \neq p_j$ such that the control function $f(z_j, \mathbf{V}_j)$ removes the dependence between p_j and ξ_j and leaves some remaining (causal) variation of p_j.

In order to resolve the difficulty associated with $E[p_j \xi_j | z_j] \neq p_j E[\xi_j | z_j]$, we require that p_j is known conditional on $(z_j, \mathbf{V}_j)$, which allows us to write $E[p_j \xi_j | z_j, \mathbf{V}_j] = p_j E[\xi_j | z_j, \mathbf{V}_j]$ and leads to the CF condition being satisfied.

Theorem 1. *If there exists control(s) $\mathbf{V}_j$ such that p_j is known conditional on $(z_j, \mathbf{V}_j)$, then the condition CF is satisfied.*

Proof. For any bounded function of (z_j, p_j), say $h(z_j, p_j)$, we have $E[h(z_j, p_j)(\xi_j - f(z_j, \mathbf{V}_j))] = 0$

due to the law of iterated expectation, because $E[h(z_j, p_j)(\xi_j - f(z_j, \mathbf{V}_j))|z_j, \mathbf{V}_j] = h(z_j, p_j)E[\xi_j - f(z_j, \mathbf{V}_j)|z_j, \mathbf{V}_j] = 0$ because p_j is known given $(z_j, \mathbf{V}_j)$ and $f(z_j, \mathbf{V}_j) = E[\xi_j|z_j, \mathbf{V}_j]$.

$\square$

We propose two variants of controls that both satisfy the CF condition. Here we discuss using

$$V_j = p_j - E[p_j|z_j] = p_j - \Pi(z_j), j = 1, \ldots, J, \tag{5}$$

with $\Pi(z_j) \equiv E[p_j|z_j]$, the expected value of p_j given z_j. In subsection 4.3 we consider an idea proposed in Matzkin (2003) as an alternative way to generate V_j. The controls for good j are then given by $\mathbf{V}_j = g_j(V_1, \ldots, V_J)$, for some known (vector) function $g_j(\cdot)$ of $(V_1, \ldots, V_J)$ chosen by the researcher. $\mathbf{V}_j$ satisfies the CF condition by Theorem 1 as long as V_j is an element of $\mathbf{V}_j$. In the simplest case $\mathbf{V}_j = V_j$, which is sufficient for identification and consistency. However, since $f(z_j, \mathbf{V}_j)$ is a new regressor in our setup, for efficiency purposes one may want to include V_k $k \neq j$ as they may also "explain" ξ_j, leading to more variation in $f(z_j, \mathbf{V}_j)$.

Having determined $\mathbf{V}_j = g_j(V_1, \ldots, V_J)$, we can then exploit the moment condition:

$$0 = E\left[\delta_j - \{c_0 + \beta_0 x_j - \alpha_0 p_j + f(z_j, \mathbf{V}_j)(1 + \gamma_0 x_j + \gamma_{p0}(\bar{y} - p_j))\}|z_j, \mathbf{V}_j\right], \tag{6}$$

where without loss of generality we let x_j be scalar. Letting $\widetilde{\xi}_j = (1 + \gamma x_j + \gamma_p(\bar{y} - p_j))\xi_j$ we now obtain

$$
\begin{aligned}
E[\widetilde{\xi}_j|z, \mathbf{V}_j] &= E[\xi_j|z_j, \mathbf{V}_j] + \gamma E[x_j \xi_j|z_j, \mathbf{V}_j] + \gamma_p E[(\bar{y} - p_j)\xi_j|z_j, \mathbf{V}_j] \\
&= E[\xi_j|z_j, \mathbf{V}_j](1 + \gamma x_j + \gamma_p(\bar{y} - p_j)) \\
&= f(z_j, \mathbf{V}_j)(1 + \gamma x_j + \gamma_p(\bar{y} - p_j)),
\end{aligned}
$$

because $x_j \in z_j$ and p_j is also known conditional on z_j and $\mathbf{V}_j$. The choice of the control function coupled with (6) thus allows us to circumvent the problem of specifying the exact relationship between p_j and ξ_j.

The structural parameters would all be identified from (6) if no linear functional relationship existed between 1, x_j, p_j, $f(z_j, \mathbf{V}_j)$, $f(z_j, \mathbf{V}_j)x_j$, and $f(z_j, \mathbf{V}_j)(\bar{y} - p_j)$. However, $f(z_j, \mathbf{V}_j)$ may contain linear functions of x_j or be collinear with p_j, in which case one will not be able to separate the coefficients (c_0, β_0, α_0) from the function $f(z_j, \mathbf{V}_j)$. We reintroduce the conditional moment restrictions to rule out this possible collinearity.

Condition 2 (CMR). $E[\xi_j|z_j] = 0$.

The CMR condition imposes

$$0 = E[\xi_j|z_j] = E[E[\xi_j|z_j, \mathbf{V}_j]|z_j] = E[f(z_j, \mathbf{V}_j)|z_j].$$

CMR imposes that the mean of $f(z_j, \mathbf{V}_j)$ is equal to zero for any value of z_j. Thus while $f(z_j, \mathbf{V}_j)$

can depend on a function of V_j and its interaction with z_j, it cannot be an additive function of z_j *only*, so functions of x_j only are also ruled out. Also, since $V_j \neq p_j$, as long as z_j includes a variable not included in x_j, $f(z_j, \mathbf{V}_j)$ will not be perfectly collinear with (x_j, p_j) because $f(z_j, \mathbf{V}_j)$ cannot be an additive function of z_j. Thus the generalized control function moments combined with the implied shape restrictions from CMR on $f(z_j, \mathbf{V}_j)$ will suffice for identification of the structural parameters θ_0. Section 4.1 provides a simple example and Section 5 proves identification formally.

Together CF and CMR can be written as a set of moment conditions

$$0 = E\left[\delta_j - \{c_0 + \beta_0' x_j - \alpha_0 p_j + f(z_j, \mathbf{V}_j)(1 + \gamma_0' x_j + \gamma_{p0}(\bar{y} - p_j))\}|z_j, \mathbf{V}_j\right] \tag{7}$$

with $f(z_j, \mathbf{V}_j)$ restricted to satisfy

$$E[f(z_j, \mathbf{V}_j)|z_j] = 0. \tag{8}$$

We use a multi-step least squares estimator based on the moment conditions from (7) and (8) to estimate θ_0 and the nonparametric function $f(z_j, \mathbf{V}_j)$, which we approximate with sieves. In the first-step we obtain consistent estimates of $\mathbf{V}_j = g_j(V_1, \ldots, V_J)$ using a consistent estimator for $\Pi(z_j)$ $j = 1, \ldots, J$ and $V_j = p_j - \Pi(z_j)$. In the second step we construct the approximation of $f(z_j, \mathbf{V}_j)$ that it satisfies (8). For example, we can approximate $f(z_j, \mathbf{V}_j)$ as

$$f(z_j, \mathbf{V}_j) = \sum_{l_1=1}^{\infty} \pi_{l_1,0}(\varphi_{l_1}(\mathbf{V}_j) - E[\varphi_{l_1}(\mathbf{V}_j)|z_j]) + \sum_{l=2}^{\infty} \sum_{l_1 \geq 1, l_2 \geq 1 \text{ s.t. } l_1+l_2=l} \pi_{l_1,l_2} \phi_{l_2}(z_j)(\varphi_{l_1}(\mathbf{V}_j) - E[\varphi_{l_1}(\mathbf{V}_j)|z_j])$$

where $\varphi_{l_1}(\mathbf{V}_j)$ and $\phi_{l_2}(z_j)$ denote approximating functions of $\mathbf{V}_j$ and z_j (e.g., tensor products polynomials or splines), with plug-in consistent estimates of $E[\varphi_{l_1}(\mathbf{V}_j)|z_j]$. In the final step we estimate θ_0 and $f(z_j, \mathbf{V}_j)$ simultaneously using non-linear least squares.[5] Kim and Petrin (2010c) provide conditions for the consistency and the asymptotic normality of the general sieve multi-step estimator.

4.1 Example

While our general approach allows $f(z_j, \mathbf{V}_j)$ to come from any class of functions that can be consistently approximated by sieves, here we consider a simple example to illustrate how the CMR restriction yields identification. For some parameter values $\pi = (\pi_0, \pi_1', \pi_2, \pi_3')'$ we assume $f(z_j, \mathbf{V}_j)$ can be written as

$$f(z_j, \mathbf{V}_j) = \pi_0 + \pi_1' z_j + \pi_2 V_j + \pi_3' z_j V_j$$

---

[5]Alternatively one can estimate the model parameters in two steps using the unconstrained approximation $\tilde{f}(z_j, \mathbf{V}_j) = \sum_{l_1=1}^{\infty} \pi_{l_1,0}\varphi_{l_1}(\mathbf{V}_j) + \sum_{l=2}^{\infty} \sum_{l_1 \geq 1, l_2 \geq 1 \text{ s.t. } l_1+l_2=l} \pi_{l_1,l_2} \phi_{l_2}(z_j)\varphi_{l_1}(\mathbf{V}_j)$. If one wanted an estimate of $f(z_j, \mathbf{V}_j)$ one would use a standard estimator to approximate $E[\tilde{f}(z_j, \mathbf{V}_j)|z_j]$ and then calculate

$$f(z_j, \mathbf{V}_j) = \tilde{f}(z_j, \mathbf{V}_j) - E[\tilde{f}(z_j, \mathbf{V}_j)|z_j].$$

with $z_j = (x_j, z_{2j})'$. Letting $\pi_3' z_j = \pi_{31} x_j + \pi_{32} z_{2j}$ the CMR in this case implies

$$
\begin{aligned}
f(z_j, \mathbf{V}_j) &= f(z_j, \mathbf{V}_j) - E[f(z_j, \mathbf{V}_j)|z_j] &(9) \\
&= (\pi_0 + \pi_1' z_j + \pi_2 V_j + \pi_3' z_j V_j) - (\pi_0 + \pi_1' z_j + \pi_2 E[V_j|z_j] + \pi_3' z_j E[V_j|z_j]) \\
&= \pi_2 V_j + \pi_3' z_j V_j,
\end{aligned}
$$

because $V_j = p_j - \Pi(z_j)$ so $E[V_j|z_j] = 0$. Thus $f(z_j, \mathbf{V}_j)$ is a function of V_j and its interaction with z_j, but conditional on these terms is not an additive function of p_j nor z_j alone.

Identification follows from plugging (9) in (7) and rearranging to obtain

$$
\begin{aligned}
0 &= E[\delta_j - \{c_0 + \beta_0 x_j - \alpha_0 p_j + \pi_2 V_j + (\pi_2 \gamma_0 + \pi_{31}) x_j V_j + \pi_2 \gamma_{p0} V_j (\bar{y} - p_j) \\
&\quad + \pi_{31} \gamma_0 x_j^2 V_j + \pi_{31} \gamma_{p0} x_j V_j (\bar{y} - p_j) + \pi_{32} z_{2j} V_j + \pi_{32} \gamma_0 z_{2j} x_j V_j + \pi_{32} \gamma_{p0} z_{2j} V_j (\bar{y} - p_j)\}|z_j, \mathbf{V}_j].
\end{aligned}
$$

The unconstrained regression of δ_j on $1, x_j, p_j, V_j, x_j V_j, V_j(\bar{y} - p_j), x_j^2 V_j, x_j V_j(\bar{y} - p_j), z_{2j} V_j, z_{2j} x_j V_j$, and $z_{2j} V_j(\bar{y} - p_j)$ then identifies the coefficients $(c_0, \beta_0, \alpha_0, \pi_2)$ and the composite coefficients $(\pi_2 \gamma_0 + \pi_{31}, \pi_2 \gamma_{p0}, \pi_{31} \gamma_0, \pi_{31} \gamma_{p0}, \pi_{32}, \pi_{32} \gamma_0, \pi_{32} \gamma_{p0})$ unless the regressors are "multicollinear". $(\gamma_0, \gamma_{p0}, \pi_{31}, \pi_{32})$ are then identified by the composite coefficients.

4.2 Identification and Higher-order CMRs

If $\gamma_p \neq 0$ then the higher order moments of ξ_j conditional on z_j do not help with identification. The problem is the same as that encountered with the conditional mean, where moment conditions are satisfied for multiple values of the parameters. For example, consider the conditional homoskedasticity assumption where $E[\xi_j^2|z_j] = \sigma^2$. Rewritten we have

$$
E[\xi_j^2|z_j] - \sigma^2 = E\left[\left(\frac{\delta_j - c_0 - \beta_0' x_j + \alpha_0 p_j}{1 + \gamma_0' x_j + \gamma_{p0}(\bar{y} - p_j)}\right)^2 | z_j\right] - \sigma^2 = 0,
$$

which is satisfied for any $\gamma_{k0} = \infty$ and $\sigma = 0$.

If $\gamma_p = 0$ then only exogenous variables interact with the demand error. The conditional moment restrictions $E[\xi_j|z_j] = 0$ are sufficient to identify (c_0, β_0, α_0) because the CMR implies

$$
E[\xi_j + \sum_{k=1}^{K} \gamma_k x_{jk} \xi_j \mid z_j] = E[\delta_j - (c_0 + \beta_0' x_j - \alpha_0 p_j)|z_j] = 0.
$$

Given (c_0, β_0, α_0), the entire multiplicative heteroskedastic error $\tilde{\xi}_j = \xi_j + \sum_{k=1}^{K} \gamma_k x_{jk} \xi_j$ is identified. The $\tilde{\xi}_j$ can be used with a higher-order moment restriction on ξ_j conditional on z_j to identify γ.

We illustrate assuming conditional homoskedasticity holds and (without loss of generality) there is only one exogenous characteristic, so the entire identified error is $\tilde{\xi}_j = \xi_j(1 + \gamma x_j)$. Taking the conditional expectation of this squared error yields

$$
E[\tilde{\xi}_j^2|z_j] = \sigma^2 + 2\sigma^2 \gamma x_j + \sigma^2 \gamma^2 x_j^2.
$$

10

If we consider the regression model

$$\tilde{\xi}_j^2 = \pi_0 + \pi_1 x_j + \pi_2 x_j^2 + \eta_j$$

with $E[\eta_j | z_j] = 0$ by construction, then γ is overidentified because $\gamma^2 = \pi_2/\pi_0$ and $\gamma = \pi_1/2\pi_0$.

4.3 Matzkin (2003) Controls

We can also use the controls proposed in Matzkin (2003), as done in Florens, Heckman, Meghir, and Vytlacil (2008) and Imbens and Newey (2003). Assuming p_j is continuous, we can always rewrite p_j as a function of z_j and a continuous single error term $\tilde{V}_j$ - $p_j = \tilde{h}(z_j, \tilde{V}_j)$ - such that $\tilde{V}_j$ is independent of z_j and $\tilde{h}(z_j, \tilde{V}_j)$ is increasing in $\tilde{V}_j$.[6] Normalizing $\tilde{V}_j$ to be uniform over the unit interval $[0, 1]$ we obtain the new control

$$\tilde{V}_j = F_{p_j|z_j}(p_j|z_j)$$

where $F_{p_j|z_j}$ denotes the conditional cumulative distribution function of p_j given z_j. The control $\tilde{V}_j$ satisfies the requirement in Theorem 1 because conditional on $(z_j, \tilde{V}_j)$, p_j is known, given as $p_j = F_{p_j|z_j}^{-1}(\tilde{V}_j|z_j) \equiv \tilde{h}(z_j, \tilde{V}_j)$. One can then proceed as described above constructing $\tilde{\mathbf{V}}_j = \tilde{g}_j(\tilde{V}_1, \ldots, \tilde{V}_J)$. Identification also holds for $\tilde{\mathbf{V}}_j$ (see Kim and Petrin (2010c) for the latter case).

4.4 Alternative Approaches

We are aware of three other approaches that allow for some form of non-separable demands with endogenous prices in discrete choice settings. Bajari and Benkard (2005) and Kim and Petrin (2010a) use the structure from Imbens and Newey (2009) and place restrictions on demand and supply such that it is possible to invert out from the pricing equations the demand errors. Once the demand errors have been recovered from the inversion, they can enter utility in any non-separable fashion that the practitioner desires because the variable is now observed. The tradeoff is that they require the controls $(V_1, \ldots, V_J)$ to be one-to-one with $\xi = (\xi_1, \ldots, \xi_J)$ conditional on $Z = (z_1, \ldots, z_J)$, and they also need full independence of ξ and Z, two important features of the econometric setup from Imbens and Newey (2009). We require neither assumption but our non-separable setup is not fully general.

In the case where a special type of characteristic exists, Berry and Haile (2010) show how to use it in conjunction with conditional moment restrictions to achieve identification in differentiated products models with market level data. This special characteristic - call it $x_j^{(1)}$ - must be perfectly substitutable with ξ_j, and the coefficient on the special characteristic must be known.[7] The approach allows for non-parametric identification in the variables $((x_j^{(1)} + \xi_j), x_j^{(2)}, p_j)$.

We show how in our parametric setup from (2) identification using the CMRs is achieved when this special characteristic exists. Substituting in the special characteristic to the mean utility we

---

[6]This does not imply that p_j and ξ_j are independent given $\tilde{V}_j$ nor that p_j and ξ_j are independent given $(\tilde{V}_1, \ldots, \tilde{V}_J)$ even if ξ_j is independent of z_j.

[7]This characteristic is related to but not the same as the special regressor from Lewbel (2000).

have

$$\delta_j = c_0 + x_j^{(1)} + \beta_0^{(2)\prime} x_j^{(2)} - \alpha_0 p_j + \xi_j + \gamma_0' x_j^{(2)}(x_j^{(1)} + \xi_j) + \gamma_{p0}(\bar{y} - p_j)(x_j^{(1)} + \xi_j),$$

with the other regressors given as $x_j^{(2)}$ and where for transparency we suppress interactions between $x_j^{(2)}$ and $(\bar{y} - p_j)$. Solving for ξ_j and taking expectations conditional on z_j, we obtain

$$0 = E[\xi_j | z_j] = -x_j^{(1)} + E\left[\frac{\delta_j - c_0 - \beta_0^{(2)\prime} x_j^{(2)} + \alpha_0 p_j}{1 + \gamma_0' x_j^{(2)} + \gamma_{p0}(\bar{y} - p_j)} \Big| z_j\right],$$

so this setup rules out any $\gamma_{k0} = \pm\infty$ unless $x_j^{(1)} = 0$. Note that if we did not know the coefficient on the special characteristic we would have to estimate it and the moment condition would become

$$0 = E[\xi_j | z_j] = -\beta_0^{(1)} x_j^{(1)} + E\left[\frac{\delta_j - c_0 - \beta_0^{(2)\prime} x_j^{(2)} + \alpha_0 p_j}{1 + \gamma_0' x_j^{(2)} + \gamma_{p0}(\bar{y} - p_j)} \Big| z_j\right],$$

which is satisfied for $\beta_0^{(1)} = 0$ and any $\gamma_{k0} = \pm\infty$, leading to failure of identification.

5 Identification

In this section we show global identification for the model with $\mu_{ij} = 0$. In the next section we provide conditions under which our sieve estimator is consistent. In the appendix we provide the consistency proof for the random coefficients setup with $\mu_{ij} \neq 0$.

We study identification using the moment conditions (7) and (8). We use controls $\mathbf{V}_j$ that both satisfy the CF condition and are possibly a function of $(p_j - \Pi(z_j))$ for $j = 1, \ldots, J$. We write this function $\mathbf{V}_j = g_j(p_1 - \Pi(z_1), \ldots, p_J - \Pi(z_J)) = g_j(V_1, \ldots, V_J)$.[8] $\mathbf{V}_j$ is identified from the first step regression of (5), and we treat $\Pi(z_j)$ and $\mathbf{V}_j$ as known throughout the discussion.[9]

If θ_0 and $f_0(z_j, \mathbf{V}_j)$ are identified they must be the unique solution to (7) and (8). $E[\delta_j | z_j, \mathbf{V}_j]$ is unique with probability one, which implies if there exists any other function $\bar{\theta}$ and $\bar{f}(z_j, \mathbf{V}_j)$ that satisfies (7) and (8) it must be that

$$\Pr\{c_0 + \beta_0' x_j - \alpha_0 p_j + f_0(z_j, \mathbf{V}_j)(1 + \gamma_0' x_j + \gamma_{p0}(\bar{y} - p_j)) = \bar{c} + \bar{\beta}' x_j - \bar{\alpha} p_j + \bar{f}(z_j, \mathbf{V}_j)(1 + \bar{\gamma}' x_j + \bar{\gamma}_p(\bar{y} - p_j))\} = 1. \tag{10}$$

Therefore, identification means we must have $\theta_0 = \bar{\theta}$ and $f_0(z_j, \mathbf{V}_j) = \bar{f}(z_j, \mathbf{V}_j)$ with probability one whenever (10) holds.

Our proof uses the unconstrained version of the moment condition (7)

$$0 = E\left[\delta_j - \{c_0 + \beta_0' x_j - \alpha_0 p_j + f_0(z_j, \mathbf{V}_j) + f_x(z_j, \mathbf{V}_j)' x_j + f_p(z_j, \mathbf{V}_j)(\bar{y} - p_j)\} | z_j, \mathbf{V}_j\right]$$

---

[8]It is possible to modify this proof to allow for more general $\mathbf{V}_j$ as defined in Matzkin (2003) (see Kim and Petrin (2010c)).

[9]While we proceed assuming price p_j is endogenous this is not necessary. We can allow for settings where the practitioner does not know whether the variable is exogenous or endogenous (see Kim and Petrin (2010b)).

from which we will show (c_0, β_0, α_0), $f_0(z_j, \mathbf{V}_j)$, $f_x(z_j, \mathbf{V}_j) = \gamma_0 f_0(z_j, \mathbf{V}_j)$, and $f_p(z_j, \mathbf{V}_j) = \gamma_{p0} f_0(z_j, \mathbf{V}_j)$ are identified, and thus so are (γ_0, γ_{p0}). Working with differences $\kappa(z_j, \mathbf{V}_j) = f_0(z_j, \mathbf{V}_j) - \bar{f}(z_j, \mathbf{V}_j)$, $\kappa_x(z_j, \mathbf{V}_j) = f_x(z_j, \mathbf{V}_j) - \bar{f}_x(z_j, \mathbf{V}_j)$, and $\kappa_p(z_j, \mathbf{V}_j) = f_p(z_j, \mathbf{V}_j) - \bar{f}_p(z_j, \mathbf{V}_j)$ we can write (10) as

$$\Pr\{\psi_0 + \psi_1' x_j + \psi_2 p_j + \kappa(z_j, \mathbf{V}_j) + \kappa_x'(z_j, \mathbf{V}_j) x_j + \kappa_p(z_j, \mathbf{V}_j)(\bar{y} - p_j) = 0\} = 1. \qquad (11)$$

If (11) holds, for identification we must have $\psi_0 = 0$, $\psi_1 = 0$, $\psi_2 = 0$, $\kappa(z_j, \mathbf{V}_j) = 0$, $\kappa_x(z_j, \mathbf{V}_j) = 0$, and $\kappa_p(z_j, \mathbf{V}_j) = 0$ with probability one. We formalize this identification statement in Theorem 2.

Theorem 2 (Identification). *Let*

$$\Psi(x_j, p_j, \cdot; \psi) = \psi_0 + \psi_1' x_j + \psi_2 p_j + \kappa(z_j, \mathbf{V}_j) + \kappa_x'(z_j, \mathbf{V}_j) x_j + \kappa_p(z_j, \mathbf{V}_j)(\bar{y} - p_j),$$

and assume the CF condition holds. If (x_j, p_j) and $(z_j, \mathbf{V}_j)$ do not have a functional relationship of the form

$$\Pr\left\{\Psi(x_j, p_j, \kappa(z_j, \mathbf{V}_j), \kappa_x(z_j, \mathbf{V}_j), \kappa_p(z_j, \mathbf{V}_j); \psi) = 0\right\} = 1 \qquad (12)$$

then the structural parameters $\theta_0 = (c_0, \beta_0', \alpha_0, \gamma_0', \gamma_{p0})'$ are identified.

Proof. The CF condition allows one to move from equation (2) to equation (7) (and thus to equation (12)). If there exists an additive functional relationship between $1, x_j, p_j, \kappa(z_j, \mathbf{V}_j), x_{j1}\kappa_{x_1}(z_j, \mathbf{V}_j),$ ..., $x_{jK}\kappa_{x_K}(z_j, \mathbf{V}_j)$, and $(\bar{y} - p_j)\kappa_p(z_j, \mathbf{V}_j)$ then (12) must be satisfied. The contrapositive proves the statement. $\qquad \square$

We now use Theorem 2 to show global identification when $\Pi(z_j)$ and $f(z_j, \mathbf{V}_j)$ are differentiable. Let $z_j = (x_j', z_{2j}')'$ and let subscripts with (2), (p_j), and (v_j) denote partial differentiation with respect to z_{2j}, p_j, and V_j. The proof uses differentiability and completeness (from Newey and Powell (2003)) to show that equation (12) implies $\kappa(z_j, \mathbf{V}_j)$ is only a function of (p_j, x_j). Writing $\kappa(z_j, \mathbf{V}_j) = \tilde{\kappa}(p_j, x_j)$ and using the CMR implies $E[\kappa(z_j, \mathbf{V}_j)|z_j] = E[\tilde{\kappa}(p_j, x_j)|z_j] = 0$. The completeness condition then implies $\kappa(z_j, \mathbf{V}_j) = \tilde{\kappa}(p_j, x_j) = 0$ almost surely. The same logic yields $\kappa_x(z_j, \mathbf{V}_j) = 0$ a.s., $\kappa_p(z_j, \mathbf{V}_j) = 0$ a.s., and $(\psi_0, \psi_1, \psi_2) = 0$.[10]

Theorem 3. *Assume $\Pi(z_j)$ and $f(z_j, \mathbf{V}_j)$ are differentiable and the one-sided derivatives are continuous at the boundary of the support of $(z_j, \mathbf{V}_j)$. Assume the CF and CMR conditions hold. If for all functions $B(p_j, x_j)$ with finite expectation, $E[B(p_j, x_j)|z_j] = 0$ a.s. implies $B(p_j, x_j) = 0$ a.s. (completeness condition), then θ_0 is identified.*

Proof. Given differentiability, it suffices to consider differentiable $\kappa_.(z_j, \mathbf{V}_j)$ where $\kappa_.(z_j, \mathbf{V}_j)$ is generic notation for $\kappa(z_j, \mathbf{V}_j), \kappa_x(z_j, \mathbf{V}_j),$ or $\kappa_p(z_j, \mathbf{V}_j)$. Taking derivatives of $\Psi(x_j, p_j, \cdot; \psi)$, we

---

[10]We also maintain that the one-sided derivatives of $\Psi(x_j, p_j, \cdot; \psi)$ are continuous at the boundary of the support of $(z_j, \mathbf{V}_j)$, although instead one may alternatively assume that the boundary of the support of $(z_j, \mathbf{V}_j)$ has zero probability (this may require a trimming device to deal with the boundary of the support in the estimation).

obtain

$$
\begin{aligned}
\Psi_{(2)}(\cdot;\psi) &= \Pi_{(2)}(z_j)(\psi_2 - \kappa_p(z_j,\mathbf{V}_j)) + \kappa_{(2)}(z_j,\mathbf{V}_j) + \kappa'_{x,(2)}(z_j,\mathbf{V}_j)x_j + \kappa_{p,(2)}(z_j,\mathbf{V}_j)(\bar{y}-p_j) \\
\Psi_{(v_j)}(\cdot;\psi) &= \psi_2 - \kappa_p(z_j,\mathbf{V}_j) + \kappa_{(v_j)}(z_j,\mathbf{V}_j) + \kappa'_{x,(v_j)}(z_j,\mathbf{V}_j)x_j + \kappa_{p,(v_j)}(z_j,\mathbf{V}_j)(\bar{y}-p_j).
\end{aligned}
$$

If there is an additive functional relationship then $\Psi(\cdot;\psi)=0$ with probability one, and also

$$
\Psi_{(2)}(\cdot;\psi)=0 \text{ and } \Psi_{(v_j)}(\cdot;\psi)=0.
$$

Premultiplying $\Psi_{(v_j)}(\cdot;\psi)$ by $\Pi_{(2)}(z_j)$ and combining $\Psi_{(2)}(\cdot;\psi)=0$ and $\Pi_{(2)}(z_j)\times\Psi_{(v_j)}(\cdot;\psi)=0$, we obtain

$$
\begin{aligned}
&\Pi_{(2)}(z_j)\{\kappa_{(v_j)}(z_j,\mathbf{V}_j) + \kappa'_{x,(v_j)}(z_j,\mathbf{V}_j)x_j + \kappa_{p,(v_j)}(z_j,\mathbf{V}_j)(\bar{y}-p_j) - \kappa_p(z_j,\mathbf{V}_j)\} \quad (13)\\
&= \kappa_{(2)}(z_j,\mathbf{V}_j) + \kappa'_{x,(2)}(z_j,\mathbf{V}_j)x_j + \kappa_{p,(2)}(z_j,\mathbf{V}_j)(\bar{y}-p_j) - \Pi_{(2)}(z_j)\kappa_p(z_j,\mathbf{V}_j).
\end{aligned}
$$

Letting

$$
\eta(\cdot) = \kappa(z_j,\mathbf{V}_j) + \kappa'_x(z_j,\mathbf{V}_j)x_j + \kappa_p(z_j,\mathbf{V}_j)(\bar{y}-p_j)
$$

we have $\frac{\partial\eta(\cdot)}{\partial z_{2j}}$ equal to the right hand side of equation (13). $\mathbf{V}_j = g_j(p_1 - \Pi(z_1),\ldots,p_J - \Pi(z_J))$ and $\frac{\partial V_j}{\partial p_j}=1$ imply

$$
\kappa_{\cdot,(v_j)}(z_j,\mathbf{V}_j) = \kappa_{\cdot,(p_j)}(z_j,\mathbf{V}_j),
$$

Using (13) it follows that

$$
\begin{aligned}
\frac{\partial\eta(\cdot)}{\partial z_{2j}} &= \Pi_{(2)}(z_j)\{\kappa_{(p_j)}(z_j,\mathbf{V}_j) + \kappa'_{x,(p_j)}(z_j,\mathbf{V}_j)x_j + \kappa_{p,(p_j)}(z_j,\mathbf{V}_j)(\bar{y}-p_j) - \kappa_p(z_j,\mathbf{V}_j)\} \\
&= \frac{\partial p_j}{\partial z_{2j}}\frac{\partial\eta(\cdot)}{\partial p_j} = \frac{\partial\eta(\cdot)}{\partial p_j}\frac{\partial p_j}{\partial z_{2j}}
\end{aligned}
$$

implying (a) z_{2j} affects $\eta(\cdot)$ and thus $\kappa_{\cdot}(z_j,\mathbf{V}_j)$ only through p_j. We also know that for all $j'\neq j$, $\kappa_{\cdot}(z_j,\mathbf{V}_j)$ is not a function of $z_{2j'}$, so

$$
\partial\kappa_{\cdot}(z_j,\mathbf{V}_j)/\partial z_{2j'} = \Pi_{(2)}(z_{j'})\kappa_{\cdot,(v_{j'})}(z_j,\mathbf{V}_j) = 0.
$$

Completeness implies the full rank of $\Pi_{(2)}(z_{j'})$, which then implies that $\kappa_{\cdot,(v_{j'})}(z_j,\mathbf{V}_j)=0$ for all $j'\neq j$ so (b) $\kappa_{\cdot}(z_j,\mathbf{V}_j)$ is not a function of $V_{j'}$ for $j'\neq j$. Combining these two findings (a) and (b), we conclude

$$
\begin{aligned}
\kappa_{\cdot}(z_j,\mathbf{V}_j) &= \kappa_{\cdot}(z_j,V_j) = \kappa_{\cdot}(z_j,p_j-\Pi(z_j)) \\
&= \tilde{\kappa}_{\cdot}(p_j,z_j) = \tilde{\kappa}_{\cdot}(p_j,x_j)
\end{aligned}
$$

for some function $\tilde{\kappa}_{\cdot}(\cdot)$, where the first equality holds by (b) and the last equality holds by (a).

The CMR condition then implies $E[\kappa_{\cdot}(z_j,\mathbf{V}_j)|z_j] = E[\tilde{\kappa}_{\cdot}(p_j,x_j)|z_j]=0$ and from the complete-

14

ness condition, it follows that $0 = \tilde{\kappa}.(p_j, x_j) = \kappa.(z_j, \mathbf{V}_j)$ with probability one. $\eta(\cdot) = 0$ coupled with $\Psi(\cdot; \psi) = 0$ implies $\psi_0 + \psi_1' x_j + \psi_2 p_j = 0$. In the special case where we condition on z_j, we obtain $E[\psi_0 + \psi_1' x_j + \psi_2 p_j | z_j] = 0$. Full rank of $\Pi_{(2)}(z_j)$ implies $\psi_0 = 0$, $\psi_1 = 0$, and $\psi_2 = 0$.[11] Therefore, $\psi_0 = 0$, $\psi_1 = 0$, $\psi_2 = 0$, $\kappa(z_j, \mathbf{V}_j) = 0, \kappa_x(z_j, \mathbf{V}_j) = 0$, and $\kappa_p(z_j, \mathbf{V}_j) = 0$ with probability one. This completes the proof.

$\square$

A simple example for a non-linear parametric setup may help to illustrate the mechanism of identification. Consider the model

$$\delta_j = c_0 - \alpha_0 p_j + \xi_j + \gamma_{p0} \xi_j (\bar{y} - p_j)$$

with $E[\xi_j | z_j, \mathbf{V}_j] = \pi_0 V_j$. Then one can use the unconstrained moment given by $E[\delta_j - (c_0 - \alpha_0 p_j + \pi_0 V_j + \gamma_{p0} \pi_0 V_j (\bar{y} - p_j)) | z_j, \mathbf{V}_j] = 0$ for identification. Specifically, $c_0, \alpha_0, \varrho_0 \equiv \gamma_{p0} \pi_0$, and π_0 are globally identified from unconstrained least squares by the uniqueness of the conditional expectation as long as $1, p_j, V_j$, and $V_j(\bar{y} - p_j)$ do not have a linear relationship. γ_{p0} is also identified from $\gamma_{p0} = \varrho_0/\pi_0$ as long as $\pi_0 \neq 0$. The theorem shows that for this non-linear model the global identification condition coincides with the local identification condition for parametric non-linear models given in Rothenberg (1971), which says the vector of derivatives with respect to parameters must not be collinear.[12]

We can also generalize the identification result to more flexible specifications of utility. Consider mean utility

$$\delta_j = h_1(x_j, p_j) + \xi_j(1 + h_2(x_j, \bar{y} - p_j))$$

where $h_1(x_j, p_j)$ and $h_2(x_j, \bar{y} - p_j)$ are nonparametric, and we normalize $h_2(\tilde{x}_j, \bar{y} - \tilde{p}_j) = c$ for some known constant c at some $(\tilde{x}_j, \tilde{p}_j)$.

Theorem 4 (Nonparametric Identification with CMR). *Assume $\Pi(z_j)$, $h_1(x_j, p_j)$, $h_2(x_j, \bar{y} - p_j)$, and $f(z_j, \mathbf{V}_j)$ are differentiable and the one-sided derivatives are continuous at the boundary of the support of $(z_j, \mathbf{V}_j)$. Assume the CF and CMR conditions hold. If for all functions $B(p_j, x_j)$ with finite expectation, $E[B(p_j, x_j) | z_j] = 0$ a.s. implies $B(p_j, x_j) = 0$ a.s., then $(h_{10}(x_j, p_j), h_{20}(x_j, \bar{y} - p_j))$ is identified up to a normalization of $h_2(\tilde{x}_j, \bar{y} - \tilde{p}_j) = c$ for some known constant c at some $(\tilde{x}_j, \tilde{p}_j)$.*

The proof strategy is essentially the same as the parametric case (see Kim and Petrin (2010c)).

6 Sieve Estimation and Consistency

We show consistency of our multi-step sieve estimator for the case when $\mu_{ij} = 0$ (see appendix

---

[11]The full rank of $\Pi_{(2)}(z_j)$ is equivalent to $\Pi_{(2)}(z_j) \neq 0$ since p_j is scalar.

[12]In this example the vector of first derivatives w.r.t. $(c_0, \alpha_0, \gamma_{p0}, \pi_0)'$ is given by $(1, -p_j, \pi_0 V_j(\bar{y} - p_j), \gamma_{p0} V_j(\bar{y} - p_j) + V_j)'$, and they are not collinear as long as $\pi_0 \neq 0$ and $1, p_j, V_j$, and $V_j(\bar{y} - p_j)$ do not have a linear relationship.

for $\mu_{ij} \neq 0$). In the first stage we estimate $\Pi(z_j)$ and obtain $\hat{V}_j = p_j - \hat{\Pi}(z_j)$ for $j = 1, \ldots, J$ and construct $\hat{\mathbf{V}}_j = g_j(\hat{V}_1, \ldots, \hat{V}_J)$. In the second step, we construct approximating basis functions using $\hat{\mathbf{V}}_j$ and z_j, where we subtract out conditional means of underlying basis functions (conditional on z_j) to approximate $f(\cdot)$ that satisfies (8). In the final step we estimate θ_0 and $f_0(\cdot)$ using a sieve method.

Let $\mathcal{F}$ denote a space of functions that includes the true function f_0, endowed with $\|\cdot\|_{\mathcal{F}}$ a pseudo-metric on $\mathcal{F}$. We first write the infeasible basis functions (and we replace them with their estimates below) for $f(\cdot)$ when $\mathbf{V}_j$ is known as

$$\tilde{\varphi}_l(\mathbf{V}_j, z_j) = \varphi_l(\mathbf{V}_j, z_j) - \bar{\varphi}_l(z_j)$$

where $\bar{\varphi}_l(z_j) = E[\varphi_l(\mathbf{V}_j, z_j)|z_j]$ and $\{\varphi_l(\mathbf{V}_j, z_j), \ l = 1, 2, \ldots\}$ denotes a sequence of approximating basis functions of $(\mathbf{V}_j, z_j)$ such as power series or splines. By subtracting out the conditional means from the underlying basis functions, we let any function $f(\cdot)$ in the sieve space defined below satisfy the requirement (8).

Let $J = \sum_{m=1}^{M} J_m$ be the sample size where J_m denotes the number products in market m and M denotes the number of markets and define the (infeasible) sieve space $\mathcal{F}_J$ as the collection of functions

$$\mathcal{F}_J = \{f : f = \sum_{l \leq L(J)} a_l \tilde{\varphi}_l(\mathbf{V}_j, z_j), \|f\|_{\mathcal{F}} < \bar{C}\}$$

for some bounded positive constant $\bar{C}$ and coefficients $(a_1, \ldots, a_{L(J)})$, with $L(J) \to \infty$ and $L(J)/J \to 0$ such that $\mathcal{F}_J \subseteq \mathcal{F}_{J+1} \subseteq \ldots \subseteq \mathcal{F}$, so we use more flexible approximations as the sample size grows.

We then replace the sequence of the infeasible basis functions $\tilde{\varphi}_l(\mathbf{V}_j, z_j)$ with their estimates as $\hat{\tilde{\varphi}}_l(\hat{\mathbf{V}}_j, z_j) = \varphi_l(\hat{\mathbf{V}}_j, z_j) - \hat{\bar{\varphi}}_l(z_j)$. We then define the sieve space constructed using the estimated basis functions as

$$\hat{\mathcal{F}}_J = \{f : f = \sum_{l \leq L(J)} a_l \hat{\tilde{\varphi}}_l(\cdot, \cdot), \|f\|_{\mathcal{F}} < \bar{C}\}. \tag{14}$$

Under weak regularity conditions $\hat{\mathcal{F}}_J \to \mathcal{F}_J$ (in the Hausdorff metric defined on the metric space $(\mathcal{F}, \|\cdot\|_{\mathcal{F}})$) as $\hat{\Pi}(\cdot) \to \Pi(\cdot)$ and $\hat{\bar{\varphi}}_l(\cdot) \to \bar{\varphi}_l(\cdot)$ (in a pseudo-metric $\|\cdot\|_s$).

Denote a sample criterion function $Q_J(\delta, z, p, \hat{\mathbf{V}}; \theta, f)$ for estimation based on the moment condition of (7). If we use nonlinear sieve least squares estimation, then the sample criterion function becomes

$$Q_J(\delta, z, p, \hat{\mathbf{V}}; \theta, f) = \frac{1}{J} \sum_{m=1}^{M} \sum_{j=1}^{J_m} \{\delta_{mj} - (c + \beta' x_{mj} - \alpha p_{mj} + f(\cdot)(1 + \gamma' x_{mj} + \gamma_p(\bar{y}_m - p_{mj})))\}^2$$

subject to $(\theta, f) \in \Theta \times \hat{\mathcal{F}}_J$. We define the corresponding population criterion function as

$$Q_J^0(\delta, z, p, \mathbf{V}; \theta, f) = \frac{1}{J} \sum_{m=1}^{M} \sum_{j=1}^{J_m} E[\{\delta_{mj} - (c + \beta' x_{mj} - \alpha p_{mj} + f(z_j, \mathbf{V}_j)(1 + \gamma' x_{mj} + \gamma_p(\bar{y}_m - p_{mj})))\}^2].$$

16

Note that below we do not require $Q_J^0(\delta, z, p, \mathbf{V}; \theta, f)$ converges when $J \to \infty$.

Then we obtain our estimator as

$$(\hat{\theta}, \hat{f}) = \mathrm{arginf}_{(\theta,f) \in \Theta \times \hat{\mathcal{F}}_J} Q_J(\delta, z, p, \hat{\mathbf{V}}; \theta, f). \tag{15}$$

We derive the consistency of our estimator under the following assumptions based on the results in Newey and Powell (2003), Chen, Linton, and van Keilegom (2003), and Chen (2006).[13] Here we abstract from the sampling error in the market shares although we allow for it in the proof of consistency with random coefficients.[14] The following assumptions are commonly imposed and standard in the sieve estimation literature, so we minimize our discussion.

We first assume identification (see Section 5):

Assumption 1 (A1). *$(\theta_0, f_0) \in \Theta \times \mathcal{F}$ is the only $(\theta, f) \in \Theta \times \mathcal{F}$[15] satisfying the moment condition (7) and (8) and $Q_J^0(\delta, z, p, \mathbf{V}; \theta_0, f_0) < \infty$.*

Next we assume that our extremum estimator solves (15).

Assumption 2 (A2). *$Q_J(\delta, z, p, \hat{\mathbf{V}}; \hat{\theta}, \hat{f}) \leq inf_{(\theta,f) \in \Theta \times \hat{\mathcal{F}}_J} Q_J(\delta, z, p, \hat{\mathbf{V}}; \theta, f) + o_p(1)$*

Denote the true functions of $\Pi(\cdot)$ and $\bar{\varphi}_l(\cdot)$ as $\Pi_0(\cdot)$ and $\bar{\varphi}_{0l}(\cdot)$, respectively, and assume $\Pi(\cdot)$ and $\bar{\varphi}_l(\cdot)$ are endowed with a pseudo-metric $\|\cdot\|_s$. Assumption A3 says that both $\Pi_0(\cdot)$ and $\bar{\varphi}_{0l}(\cdot)$ can be approximated by the first stage and the middle stage series approximations. For example, this is known to be satisfied for power series and splines approximation if $\Pi_0(\cdot)$'s and $\bar{\varphi}_{0l}(\cdot)$'s are smooth and their derivatives are bounded (e.g., belong to a Hölder class of functions).

Assumption 3 (A3). *$\left\|\hat{\Pi}(\cdot) - \Pi_0(\cdot)\right\|_s = o_p(1)$ and $\left\|\hat{\bar{\varphi}}_l(\cdot) - \bar{\varphi}_{0l}(\cdot)\right\|_s = o_p(1)$ for all l.*

Assumption 4 (A4). *The sieve space $\mathcal{F}_J$ satisfies $\mathcal{F}_J \subseteq \mathcal{F}_{J+1} \subseteq \ldots \subseteq \mathcal{F}$ for all $J \geq 1$; and for any $f \in \mathcal{F}$ there exists $\pi_J f \in \mathcal{F}_J$ such that $\|f - \pi_J f\|_{\mathcal{F}} \to 0$ as $J \to \infty$.*

We maintain the following continuity conditions, which are easy to show for our objective function.

Assumption 5 (A5). *$Q_J^0(\delta, z, p, \mathbf{V}; \theta, f)$ is continuous in $(\theta, f) \in \Theta \times \mathcal{F}$.*

Assumption 6 (A6). *$Q_J^0(\delta, z, p, \mathbf{V}; \theta, f_J)$ is continuous in $\Pi(\cdot)$ and $\bar{\varphi}_l(\cdot)$ uniformly for all $(\theta, f_J) \in \Theta \times \mathcal{F}_J$.*

Next we impose compactness on the sieve space.

---

[13]Our problem is different from Newey and Powell (2003)'s Theorem 4.1 because we use estimated regressors (functions, $\hat{\Pi}(\cdot)$ and $\hat{\bar{\varphi}}_l(\cdot)$) in the main estimation. Our problem is also different from Chen, Linton, and van Keilegom (2003) because we estimate the parametric component (θ_0) and the nonparametric component (f_0) simultaneously in the main estimation.

[14]See Berry, Levinsohn, and Pakes (1995) and Berry, Linton, and Pakes (2004) for explicit treatments of this sampling error. The contribution of this sampling error to the variance of the estimator will be negligible when the market size is large.

[15]The parameter space does not need to be a product space. We use " $\cdot \times \cdot$" for ease of notation throughout the paper.

Assumption 7 (A7). *The parameter space Θ is compact and the sieve space, $\mathcal{F}_J$, is compact under the pseudo-metric $||\cdot||_{\mathcal{F}}$.*

A sufficient condition for compactness is that the sieve space be based on power series or splines.

The last condition we add is that in the neighborhoods of $\Pi_0(\cdot)$ and $\bar{\varphi}_{0l}(\cdot)$, the difference between the sample criterion function and the population criterion function is small enough when J is large. [16]

Assumption 8 (A8). *For all positive sequences $\epsilon_J = o(1)$, we have*

$$\sup_{(\theta,f)\in\Theta\times\mathcal{F}_J,||\Pi-\Pi_0||_s\leq\epsilon_J,||\bar{\varphi}_l-\bar{\varphi}_{0l}||_s\leq\epsilon_J\forall l} \left|Q_J(\delta,z,p,\mathbf{V};\theta,f) - Q_J^0(\delta,z,p,\mathbf{V};\theta,f)\right| = o_p(1)$$

where $\mathbf{V}_{mj} = g_j(p_{m1} - \Pi(z_{m1}),\ldots,p_{mJ_m} - \Pi(z_{mJ_m}))$.

Theorem 5. *Suppose Assumptions A1-A8 are satisfied. Then $\hat{\theta} \rightarrow_p \theta_0$.*

See appendix for the proof. Kim and Petrin (2010c) also develop the asymptotic distribution of $\hat{\theta}$ in the context of the sieve estimation where both $\Pi(\cdot)$ and $\bar{\varphi}_l(\cdot)$ are nonparametrically estimated. The formulas for standard errors of $\hat{\theta}$ developed there can be also used for the parametric estimation case when the truncated semiparametric model (i.e., with fixed length of sieves) is the true parametric model.

7   Monte Carlo Evidence

We demonstrate our estimator's performance using Monte Carlo studies on simple demand/pricing models. We first consider the following demand function (i.e., mean utility of one inside good) where the endogenous price p interacts with the unobserved demand shock ξ:

$$q = c - \alpha p + \gamma p\xi + \xi.$$

Before turning to a single product monopolist setting we consider two reduced form pricing equations

$$[1]\ p\ =\ 2 + Z + (5 + Z^2 + 5Z)\xi + \varsigma$$
$$[2]\ p\ =\ Z + (5 + 5Z + \varsigma)\xi.$$

Here the instrument Z is an observed supply shifter and ς is an unobserved cost shock. In the first design [1], the instrument and the demand error are not additively separable. In the second design [2] the demand error is not additively separable from the instrument nor the supply-side error.

---

[16]Note that Assumption A8 can be easily satisfied by applying a proper law of large numbers (e.g., Chebychev's weak LLN). Define $\bar{\mathbf{W}}_J = \frac{1}{J}\sum_{m=1}^M\sum_{j=1}^{J_m}\mathbf{W}_{mj}$ and $\bar{\mu}_J^{\mathbf{W}} = \frac{1}{J}\sum_{m=1}^M\sum_{j=1}^{J_m}E[\mathbf{W}_{mj}]$ for a random vector $\mathbf{W}_{mj}$. Then it is not difficult to see that Assumption A8 holds if $||\bar{\mathbf{W}}_J - \bar{\mu}_J^{\mathbf{W}}|| = o_p(1)$ with $\mathbf{W}_{mj} = vec(\mathbf{w}_{mj}\mathbf{w}'_{mj})$ and $\mathbf{w}_{mj} = (\delta_{mj},1,x'_{mj},p_{mj},f(z_{mj},\mathbf{V}_{mj}),f(z_{mj},\mathbf{V}_{mj})x'_{mj},f(z_{mj},\mathbf{V}_{mj})(\bar{y}_m - p_{mj}))'$ for all $f \in \mathcal{F}_J$ such that $||\Pi - \Pi_0||_s \leq \epsilon_J$ and $||\bar{\varphi}_l - \bar{\varphi}_{0l}||_s \leq \epsilon_J$.

We generate a simulation data based on these designs with the following distributions: $\xi \sim U_{[-1/2,1/2]}$, $\varsigma \sim U_{[-1/2,1/2]}$, $Z = 2 + 2U_{[-1/2,1/2]}$, and they are independent where $U_{[-1/2,1/2]}$ denotes the uniform distribution supported on $[-1/2, 1/2]$. Note that in these designs, the control $V = p - E[p|Z]$ is not independent of Z. We set the true parameter values $(c_0, \alpha_0, \gamma_0) = (1, 1, 0.5)$. The data is generated with the sample sizes: $M = 1,000$ and $M = 10,000$. We take one reasonable sample size and one large sample size because we are interested both in a finite sample performance and the consistency of our proposed estimator.

In our third design we consider a single product monopolistic pricing model with a demand function (i.e., mean utility in the logit demand)

$$q(X, p, \xi; c, \beta, \alpha, \gamma) = \ln s - \ln(1 - s) = c + \beta X - \alpha p + \gamma p\xi + \xi \text{ and}$$

$$p = \text{argmax}_p \ (p - mc) \frac{\exp(q(X, p, \xi; c, \beta, \alpha, \gamma))}{1 + \exp(q(X, p, \xi; c, \beta, \alpha, \gamma))}$$

where s is the share of the inside good, X is an observed demand shifter, and we let the marginal cost be $mc = 2 + 0.5Z_2 + (2 + 2Z_2)\varsigma$. In this design we draw a demand shock $\xi \sim U_{[-1/2,1/2]}$, a supply-side shock $\varsigma \sim \xi + U_{[-1/2,1/2]}$, $X = U_{[-1/2,1/2]}$, and an observed supply shifter $Z_2 = X + 2 + 2U_{[-1/2,1/2]}$. We set the true parameter values $(c_0, \beta_0, \alpha_0, \gamma_0) = (-2, 1, 1, 0.5)$. The data is generated with the sample sizes: $M = 2,000$ and $M = 10,000$. We let $Z = (X, Z_2)'$.

We estimate the models using three methods: OLS, 2SLS, and our estimator (CMRCF). Our estimator is implemented in three steps. First we estimate $\hat{V} = p - (\hat{\pi}_0 + \hat{\pi}_1' Z + \hat{\pi}_2' Z^2 + \hat{\pi}_3' Z^3)$ using OLS and construct approximating functions $\tilde{V}_1 = \hat{V}$, $\tilde{V}_2 = \hat{V}^2 - \hat{E}[\hat{V}^2|Z]$, and others are defined similarly where $\hat{E}[\cdot|Z]$ is implemented by the OLS estimation on $(1, Z, Z^2, Z^3)$.[17] In the last step we estimate the model parameters using nonlinear least squares:

$$(\hat{c}, \hat{\beta}, \hat{\alpha}, \hat{\gamma}, \hat{a}) = \text{argmin} \sum_{m=1}^{M} \{q_m - (c + \beta X_m - \alpha p_m + \gamma p_m (\sum_{l=1}^{L_M} a_l \tilde{V}_{ml}) + \sum_{l=1}^{L_M} a_l \tilde{V}_{ml})\}^2 / M$$

where we let $\beta = 0$ in designs [1] and [2].

For the design [1] we use the controls $(\tilde{V}_1, Z\tilde{V}_1, Z^2\tilde{V}_1)$ when $M = 1,000$ and use $(\tilde{V}_1, Z\tilde{V}_1, Z^2\tilde{V}_1, Z^3\tilde{V}_1, \tilde{V}_2)$ when $M = 10,000$. For the design [2] we use $(\tilde{V}_1, Z\tilde{V}_1, Z^2\tilde{V}_1)$ with $M = 1,000$ and use $(\tilde{V}_1, Z\tilde{V}_1, Z^2\tilde{V}_1, \tilde{V}_2)$ with $M = 10,000$. Finally we use $(\tilde{V}_1, Z\tilde{V}_1, Z^2\tilde{V}_1)$ for the design [3] with both sample sizes.[18]

We report the biases and the RMSE based on 100 repetitions of the estimations: OLS, 2SLS, and our estimator. The simulation results (Tables I-III) clearly show that OLS is biased in all designs. 2SLS is also biased. Our estimator is robust regardless of different designs for the price.

In the designs [1]-[3], 2SLS estimates for the constant term (c) are biased (-69%, 21%, -18% respectively). In the designs [1]-[3] the 2SLS estimates for the coefficient on the price (α) are severely biased (38%, 21%, and -16%). The 2SLS estimates for the coefficient on the exogenous

---

[17]In the third design $Z^l = (X^l, Z_2^l)'$ for $l = 2, 3$ with abuse of notation.

[18]One can choose an *optimal* set of controls among alternatives based on the cross validation (CV) criterion, although the validity of CV may be compromised due to the presence of the first and the second step in our estimation.

demand shifter (β) in the design [3] seem not biased.

From other Monte Carlos (not reported here) we find higher coefficients on ξ in the pricing equation create larger biases for the 2SLS estimates of c and higher coefficients on the interaction term $Z\xi$ in the pricing equation generate larger biases for the 2SLS estimates of α.

Table I: Design [1], $c_0 = 1, \alpha_0 = 1, \gamma_0 = 0.5$, **Controls:** $\tilde{V}_1, Z\tilde{V}_1, Z^2\tilde{V}_1, Z^3\tilde{V}_1, \tilde{V}_2$

		mean	bias	RMSE	mean	bias	RMSE
			$M = 1,000$			$M = 10,000$	
OLS	c	1.2081	0.2081	0.2119	1.2037	0.2037	0.2040
	α	1.1501	0.1501	0.1503	1.1506	0.1506	0.1506
2SLS	c	0.3584	-0.6416	0.7461	0.3054	-0.6946	0.7007
	α	1.3634	0.3634	0.3765	1.3752	0.3752	0.3760
CMRCF	c	1.0118	0.0118	0.0523	1.0024	0.0024	0.0245
	α	0.9982	-0.0018	0.0132	0.9998	-0.0002	0.0058
	γ	0.5063	0.0063	0.1276	0.4975	-0.0025	0.0549

Table II: Design [2], $c_0 = 1, \alpha_0 = 1, \gamma_0 = 0.5$, **Controls:** $\tilde{V}_1, Z\tilde{V}_1, Z^2\tilde{V}_1, \tilde{V}_2$

		mean	bias	RMSE	mean	bias	RMSE
			$M = 1,000$			$M = 10,000$	
OLS	c	1.3596	0.3596	0.3603	1.3580	0.3580	0.3581
	α	1.1333	0.1333	0.1335	1.1337	0.1337	0.1337
2SLS	c	1.2038	0.2038	0.2195	1.2062	0.2062	0.2072
	α	1.2117	0.2117	0.2163	1.2096	0.2096	0.2099
CMRCF	c	1.0097	0.0097	0.0388	1.0018	0.0018	0.0162
	α	0.9960	-0.0040	0.0202	0.9995	-0.0005	0.0083
	γ	0.5089	0.0089	0.1499	0.5014	0.0014	0.0626

Table III: Design [3], $c_0 = -2, \beta_0 = 1, \alpha_0 = 1, \gamma_0 = 0.5$, **Controls:** $\tilde{V}_1, Z\tilde{V}_1, Z^2\tilde{V}_1$

		mean	bias	RMSE	mean	bias	RMSE
			$M = 2,000$			$M = 10,000$	
OLS	c	-2.7465	-0.7465	0.7469	-2.7484	-0.7484	0.7485
	β	0.9438	-0.0562	0.0777	0.9453	-0.0547	0.0587
	α	0.7496	-0.2504	0.2505	0.7487	-0.2513	0.2513
2SLS	c	-2.2934	-0.2934	0.3561	-2.3637	-0.3637	0.3778
	β	1.0007	0.0007	0.0673	0.9955	-0.0045	0.0274
	α	0.8617	-0.1383	0.1470	0.8437	-0.1563	0.1583
CMRCF	c	-1.9316	0.0684	0.2472	-2.0092	-0.0092	0.0978
	β	1.0048	0.0048	0.0735	1.0024	0.0024	0.0261
	α	1.0143	0.0143	0.0600	0.9942	-0.0058	0.0245
	γ	0.4929	-0.0071	0.2274	0.5067	0.0067	0.1464

8 Non-separability in the BLP Automobile Data

We revisit the original Berry, Levinsohn, and Pakes (1995) automobile data to investigate whether interaction terms are important for own- and cross-price elasticities. There are 2217 market-level observations on prices, quantities, and characteristics of automobiles sold in the 20 U.S. automobile markets indexed m beginning in 1971 and continuing annually to 1990. We let J_m denote the number products in market m and include the same characteristics: horsepower-to-weight, interior space, a/c standard, and miles per dollar. We do not use a supply side model when we estimate the demand side model so our point estimates only exactly match their estimated specifications for the cases they examine without the supply side.[19]

We decompose utility into three components as in equation (1), with the utility common to all consumers δ_{mj} given as

$$\delta_{mj} = c + \beta' x_{mj} - \alpha p_{mj} + \xi_{mj} + \sum_{k=1}^{4} \gamma_k x_{mjk} \xi_{mj} + \gamma_p (\bar{y}_m - p_{mj}) \xi_{mj}.$$

When $(\gamma_1, \gamma_2, \gamma_3, \gamma_4, \gamma_p) \neq 0$ either characteristics or price are not separable from the demand error. We parameterize $\mu_{ij}(\sigma)$ as

$$\mu_{ij} = \sigma_c \nu_{ic} + \sum_{k=1}^{4} \sigma_k \nu_{ik} x_{jk}$$

with $\nu_i = (\nu_{ic}, \nu_{i1}, \ldots, \nu_{i4})$ mean-zero standard normal and $\sigma = (\sigma_c, \sigma_1, \ldots, \sigma_4)$ the standard deviation parameters associated with the taste shocks. The induced vector of tastes for each car j for consumer i is given as $\mu_i(\sigma) = (\mu_{i1}(\sigma), \ldots, \mu_{iJ}(\sigma))$ with density $f(\mu_i(\sigma))$. Letting $\delta_m = (\delta_{m1}, \ldots, \delta_{mJ_m})$ the market share of product j is then

$$s_{mj}(\delta_m) = \int \frac{e^{\delta_{mj} + \mu_{ij}}}{\sum_{k=0}^{J_m} e^{\delta_{mk} + \mu_{ik}}} f(\mu) d\mu,$$

and we approximate this integral with standard simulation techniques.

8.1 Controls

We use the mean projection residuals for price as the starting point for controls. Following Berry, Levinsohn, and Pakes (1995) we assume all observed product characteristics are exogenous and denote these variables for market m as Z_m. The mean projection residual is given as an estimate of

$$\tilde{\xi}_{mj} = p_{mj} - E[p_{mj} \mid Z_m].$$

There are many instruments so we follow Berry, Levinsohn, and Pakes (1995) and Pakes (1996), reducing this set to 15 instruments for each good j that we denote $\tilde{z}_{mj}$. These instruments include

---

[19]We focus on the demand side for three reasons: it makes the comparison more transparent, most researchers do not impose a supply side model when estimating demands, and the results are easier to replicate.

j's product characteristics, the sum of each of the product characteristics across all goods in market m produced by the same firm producing j, and the sum in market m of each of the product characteristics across all other firms not producing j. Our first control is then given as

$$\tilde{\xi}_{mj} = p_{mj} - E[\,p_{mj} \,|\, \tilde{z}_{mj}],$$

and we estimate the expectation using ordinary least squares.

The control function in our setup is given as $f(z_j, \mathbf{V}_j) = E[\xi_j | z_j, \mathbf{V}_j]$ and for consistency setting $\mathbf{V}_j = \tilde{\xi}_{mj}$ is sufficient. However, $f(z_j, \mathbf{V}_j)$ is a new regressor in our setting, and more variation in this regressor can help to improve precision of the parameter estimates. We add two additional controls that may lead to an increase in the variation of $E[\xi_j | z_j, \mathbf{V}_j]$. Following the logic used in refining the instrument set, we use

$$\tilde{\xi}_{(1)mj} = \sum_{k \neq j, k \in J_f} \tilde{\xi}_{mk}$$

and

$$\tilde{\xi}_{(2)mj} = \sum_{k \notin J_f} \tilde{\xi}_{mk},$$

where J_f is the set of products produced by the firm that produces the product j. These controls are respectively the sum of all of the other residuals of the products made by the same firm, given by $\tilde{\xi}_{(1)mj}$, and the sum of all the residuals of all the products made by other firms, given by $\tilde{\xi}_{(2)mj}$.

Based on these $\tilde{\xi}_{mj}$, $\tilde{\xi}_{(1)mj}$, and $\tilde{\xi}_{(2)mj}$, we generate the following nine controls that we use for our estimation:

$$
\begin{aligned}
V_{1mj} &= \tilde{\xi}_{mj}, V_{2mj} = \tilde{\xi}_{mj}^2 - E[\tilde{\xi}_{mj}^2 | \tilde{z}_{mj}], V_{3mj} = \tilde{\xi}_{mj}^3 - E[\tilde{\xi}_{mj}^3 | \tilde{z}_{mj}], \\
V_{4mj} &= \tilde{\xi}_{(1)mj}, V_{5mj} = \tilde{\xi}_{(1)mj}^2 - E[\tilde{\xi}_{(1)mj}^2 | \tilde{z}_{mj}], V_{6mj} = \tilde{\xi}_{(1)mj}^3 - E[\tilde{\xi}_{(1)mj}^3 | \tilde{z}_{mj}], \\
V_{7mj} &= \tilde{\xi}_{(2)mj}, V_{8mj} = \tilde{\xi}_{(2)mj}^2 - E[\tilde{\xi}_{(2)mj}^2 | \tilde{z}_{mj}], V_{9mj} = \tilde{\xi}_{(2)mj}^3 - E[\tilde{\xi}_{(2)mj}^3 | \tilde{z}_{mj}].
\end{aligned}
$$

Our model for δ_{mj} then becomes

$$\delta_{mj} = c + \beta' x_{mj} - \alpha p_{mj} + f(\tilde{z}_{mj}, \hat{\mathbf{V}}_{mj})(1 + \gamma' x_{mj} + \gamma_p(\bar{y}_m - p_{mj})),$$

where we approximate $f(\tilde{z}_{mj}, \hat{\mathbf{V}}_{mj}) = \sum_{l=1}^9 \pi_l \hat{V}_{lmj}$ with parameters $\pi = (\pi_1, \ldots, \pi_9)$ to be estimated.

8.2 Estimation

Letting $\theta = (c, \beta', \alpha, \gamma', \gamma_p)'$ we have three sets of parameters to identify given by (σ, θ, π). Estimation proceeds as in Berry, Levinsohn, and Pakes (1995). Given a value of σ, we use the contraction mapping to solve for the vector $\tilde{\delta}_m(\sigma)$ that satisfies $s(\sigma, \delta(\sigma)) = s^{Data}$. $\tilde{\delta}_m(\sigma)$ then becomes the regressand in the non-linear least squares objective function given as

$$Q_J(\theta(\sigma), \pi(\sigma); \sigma) = \min_{\theta, \pi} \frac{1}{J} \sum_{m=1}^{M} \sum_{j=1}^{J_m} \{\tilde{\delta}_{mj}(\sigma) - (c + \beta' x_{mj} - \alpha p_{mj} + f(\tilde{z}_{mj}, \hat{\mathbf{V}}_{mj})(1 + \gamma' x_{mj} + \gamma_p(\bar{y}_m - p_{mj})))\}^2$$

with $J = \sum_{m=1}^{M} J_m$ and $f(\tilde{z}_{mj}, \hat{\mathbf{V}}_{mj}) = \sum_{l=1}^{9} \pi_l \hat{V}_{lmj}$. This procedure is used iteratively to minimize $Q_J(\theta(\sigma), \pi(\sigma); \sigma)$ over σ, yielding parameter estimates $(\hat{\sigma}, \hat{\theta}, \hat{\pi}) = (\hat{\sigma}, \hat{\theta}(\hat{\sigma}), \hat{\pi}(\hat{\sigma}))$ such that $\hat{\sigma} = \text{argmin}_\sigma Q_J(\theta(\sigma), \pi(\sigma); \sigma)$.

8.3   Results

The first three columns of Table 1 report results for different specifications in the case where $\mu_{ij} = 0$, so the dependent variable is $\delta_{mj} = \ln(s_{mj}) - \ln(s_{m0})$, where s_{mj} and s_{m0} denote respectively the observed market shares in market m for good j and for the outside good. Column 4 reports results with $\mu_{ij} \neq 0$, with the market vector δ_m then recovered from matching observed to predicted market shares conditional on all parameters not entering into mean utility. Table 2 reports the implied demand elasticities.

The results for the separable error and exogenous price case are in Column 1 of Table 1 and they replicate those results from the first column of Table III in BLP. The price coefficient increases from -0.088 to -0.136 when we move from OLS to 2SLS, suggesting prices are endogenous, as noted in Berry, Levinsohn, and Pakes (1995).[20]

Column 3 includes our CMRCF results where we do not impose $(\gamma, \gamma_p) = 0$. The additively separable specification is rejected at 5% as the p-value for $H_0 : (\gamma_0, \gamma_{p0}) = 0$ is 0.019, although no single interaction term is significant on its own. The point estimate on the interaction term for price is negative but not significant, and thus only suggestive that the marginal utility of income declines as the demand error increases.

Most relevant for estimates of price elasticities is the bias in the 2SLS price coefficient estimate induced by the correlation between the instrumented price and the interaction term in the error. The price coefficient α increases from -0.136 to -0.232 and is also significantly different from the coefficient from 2SLS. The sign of the bias coupled with a negative estimate for the interaction term on price suggests that there is positive correlation between $-\hat{p}_j$ and $(\bar{y} - p_j)\xi_j$ conditional on x_j in the automobile data.

Column 4 allows for random coefficients in the non-separable specification. Horsepower/weight and miles-per-dollar have significant $\sigma'_k s$, but with the exception of the point estimate for β_k on Horsepower/weight, all of the other point estimates from Column 3 are largely the same. The

---

[20]While it does not change the substance of either their findings or our findings, we were not able to exactly replicate the results for their 2SLS estimator using the optimal instruments described in their paper. We find a price coefficient that is somewhat smaller than their original reported finding of -0.21. While we can only speculate as to the source of the difference, we suspect it lies in the instruments they used for these results, as we are able to replicate the OLS point estimates and standard deviations in their paper. Also consistent with this hypothesis is the fact that our estimate of -0.13 falls well within +/- two standard deviations of their estimate, as their standard deviation was -0.12. The significantly smaller standard deviation on our price coefficient also suggests the instruments they used for that specification - whatever they might of been - were not nearly as "optimal" as the instruments they propose in the paper, for which we find a much smaller standard deviation on the price coefficient.

Table 1: Estimated Parameters for Automobile Demand

No Correction, 2SLS, CMRCF (w/ Interactions), RandomCoefficient-CMRCF (w/ Interactions)

Dependent Variable is $\widehat{\delta}_{mj}$

Parameter	Variable	No Correction*	2SLS (No Interactions)	CMRCF (w/ Interactions)	RC-CMRCF (w/Interactions)
Term on Price	price	-0.088	-0.136	-0.232	-0.233
		(0.004)	(0.011)	(0.019)	(0.019)
Mean	Constant	-10.071	-9.915	-9.661	-9.332
Parameters		(0.252)	(0.263)	(0.290)	(0.317)
	HP/Weight	-0.122	1.226	2.813	1.019
		(0.277)	(0.404)	(0.527)	(1.074)
	Air	-0.034	0.486	1.384	1.410
		(0.072)	(0.133)	(0.179)	(0.186)
	MP\$	0.265	0.172	0.104	0.121
		(0.043)	(0.049)	(0.054)	(0.056)
	Size	2.342	2.292	2.365	2.409
		(0.125)	(0.129)	(0.140)	(0.148)
Interaction	$(\bar{y}$-price$)\cdot\xi$			-0.043	-0.042
Parameters				(0.041)	(0.044)
	HP/Weight$\cdot\xi$			1.158	1.258
				(1.809)	(2.071)
	Air$\cdot\xi$			0.514	0.627
				(0.636)	(0.814)
	MP\$$\cdot\xi$			-0.096	-0.145
				(0.116)	(0.153)
	Size$\cdot\xi$			0.262	0.368
				(0.728)	(0.993)
Std. Deviations	Constant				0.042 (0.228)
	HP/Weight				2.214 (0.792)
	Air				0.034 (0.578)
	MP\$				0.083 (0.014)
	Size				0.001 (0.181)
Control Ftns	V_1			2.605	2.388
				(2.683)	(2.702)
	V_2			-0.898	-0.716
				(1.606)	(1.453)
	V_3			0.114	0.087
				(0.774)	(0.689)
	V_4			-0.530	-0.485
				(0.533)	(0.536)
	V_5			0.063	0.067
				(0.132)	(0.132)
	V_6			0.799	0.740
				(0.805)	(0.820)
	V_7			-0.062	-0.062
				(0.099)	(0.099)
	V_8			0.216	0.214
				(0.233)	(0.248)
	V_9			-0.081	-0.083
				(0.192)	(0.187)

The data are identical to BLP (1995). Column 1 replicates estimates for the model of their first column of results in their Table III. The second column uses the same instruments from BLP and estimates 2SLS for the characteristics used in Column 1. The third column reports estimates of our CMRCF approach. The last column reports the CMRCF estimates of the random coefficients model with interactions. We do not impose a supply side model during estimations. Standard errors reported for our CMRCF and RC-CMRCF estimators are robust to heteroskedasticity and account for the "first and second-stage estimates" following Kim and Petrin (2010c). The p-value for H_0 :all the interaction parameters equal to zero is 0.019 for the CMRCF and is 0.036 for the RC-CMRCF.

presence of the random coefficients does not change the fact that $H_0 : (\gamma_0, \gamma_{p0}) = 0$ is rejected at 5% as the p-value is 0.036, and the coefficient on the price coefficient changes from 0.232 to 0.233 and the price interaction term from -0.043 to -0.042.

Table 2 translates these estimates into elasticities. Berry, Levinsohn, and Pakes (1995) report elasticities for selected automobiles from 1990, so we do the same, choosing every fourth automobile from their Table III, in which vehicles are sorted in order of ascending price. The first column uses the uncorrected logit specification from Column 1 of Table III in BLP (1995).[21] Ignoring price endogeneity severely biases price elasticities towards zero. As we control the endogeneity using the 2SLS the price elasticities change significantly and become more elastic, as the median elasticity moves from -0.77 to -1.18. However, biggest change comes when we move from 2SLS to our CMRCF approach allowing for interactions, as the median elasticity increases from -1.18 to -2.05, and the mean elasticity increases from -1.60 to -2.65. Adding the random coefficients to the non-separable specification has very little effect on the elasticities reported in Table 2, as is clear from examining columns three and four.

Table 2
Automobile Elasticities: No Correction, 2SLS (without Interactions),
CMRCF, and RandomCoefficient-CMRCF (with Interactions)

	No Correction[1]	2SLS	CMRCF	RC-CMRCF
Interactions	No	No	Yes	Yes
Results for 1971-1990				
Median	-0.77	-1.18	-2.05	-2.04
Mean	-0.75	-1.60	-2.65	-2.65
Standard Deviation	0.34	1.17	1.69	1.70
No. of Inelastic Demands	68%	21%	4%	4%
Elasticities from 1990				
Median	-0.93	-1.43	-2.80	-2.79
Mean	-0.91	-1.90	-3.23	-3.24
Standard Deviation	0.46	1.28	1.85	1.87
No. of Inelastic Demands	53%	12%	2%	2%
1990 Models (from BLP, Table VI):				
Mazda 323	-0.44	-0.69	-1.61	-1.61
Honda Accord	-0.81	-1.26	-1.42	-1.47
Acura Legend	-1.67	-2.57	-4.17	-4.20
BMW 735i	-3.39	-5.09	-7.14	-7.26

The uncorrected specification is that from Table III of BLP (1995). 1990 is the year BLP focus on for the individual models; we choose every fourth automobile from their Table VI (the other elasticities were also very similar).

---

[21] Because the data sets are the same, these are the same elasticities that result from the coefficients of their Table III.

9   Conclusion

We show how to allow for interactions in the utility function between the unobserved demand factor and observed factors including price in a discrete choice demand setting. We start by noting that when endogenous variables interact with the demand error the inversion and contraction from Berry (1994) and Berry, Levinsohn, and Pakes (1995) can still be used to recover mean utility. However, the standard IV approach is no longer consistent because the price interaction term is correlated with the instrumented price. Furthermore, the conditional mean restrictions (CMR) used for identification in Berry (1994) and Berry, Levinsohn, and Pakes (1995) are no longer sufficient for identification.

We show how to consistently estimate demand parameters while allowing for both endogenous and exogenous variables to interact with the error. We couple the standard CMRs with new moment conditions that we call "generalized control function moments." We require only the use of the exact same instruments used in the separable setting. Our approach thus extends the non-separable demand literature as we do not require that our controls be one-to-one with the unobserved factors, as in Bajari and Benkard (2005) or Kim and Petrin (2010a).

We develop a sieve semiparametric estimator for the nonseparable demand models. Given mean utility it is a simple three-step estimator to recover the parameters subsumed in the mean utility term, including those parameters on the interaction terms. Monte Carlos suggest standard IV estimators in the non-separable setting perform poorly, while our approach is consistent. Using the same automobile data as was used in Berry, Levinsohn, and Pakes (1995), our estimates reveal that the interactions terms are significant and the demand elasticities become 60% more elastic relative to the standard IV estimator, primarily because the coefficient on price changes substantially when the interaction terms are included.

A Proof of Consistency (Theorem 5)

We prove the consistency by extending Chen (2006)'s consistency proof for sieve extremum estimators allowing for pre-step estimates. We first show that any (infeasible) estimator, $(\tilde{\theta}, \tilde{f})$ defined as any sequence that satisfies the following is consistent:

$$Q_J(\delta, z, p, \hat{\mathbf{V}}; \tilde{\theta}, \tilde{f}) \leq \inf_{(\theta, f) \in \Theta \times \mathcal{F}_J} Q_J(\delta, z, p, \hat{\mathbf{V}}; \theta, f) + o_p(1). \tag{16}$$

Let $\varepsilon > 0$ be any small real numbers. Any estimator $(\tilde{\theta}, \tilde{f})$ that satisfies (16) also satisfies that with probability approaching to one (w.p.a.1), $Q_J(\delta, z, p, \hat{\mathbf{V}}; \tilde{\theta}, \tilde{f}) < Q_J(\delta, z, p, \hat{\mathbf{V}}; \theta, f_J) + \frac{\varepsilon}{6}$ for all $(\theta, f_J) \in \Theta \times \mathcal{F}_J$. From the fact that $\theta_0 \in \Theta$ and $\pi_J f_0 \in \mathcal{F}_J$, it follows that $Q_J(\delta, z, p, \hat{\mathbf{V}}; \tilde{\theta}, \tilde{f}) < Q_J(\delta, z, p, \hat{\mathbf{V}}; \theta_0, \pi_J f_0) + \frac{\varepsilon}{6}$. Then by Assumption A8 and the consistency of the pre-stage estimators (A3), we have w.p.a.1, $Q_J^0(\delta, z, p, \hat{\mathbf{V}}; \tilde{\theta}, \tilde{f}) - Q_J(\delta, z, p, \hat{\mathbf{V}}; \tilde{\theta}, \tilde{f}) < \frac{\varepsilon}{6}$ and $Q_J^0(\delta, z, p, \hat{\mathbf{V}}; \theta_0, \pi_J f_0) - Q_J(\delta, z, p, \hat{\mathbf{V}}; \theta_0, \pi_J f_0) > -\frac{\varepsilon}{6}$. It follows that w.p.a.1,

$$\begin{aligned}
Q_J^0(\delta, z, p, \hat{\mathbf{V}}; \tilde{\theta}, \tilde{f}) - \frac{\varepsilon}{6} \quad &< \quad Q_J(\delta, z, p, \hat{\mathbf{V}}; \tilde{\theta}, \tilde{f}) \\
&< \quad Q_J(\delta, z, p, \hat{\mathbf{V}}; \theta_0, \pi_J f_0) + \frac{\varepsilon}{6} < Q_J^0(\delta, z, p, \hat{\mathbf{V}}; \theta_0, \pi_J f_0) + \frac{\varepsilon}{6} + \frac{\varepsilon}{6}.
\end{aligned}$$

Next we note that by the continuity assumption (A6) and the consistency of the pre-stage estimators (A3), we have w.p.a.1, $Q_J^0(\delta, z, p, \mathbf{V}; \tilde{\theta}, \tilde{f}) - Q_J^0(\delta, z, p, \hat{\mathbf{V}}; \tilde{\theta}, \tilde{f}) < \frac{\varepsilon}{6}$ and $Q_J^0(\delta, z, p, \mathbf{V}; \theta_0, \pi_J f_0) - Q_J^0(\delta, z, p, \hat{\mathbf{V}}; \theta_0, \pi_J f_0) > -\frac{\varepsilon}{6}$. It follows that w.p.a.1,

$$Q_J^0(\delta, z, p, \mathbf{V}; \tilde{\theta}, \tilde{f}) - \frac{\varepsilon}{6} < Q_J^0(\delta, z, p, \mathbf{V}; \theta_0, \pi_J f_0) + \frac{\varepsilon}{6} + \frac{3\varepsilon}{6}.$$

By A1 and A5 (continuity) and the fact that $\|f_0 - \pi_J f_0\|_{\mathcal{F}} \to 0$ as $J \to \infty$, for all $J > J_0$ large enough we have $Q_J^0(\delta, z, p, \mathbf{V}; \theta_0, \pi_J f_0) < Q_J^0(\delta, z, p, \mathbf{V}; \theta_0, f_0) + \frac{\varepsilon}{6}$. It follows that

$$Q_J^0(\delta, z, p, \mathbf{V}; \tilde{\theta}, \tilde{f}) < Q_J^0(\delta, z, p, \mathbf{V}; \theta_0, f_0) + \varepsilon. \tag{17}$$

Next note that for any $\epsilon > 0$, by A4, A5(continuity), A7 (compactness),

$$\inf_{\{(\theta, f) \in \Theta \times \mathcal{F}_J : \|\theta - \theta_0\| + \|f - f_0\|_{\mathcal{F}} \geq \epsilon\}} Q_J^0(\delta, z, p, \mathbf{V}; \theta, f)$$

exists (it can vary by J). Then by A1 (identification) and the fact that $\mathcal{F}_J \subset \mathcal{F}$, it must be that

$$Q_J^0(\delta, z, p, \mathbf{V}; \theta_0, f_0) < \inf_{\{(\theta, f) \in \Theta \times \mathcal{F}_J : \|\theta - \theta_0\| + \|f - f_0\|_{\mathcal{F}} \geq \epsilon\}} Q_J^0(\delta, z, p, \mathbf{V}; \theta, f).$$

Take ε small enough that $\inf_{\{(\theta, f) \in \Theta \times \mathcal{F}_J : \|\theta - \theta_0\| + \|f - f_0\|_{\mathcal{F}} \geq \epsilon\}} Q_J^0(\delta, z, p, \mathbf{V}; \theta, f) - Q_J^0(\delta, z, p, \mathbf{V}; \theta_0, f_0) \geq \varepsilon$. Then from (17) it follows that w.p.a.1, $Q_J^0(\cdot; \tilde{\theta}, \tilde{f}) < \inf_{\{(\theta, f) \in \Theta \times \mathcal{F}_J : \|\theta - \theta_0\| + \|f - f_0\|_{\mathcal{F}} \geq \epsilon\}} Q_J^0(\cdot; \theta, f)$. Then by A5 (continuity) and the fact that $(\tilde{\theta}, \tilde{f}) \in \Theta \times \mathcal{F}_J$, we conclude $\|\tilde{\theta} - \theta_0\| + \|\tilde{f} - f_0\|_{\mathcal{F}} < \epsilon$. This proves any estimator, $(\tilde{\theta}, \tilde{f})$ that satisfies (16) is consistent. Next we note that our estimator

$(\hat{\theta}, \hat{f})$ satisfies the following, so is consistent:

$$
\begin{aligned}
Q_J(\delta, z, p, \hat{\mathbf{V}}; \hat{\theta}, \hat{f}) &\leq \inf_{(\theta, f) \in \Theta \times \hat{\mathcal{F}}_J} Q_J(\delta, z, p, \hat{\mathbf{V}}; \theta, f) + o_p(1) \\
&= \inf_{(\theta, f) \in \Theta \times \mathcal{F}_J} Q_J(\delta, z, p, \hat{\mathbf{V}}; \theta, f) + o_p(1)
\end{aligned}
$$

where the first inequality holds by Assumption A2 (extremum estimator) and the second equality holds because $\hat{\mathcal{F}}_J \to \mathcal{F}_J$ by Assumption A3 and $Q_J(\delta, z, p, \hat{\mathbf{V}}; \theta, f)$ is continuous in f.

B Consistency Theorem for Random Coefficients Logit Models

We extend Berry, Linton, and Pakes (2004)'s consistency theorem to the case of our estimator. In proofs of consistency for estimators of random coefficients models, when the asymptotics are in the number of products Berry, Linton, and Pakes (2004) argue against maintaining uniform convergence of the objective function and our proof accordingly avoids it. We heavily borrow notation and regularity conditions from Berry, Linton, and Pakes (2004). The key complication is that our estimator is a multi-step estimator, so we need to control the approximation errors from pre-step estimators together with two other sources of errors, the sampling error in the observed shares and the simulation error in the simulated distribution of the random coefficients.

For transparency here we suppress the market index and assume the data is from a single market, $M = 1$. The theorem naturally extends to the data of the multiple markets.

Let $\nu(x, p, \xi, \lambda, \theta, \theta_\lambda)$ be a $J \times 1$ share function specific to a household type λ and $P(\lambda)$ be the distribution of λ that represents household heterogeneity. Then the vector of aggregate market shares predicted by the random coefficients model at given values of the parameters θ in the mean utility, the parameter θ_λ in $P(\lambda)$, x, p, ξ, and P are

$$
\sigma(\delta(x, p, \xi, \theta), x, p, \theta_\lambda, P) = \int \nu(x, p, \xi, \lambda, \theta, \theta_\lambda) dP(\lambda).
$$

Note that our notation allows for random coefficients on (x, p) but not ξ, so ξ appears only in the mean utility. The function $\sigma(\cdot)$ maps the appropriate product space to the $J + 1$ dimensional unit simplex for shares,

$$
\mathcal{S}_J = \{(s_0, \ldots, s_J)' | 0 \leq s_j \leq 1 \text{ for } j = 0, \ldots, J, \text{ and } \sum_{j=0}^{J} s_j = 1\}.
$$

The actual market shares in the population are given by evaluating $\sigma(\delta(\cdot, \theta), x, p, \theta_\lambda, P)$ at $(\theta_0, \theta_{\lambda 0}, P^0)$, the true value of θ, θ_λ, and P.

As the first source of error in the implementation of the random coefficients model we approximate P^0 by simulating it with P^R, the empirical measure of some i.i.d. sample $\lambda_1, \ldots, \lambda_R$ such that

(e.g.)

$$\sigma(\delta(\cdot,\theta),x,p,\theta_\lambda,P^R) = \int \nu(x,p,\xi,\lambda,\theta,\theta_\lambda)dP^R(\lambda) = \frac{1}{R}\sum_{r=1}^{R}\nu(x,p,\xi,\lambda_r,\theta,\theta_\lambda).$$

The second source of error is the sampling error in observed market shares s^n, typically constructed from n i.i.d. draws from the population of consumers. We denote by s^0, the population market shares. We assume Berry, Linton, and Pakes (2004)'s Assumption A1 that regulates the simulation errors and the sampling errors.

Assumption 9. *The market shares $s_j^n = \frac{1}{n}\sum_{i=1}^{n}1(C_i = j)$, where C_i is the choice of the i-th consumer, and C_i are i.i.d. across i. For any fixed $(x,p,\xi,\theta,\theta_\lambda)$,*

$$\sigma_j(\delta(\cdot,\theta),x,p,\theta_\lambda,P^R) - \sigma_j(\delta(\cdot,\theta),x,p,\theta_\lambda,P^0) = \frac{1}{R}\sum_{r=1}^{R}\varepsilon_{j,r}(\delta(\cdot,\theta),x,p,\theta_\lambda),$$

where $\varepsilon_{j,r}(\delta(\cdot,\theta),x,p,\theta_\lambda)$ is bounded, continuous, and differentiable in $\delta(\cdot,\theta)$, θ, and θ_λ.

The following regularity conditions on $\sigma(\delta(\cdot,\theta),x,p,\theta_\lambda,P)$ are from Assumption A2 of Berry, Linton, and Pakes (2004) but the key difference is that we require some conditions to hold in terms of the mean utility $\delta(\cdot,\theta)$ rather than ξ. As consequences, our assumption can be weaker than Assumption A2 of Berry, Linton, and Pakes (2004) because we allow that $\delta_j(\cdot,\theta)$ is not necessarily monotonic in ξ_j (so $\sigma_j(\delta(\cdot,\theta),x,p,\theta_\lambda,P)$ does not need to be monotonic in ξ_j) in our models with interactions. But note that we specifically consider the random coefficients logit model while Berry, Linton, and Pakes (2004) is applicable to other models too.

Assumption 10. *(i) For every finite J, for all finite δ and $\theta_\lambda \in \Theta_\lambda$, and for all P in a neighborhood of P^0, $\frac{\partial \sigma_j(\delta,\cdot,\theta_\lambda,P)}{\partial \delta_k}$ exists, and is continuously differentiable in both δ and θ_λ, with $\frac{\partial \sigma_j(\delta,\cdot,\theta_\lambda,P)}{\partial \delta_j} > 0$, and for $k \neq j$, $\frac{\partial \sigma_j(\delta,\cdot,\theta_\lambda,P)}{\partial \delta_k} \leq 0$. The matrix $\frac{\partial \sigma(\delta,\cdot,\theta_\lambda,P)}{\partial \delta'}$ is invertible for all J; (ii) $s_j^0 > 0$ for all j; (iii) For every finite J, for all $\theta \in \Theta$, $\delta(\cdot,\theta)$ is continuously differentiable in θ.*

We outline our estimation procedure for which we show the consistency specifically. We first obtain the mean utility $\delta^* = \delta^*(x,p,\theta_\lambda,s,P)$ that solves

$$s - \sigma(\delta^*,x,p,\theta_\lambda,P) = 0$$

and this solution is unique under the conditions outlined in BLP (1995), which we maintain here. Therefore s and δ^* are in one-to-one relation for any θ_λ and P. Then by the implicit function theorem, Dieudonne (1969)(Theorem 10.2.1), and Assumption 9, the mapping $\delta^*(x,p,\theta_\lambda,s,P)$ is continuously differentiable in θ_λ,s,P, in some neighborhood. Moreover, Assumption 10 above allows us to expand the inverse map from $(x,p,\theta_\lambda,s^n,P)$ to $\delta^*(\cdot)$ around s^0, which in turn facilitates controlling the sampling error to show the consistency. As the convention, we let the true value of δ^*, $\delta^{*0} = \delta^*(x,p,\theta_{\lambda 0},s^0,P^0)$ be the solution to

$$s^0 - \sigma(\delta^*,x,p,\theta_{\lambda 0},P^0) = 0.$$

29

Note that in the additive models where ξ_j enters additively in $\delta_j(\cdot, \theta)$, inverting ξ is equivalent to inverting δ^* but in non-additive models it is convenient to characterize the inversion in term of the mean utility. First some regularity conditions below regarding $\delta^*(\cdot)$ does not depend on θ, the parameters in the mean utility. Given θ_λ, i.e., $\delta^*(\cdot, \theta_\lambda, \cdot)$ the estimation procedure of the mean utility parameters is identical to that of the model (in Section 6) without random coefficients when we treat $\delta^*(\cdot, \theta_\lambda, \cdot)$ as the actual mean utility.

We need to specify the mean utility for estimation (e.g.)

$$\delta_j^* = \delta_j(x_j, p_j, \xi_j, \theta) = c + \beta' x_j - \alpha p_j + \xi_j + \gamma' x_j \xi_j + \gamma_p(\bar{y} - p_j)\xi_j.$$

Then the moment condition (that corresponds to (7) for the model without random coefficients) we use becomes

$$0 = E\left[\delta_j^*(x_j, p_j, \theta_{\lambda 0}, s^0, P^0) - \{c_0 + \beta_0' x_j - \alpha_0 p_j + f_0(z_j, \mathbf{V}_j)(1 + \gamma_0' x_j + \gamma_{p0}(\bar{y} - p_j))\}|z_j, \mathbf{V}_j\right]$$

for $j = 1, \ldots, J$ where $f_0(z_j, \mathbf{V}_j) = E[\xi_j | z_j, \mathbf{V}_j]$. In the estimation we replace s^0 with s^n and $\delta^*(\cdot, s^0, P^0)$ with $\delta^*(\cdot, s^n, P^R)$. We also approximate $\mathbf{V}_j$ with $\hat{\mathbf{V}}_j$ and approximate f_0 with a function in the sieve space $\hat{\mathcal{F}}_J$ defined in (14). Therefore we obtain the estimator $(\hat{\theta}, \hat{\theta}_\lambda, \hat{f}(\cdot))$ that minimizes the sample criterion function such that

$$(\hat{\theta}, \hat{\theta}_\lambda, \hat{f}) = \text{arginf}_{(\theta, \theta_\lambda, f) \in \Theta \times \Theta_\lambda \times \hat{\mathcal{F}}_J} Q_J(\delta^*(\cdot, \theta_\lambda, s^n, P^R), z, p, \hat{\mathbf{V}}; \theta, f) \tag{18}$$

where we define (e.g.) in the case of the sieve least squares estimation

$$Q_J(\delta^*(\cdot, \theta_\lambda, s, P), \cdot, \mathbf{V}; \theta, f) \equiv \frac{1}{J}\sum_{j=1}^{J}\{\delta_j^*(x_j, p_j, \theta_\lambda, s, P) - (c + \beta' x_j - \alpha p_j + f(z_j, \mathbf{V}_j)(1 + \gamma' x_j + \gamma_p(\bar{y} - p_j)))\}^2.$$

Also define the population criterion function $Q_J^0(\delta^*(\cdot, \theta_\lambda, s, P), \cdot, \mathbf{V}; \theta, f) = E[Q_J(\delta^*(\cdot, \theta_\lambda, s, P), \cdot, \mathbf{V}; \theta, f)]$.

To obtain the consistency theorem we need to add further assumptions (Assumption 11 below) that control the way in which s^n approaches s^0 and $\sigma(\delta^*(\cdot), \cdot, \theta_\lambda, P^R)$ approaches to $\sigma(\delta^*(\cdot), \cdot, \theta_\lambda, P^0)$ (corresponding to Assumption A3 in Berry, Linton, and Pakes (2004)) and add identification conditions (Assumption 17 and 18 below, corresponding to Assumptions A3-A6 in Berry, Linton, and Pakes (2004)). We also add restrictions on the rate at which s_j^0 approaches to zero (Condition S in Berry, Linton, and Pakes (2004)).

Condition 3 (S)**.** *There exist positive finite constants $\underline{c}$ and $\bar{c}$ such that with probability one*

$$\underline{c}/J \leq s_j^0 \leq \bar{c}/J, \;\; j = 0, 1, \ldots, J.$$

In the following we work on the product space $\Theta \times \Theta_\lambda \times \mathcal{F} \times \mathcal{S}_J \times \mathbf{P}$ where $\mathbf{P}$ is the set of probability measures and endow the marginal spaces with (pseudo) metrics: $\rho_P(P, \tilde{P}) = \sup_{B \in \mathcal{B}} |P(B) - \tilde{P}(B)|$, where $\mathcal{B}$ is the class of all Borel sets on $\mathbb{R}^{\dim(\lambda)}$, the Euclidean metric $\rho_E(\cdot, \cdot)$ on Θ and Θ_λ, the

pseudo metric $||\cdot||_{\mathcal{F}}$ on $\mathcal{F}$, and a metric ρ_{s^0} on $\mathcal{S}_J$, defined by

$$\rho_{s^0}(s,\tilde{s}) = \max_{0\leq j\leq J}\left|\frac{s_j - \tilde{s}_j}{s_j^0}\right|.$$

The same metric is used for $\sigma_j(\cdot)$ in place of s_j.

We also use the metric $\rho_\delta(\delta^*,\tilde{\delta}^*) = J^{-1}\sum_{j=1}^J(\delta_j^* - \tilde{\delta}_j^*)^2$. Lastly define for each $\epsilon > 0$, the following neighborhoods of θ_0, $\theta_{\lambda 0}$, f_0, P^0, and s^0: $\mathcal{N}_{\theta_0}(\epsilon) = \{\theta : \rho_E(\theta,\theta_0) < \epsilon\}$, $\mathcal{N}_{\theta_{\lambda 0}}(\epsilon) = \{\theta_\lambda : \rho_E(\theta_\lambda,\theta_{\lambda 0}) < \epsilon\}$, $\mathcal{N}_{f_0,J}(\epsilon) = \{f : ||f - f_0||_{\mathcal{F}} < \epsilon, f \in \mathcal{F}_J\}$, $\mathcal{N}_{P^0}(\epsilon) = \{P : \rho_P(P,P^0) < \epsilon\}$, and $\mathcal{N}_{s^0}(\epsilon) = \{s : \rho_{s^0}(s,s^0) < \epsilon\}$. Also for each θ_λ and $\epsilon > 0$, define $\mathcal{N}_{\delta^{*0}}(\theta_\lambda,\epsilon) = \{\delta^* : \rho_\delta(\delta^*,\delta^*(\cdot,\theta_\lambda,s^0,P^0)) < \epsilon\}$. We further denote $\mathcal{N}_{\Pi_0}(\epsilon) = \{\Pi : ||\Pi - \Pi_0||_s < \epsilon\}$ and $\mathcal{N}_{\bar{\varphi}_{0l}}(\epsilon) = \{\bar{\varphi}_l : ||\bar{\varphi}_l - \bar{\varphi}_{0l}||_s < \epsilon\}$ for the pseudo metric $||\cdot||_s$.

Assumption 11. *The random sequences s^n and $\sigma^R(\theta_\lambda)$ are consistent with respect to the corresponding metrics,*

$$\text{(a) } \rho_{s^0}(s^n,s^0) \to_p 0; \quad \text{(b) } \sup_{\theta_\lambda\in\Theta_\lambda} \rho_{\sigma(\theta_\lambda)}(\sigma^R(\theta_\lambda),\sigma(\theta_\lambda)) \to_p 0$$

where $\sigma^R(\theta_\lambda) = \sigma(\delta^*(\cdot,\theta_\lambda,s^0,P^0),\cdot,\theta_\lambda,P^R)$ and $\sigma(\theta_\lambda) = \sigma(\delta^*(\cdot,\theta_\lambda,s^0,P^0),\cdot,\theta_\lambda,P^0)$. Furthermore suppose that the true market shares satisfy

$$\text{(c) } \frac{1}{nJ}\sum_{j=0}^J \frac{s_j^0(1-s_j^0)}{(s_j^0)^2} \to_p 0; \quad \text{(d) } \sup_{\theta_\lambda\in\Theta_\lambda} \frac{1}{RJ}\sum_{j=0}^J \frac{\sigma_j(\theta_\lambda)(1-\sigma_j(\theta_\lambda))}{(\sigma_j(\theta_\lambda))^2} \to_p 0.$$

Next we assume that our estimator is an extremum estimator that solves (18).

Assumption 12. $Q_J(\delta^*(\cdot,\hat{\theta}_\lambda,s^n,P^R),\cdot,\hat{\mathbf{V}};\hat{\theta},\hat{f}) \leq \inf_{(\theta,\theta_\lambda,f)\in\Theta\times\Theta_\lambda\times\hat{\mathcal{F}}_J}Q_J(\delta^*(\cdot,\theta_\lambda,s^n,P^R),\cdot,\hat{\mathbf{V}};\theta,f) + o_p(1)$.

The next condition is that in the small neighborhoods of $\Pi_0(\cdot)$ and $\bar{\varphi}_{0l}(\cdot)$, the difference between the sample criterion function and the population criterion function is small enough when J is large.

Assumption 13. *For any $C > 0$ there exists $\epsilon > 0$ such that*

$$\lim_{J\to\infty} \Pr\{ \sup_{(\theta,\theta_\lambda,f)\in\Theta\times\Theta_\lambda\times\mathcal{F}_J,\Pi\in\mathcal{N}_{\Pi_0}(\epsilon),\bar{\varphi}_{0l}\in\mathcal{N}_{\bar{\varphi}_{0l}}(\epsilon)\forall l} |Q_J(\delta^*(\cdot,\theta_\lambda,s^0,P^0),\cdot,\mathbf{V};\theta,f)$$
$$-Q_J^0(\delta^*(\cdot,\theta_\lambda,s^0,P^0),\cdot,\mathbf{V};\theta,f)| > C\} = 0$$

where $\mathbf{V}_j = g_j(p_1 - \Pi(z_1),\ldots,p_J - \Pi(z_J))$.

Although it is obvious, we add the following continuity conditions.

Assumption 14. $Q_J^0(\delta^*(\cdot,\theta_\lambda,s,P),\cdot,\mathbf{V};\theta,f)$ *is continuous in* $(\theta,\theta_\lambda,f) \in \Theta\times\Theta_\lambda\times\mathcal{F}_J$.

Assumption 15. $Q_J^0(\delta^*(\cdot,\theta_\lambda,s,P),\cdot,\mathbf{V};\theta,f_J)$ *is continuous in* $\Pi(\cdot)$ *and* $\bar{\varphi}_l(\cdot)$ *uniformly for all* $(\theta,\theta_\lambda,f_J) \in \Theta\times\Theta_\lambda\times\mathcal{F}_J$.

We also assume our parameter space is compact.

Assumption 16. *The parameter space $\Theta \times \Theta_\lambda$ is compact and the sieve space, $\mathcal{F}_J$, is compact under the pseudo-metric $|| \cdot ||_{\mathcal{F}}$.*

The next condition ensures that we can, at least asymptotically, distinguish the δ^* that sets the models predictions for shares equal to the actual shares from other values of δ^*.

Assumption 17. *For all ϵ, there exists $C(\epsilon) > 0$ such that $\sigma(\delta^*, x, p, \theta_\lambda, P)$*

$$\lim_{J \to \infty} \Pr\{\inf_{\theta_\lambda \in \Theta_\lambda} \inf_{\delta^* \notin \mathcal{N}_{\delta^{*0}}(\theta_\lambda, \epsilon)} ||J^{-1/2} \log \sigma(\delta^*, \cdot, \theta_\lambda, P^0) - J^{-1/2} \log \sigma(\delta^*(\cdot, \theta_\lambda, s^0, P^0), \cdot, \theta_\lambda, P^0)|| > C(\epsilon)\} = 1.$$

The following is the key condition for the consistency on the limiting behavior of the population criterion function for $(\theta, \theta_\lambda, f)$ outside a neighborhood of $(\theta_0, \theta_{\lambda 0}, f_0)$.

Assumption 18. *(i) $Q_J^0(\delta^*(\cdot, \theta_{\lambda 0}, s^0, P^0), \cdot, \mathbf{V}; \theta_0, f_0) < \infty$; (ii) For all $\epsilon > 0$, there exists $C(\epsilon) > 0$ such that*

$$\lim_{J \to \infty} \{\inf_{\theta \notin \mathcal{N}_{\theta_0}(\epsilon), \theta_\lambda \notin \mathcal{N}_{\theta_{\lambda 0}}(\epsilon), f \notin \mathcal{N}_{f_0, J}(\epsilon)} Q_J^0(\delta^*(\cdot, \theta_\lambda, s^0, P^0), \cdot, \mathbf{V}; \theta, f) - Q_J^0(\delta^*(\cdot, \theta_{\lambda 0}, s^0, P^0), \cdot, \mathbf{V}; \theta_0, f_0)\} \geq C(\epsilon).$$

We now state our consistency theorem.

Theorem 6. *Suppose Assumptions 9-18 hold for some $n(J), R(J) \to \infty$. Further suppose Assumptions A3 and A4 hold. Then $\hat{\theta} \to_p \theta_0$ and $\hat{\theta}_\lambda \to_p \theta_{\lambda 0}$.*

B.1 Proof of General Consistency (Theorem 6)

In proving Theorem 6 we use a strategy very close to Berry, Linton, and Pakes (2004). We first show that the estimator, $(\tilde{\theta}, \tilde{\theta}_\lambda, \tilde{f})$ defined as any sequence that satisfies the following is consistent:

$$Q_J(\delta^*(\cdot, \tilde{\theta}_\lambda, s^0, P^0), z, p, \hat{\mathbf{V}}; \tilde{\theta}, \tilde{f}) = \inf_{(\theta, \theta_\lambda, f) \in \Theta \times \Theta_\lambda \times \mathcal{F}_J} Q_J(\delta^*(\cdot, \theta_\lambda, s^0, P^0), z, p, \hat{\mathbf{V}}; \theta, f) + o_p(1). \tag{19}$$

Let $\varepsilon > 0$ be any small real numbers. Note that any estimator $(\tilde{\theta}, \tilde{\theta}_\lambda, \tilde{f})$ satisfying (19) also satisfies that with probability approaching to one (w.p.a.1), $Q_J(\delta^*(\cdot, \tilde{\theta}_\lambda, s^0, P^0), \cdot, \hat{\mathbf{V}}; \tilde{\theta}, \tilde{f}) < Q_J(\delta^*(\cdot, \theta_\lambda, s^0, P^0), \cdot, \hat{\mathbf{V}}; \theta, f_J) + \frac{\varepsilon}{6}$ for all $(\theta, \theta_\lambda, f_J) \in \Theta \times \Theta_\lambda \times \mathcal{F}_J$. Then from the fact that $(\theta_0, \theta_{\lambda 0}) \in \Theta \times \Theta_\lambda$ and $\pi_J f_0 \in \mathcal{F}_J$, it follows that

$$Q_J(\delta^*(\cdot, \tilde{\theta}_\lambda, s^0, P^0), z, p, \hat{\mathbf{V}}; \tilde{\theta}, \tilde{f}) < Q_J(\delta^*(\cdot, \theta_{\lambda 0}, s^0, P^0), z, p, \hat{\mathbf{V}}; \theta_0, \pi_J f_0) + \frac{\varepsilon}{6}.$$

Then by Assumption 13 and the consistency of the pre-stage estimators (A3), we have w.p.a.1, $Q_J^0(\delta^*(\cdot, \tilde{\theta}_\lambda, s^0, P^0), \cdot, \hat{\mathbf{V}}; \tilde{\theta}, \tilde{f}) - Q_J(\delta^*(\cdot, \tilde{\theta}_\lambda, s^0, P^0), \cdot, \hat{\mathbf{V}}; \tilde{\theta}, \tilde{f}) < \frac{\varepsilon}{6}$ and

$$Q_J^0(\delta^*(\cdot, \theta_{\lambda 0}, s^0, P^0), z, p, \hat{\mathbf{V}}; \theta_0, \pi_J f_0) - Q_J(\delta^*(\cdot, \theta_{\lambda 0}, s^0, P^0), z, p, \hat{\mathbf{V}}; \theta_0, \pi_J f_0) > -\frac{\varepsilon}{6}.$$

32

It follows that w.p.a.1,

$$Q_J^0(\delta^*(\cdot, \tilde{\theta}_\lambda, s^0, P^0), z, p, \hat{\mathbf{V}}; \tilde{\theta}, \tilde{f}) - \frac{\varepsilon}{6} < Q_J(\delta^*(\cdot, \tilde{\theta}_\lambda, s^0, P^0), z, p, \hat{\mathbf{V}}; \tilde{\theta}, \tilde{f})$$

$$< Q_J(\delta^*(\cdot, \theta_{\lambda 0}, s^0, P^0), z, p, \hat{\mathbf{V}}; \theta_0, \pi_J f_0) + \frac{\varepsilon}{6} < Q_J^0(\delta^*(\cdot, \theta_{\lambda 0}, s^0, P^0), z, p, \hat{\mathbf{V}}; \theta_0, \pi_J f_0) + \frac{\varepsilon}{6} + \frac{\varepsilon}{6}.$$

Next we note that by the continuity assumption (Assumption 15) and the consistency of the pre-stage estimators (A3), we have w.p.a.1, $Q_J^0(\delta^*(\cdot, \tilde{\theta}_\lambda, s^0, P^0), \cdot, \mathbf{V}; \tilde{\theta}, \tilde{f}) - Q_J^0(\delta^*(\cdot, \tilde{\theta}_\lambda, s^0, P^0), \cdot, \hat{\mathbf{V}}; \tilde{\theta}, \tilde{f}) < \frac{\varepsilon}{6}$ and $Q_J^0(\delta^*(\cdot, \theta_{\lambda 0}, s^0, P^0), \cdot, \mathbf{V}; \theta_0, \pi_J f_0) - Q_J^0(\delta^*(\cdot, \theta_{\lambda 0}, s^0, P^0), \cdot, \hat{\mathbf{V}}; \theta_0, \pi_J f_0) > -\frac{\varepsilon}{6}$. It follows that w.p.a.1,

$$Q_J^0(\delta^*(\cdot, \tilde{\theta}_\lambda, s^0, P^0), z, p, \mathbf{V}; \tilde{\theta}, \tilde{f}) - \frac{\varepsilon}{6} < Q_J^0(\delta^*(\cdot, \theta_{\lambda 0}, s^0, P^0), z, p, \mathbf{V}; \theta_0, \pi_J f_0) + \frac{\varepsilon}{6} + \frac{3\varepsilon}{6}.$$

Then by Assumption 18 and Assumption 14 (continuity) and the fact that $\|f_0 - \pi_J f_0\|_{\mathcal{F}} \to 0$ as $J \to \infty$, for all $J > J_0$ large enough we have

$$Q_J^0(\delta^*(\cdot, \theta_{\lambda 0}, s^0, P^0), z, p, \mathbf{V}; \theta_0, \pi_J f_0) < Q_J^0(\delta^*(\cdot, \theta_{\lambda 0}, s^0, P^0), z, p, \mathbf{V}; \theta_0, f_0) + \frac{\varepsilon}{6}.$$

It follows that

$$Q_J^0(\delta^*(\cdot, \tilde{\theta}_\lambda, s^0, P^0), z, p, \mathbf{V}; \tilde{\theta}, \tilde{f}) < Q_J^0(\delta^*(\cdot, \theta_{\lambda 0}, s^0, P^0), z, p, \mathbf{V}; \theta_0, f_0) + \varepsilon. \tag{20}$$

Next note that for any $\epsilon > 0$, by A4, Assumption 14 (continuity), 16 (compactness),

$$\inf_{\theta \notin \mathcal{N}_{\theta_0}(\epsilon), \theta_\lambda \notin \mathcal{N}_{\theta_{\lambda 0}}(\epsilon), f \notin \mathcal{N}_{f_0, J}(\epsilon)} Q_J^0(\delta^*(\cdot, \theta_\lambda, s^0, P^0), z, p, \mathbf{V}; \theta, f)$$

exists (it can vary by J). Then by Assumption 18 and the fact that $\mathcal{F}_J \subset \mathcal{F}$, it must be that

$$Q_J^0(\delta^*(\cdot, \theta_{\lambda 0}, s^0, P^0), z, p, \mathbf{V}; \theta_0, f_0) < \inf_{\theta \notin \mathcal{N}_{\theta_0}(\epsilon), \theta_\lambda \notin \mathcal{N}_{\theta_{\lambda 0}}(\epsilon), f \notin \mathcal{N}_{f_0, J}(\epsilon)} Q_J^0(\delta^*(\cdot, \theta_\lambda, s^0, P^0), z, p, \mathbf{V}; \theta, f).$$

Take ε small enough that

$$\inf_{\theta \notin \mathcal{N}_{\theta_0}(\epsilon), \theta_\lambda \notin \mathcal{N}_{\theta_{\lambda 0}}(\epsilon), f \notin \mathcal{N}_{f_0, J}(\epsilon)} Q_J^0(\delta^*(\cdot, \theta_\lambda, s^0, P^0), \cdot, \mathbf{V}; \theta, f) - Q_J^0(\delta^*(\cdot, \theta_{\lambda 0}, s^0, P^0), \cdot, \mathbf{V}; \theta_0, f_0) \geq \varepsilon.$$

Then from (20) it follows that w.p.a.1,

$$Q_J^0(\delta^*(\cdot, \tilde{\theta}_\lambda, s^0, P^0), z, p, \mathbf{V}; \tilde{\theta}, \tilde{f}) < \inf_{\theta \notin \mathcal{N}_{\theta_0}(\epsilon), \theta_\lambda \notin \mathcal{N}_{\theta_{\lambda 0}}(\epsilon), f \notin \mathcal{N}_{f_0, J}(\epsilon)} Q_J^0(\delta^*(\cdot, \theta_\lambda, s^0, P^0), z, p, \mathbf{V}; \theta, f).$$

Then by Assumption 14 (continuity) and the fact that $(\tilde{\theta}, \tilde{\theta}_\lambda, \tilde{f}) \in \Theta \times \Theta_\lambda \times \mathcal{F}_J$, we conclude $\tilde{\theta} \in \mathcal{N}_{\theta_0}(\epsilon), \tilde{\theta}_\lambda \in \mathcal{N}_{\theta_{\lambda 0}}(\epsilon)$, and $\tilde{f} \in \mathcal{N}_{f_0, J}(\epsilon)$. Therefore we have shown that any estimator $(\tilde{\theta}, \tilde{\theta}_\lambda, \tilde{f})$ that satisfies (19) is consistent.

Next we show that the actual estimator $(\hat{\theta}, \hat{\theta}_\lambda, \hat{f})$ satisfies the following, so is consistent because it then satisfies (19) :

$$Q_J(\delta^*(\cdot, \hat{\theta}_\lambda, s^0, P^0), z, p, \hat{\mathbf{V}}; \hat{\theta}, \hat{f}) = Q_J(\delta^*(\cdot, \hat{\theta}_\lambda, s^n, P^R), z, p, \hat{\mathbf{V}}; \hat{\theta}, \hat{f}) + o_p(1) \tag{21}$$

$$\leq \inf_{(\theta, \theta_\lambda, f) \in \Theta \times \Theta_\lambda \times \hat{\mathcal{F}}_J} Q_J(\delta^*(\cdot, \theta_\lambda, s^n, P^R), z, p, \hat{\mathbf{V}}; \theta, f) + o_p(1) \tag{22}$$

$$= \inf_{(\theta, \theta_\lambda, f) \in \Theta \times \Theta_\lambda \times \mathcal{F}_J} Q_J(\delta^*(\cdot, \theta_\lambda, s^n, P^R), z, p, \hat{\mathbf{V}}; \theta, f) + o_p(1) \tag{23}$$

$$= \inf_{(\theta, \theta_\lambda, f) \in \Theta \times \Theta_\lambda \times \mathcal{F}_J} Q_J(\delta^*(\cdot, \theta_\lambda, s^0, P^0), z, p, \hat{\mathbf{V}}; \theta, f) + o_p(1) \tag{24}$$

where (22) (the first inequality) holds because $(\hat{\theta}, \hat{\theta}_\lambda, \hat{f})$ is an extremum estimator satisfying (18) and (23) (the second equality) holds because $\hat{\mathcal{F}}_J \to \mathcal{F}_J$ (in the Hausdorff metric defined on the metric space $(\mathcal{F}, \|\cdot\|_{\mathcal{F}})$) by Assumption A3 and $Q_J(\cdot, \hat{\mathbf{V}}; \theta, f)$ is continuous in f. We focus on (21) (the first equality) and (24) (the last equality). Consider that by applying the Cauchy-Schwarz inequality twice we obtain

$$\sup\nolimits_{(\theta, \theta_\lambda, f) \in \Theta \times \Theta_\lambda \times (\hat{\mathcal{F}}_J \cup \mathcal{F}_J)} |Q_J(\delta^*(\cdot, \theta_\lambda, s^n, P^R), \cdot, \hat{\mathbf{V}}; \theta, f) - Q_J(\delta^*(\cdot, \theta_\lambda, s^0, P^0), \cdot, \hat{\mathbf{V}}; \theta, f)| \tag{25}$$

$$\leq \sup\nolimits_{(\theta, \theta_\lambda, f) \in \Theta \times \Theta_\lambda \times (\hat{\mathcal{F}}_J \cup \mathcal{F}_J)} J^{-1} \sum_{j=1}^{J} (\delta_j^*(\cdot, \theta_\lambda, s^n, P^R) - \delta_j^*(\cdot, \theta_\lambda, s^0, P^0))^2 \times$$

$$\times 2 \sup\nolimits_{(\theta, \theta_\lambda, f) \in \Theta \times \Theta_\lambda \times (\hat{\mathcal{F}}_J \cup \mathcal{F}_J)} (Q_J(\delta^*(\cdot, \theta_\lambda, s^n, P^R), \cdot, \hat{\mathbf{V}}; \theta, f) + Q_J(\delta^*(\cdot, \theta_\lambda, s^0, P^0), \cdot, \hat{\mathbf{V}}; \theta, f))$$

$$\leq C \cdot \sup\nolimits_{\theta_\lambda \in \Theta_\lambda} J^{-1} \sum_{j=1}^{J} (\delta_j^*(\cdot, \theta_\lambda, s^n, P^R) - \delta_j^*(\cdot, \theta_\lambda, s^0, P^0))^2$$

for some constant C. Here the second inequality holds because any $\delta^*(\cdot)$ obtained from the contraction mapping is bounded (BLP (1995) show the random coefficients logit model satisfies the contraction mapping property), all the parameter spaces are bounded (Assumption 16), and we assume z_j and p_j are (stochastically) bounded, so $\sup_{(\theta, \theta_\lambda, f) \in \Theta \times \Theta_\lambda \times (\hat{\mathcal{F}}_J \cup \mathcal{F}_J)} Q_J(\cdot)$ is bounded. Also note that $\delta^*(\cdot, \theta_\lambda, \cdot)$ does not depend on (θ, f).

Therefore (25) is $o_p(1)$ if

$$\sup_{\theta_\lambda \in \Theta_\lambda} J^{-1} \sum_{j=1}^{J} (\delta_j^*(\cdot, \theta_\lambda, s^n, P^R) - \delta_j^*(\cdot, \theta_\lambda, s^0, P^0))^2 = o_p(1). \tag{26}$$

This in turn implies (21) immediately and also implies (24) by the triangle inequality as we argue below. Let $Q_J(\delta^*(\cdot, \theta_\lambda^{(1)}, s^n, P^R), z, p, \hat{\mathbf{V}}; \theta^{(1)}, f^{(1)}) = \operatorname{arginf}_{(\theta, \theta_\lambda, f) \in \Theta \times \Theta_\lambda \times \mathcal{F}_J} Q_J(\delta^*(\cdot, \theta_\lambda, s^n, P^R), z, p, \hat{\mathbf{V}}; \theta, f)$ and $Q_J(\delta^*(\cdot, \theta_\lambda^{(2)}, s^0, P^0), z, p, \hat{\mathbf{V}}; \theta^{(2)}, f^{(2)}) = \operatorname{arginf}_{(\theta, \theta_\lambda, f) \in \Theta \times \Theta_\lambda \times \mathcal{F}_J} Q_J(\delta^*(\cdot, \theta_\lambda, s^0, P^0), z, p, \hat{\mathbf{V}}; \theta, f)$. The minimizers $(\theta^{(1)}, \theta_\lambda^{(1)}, f^{(1)})$ and $(\theta^{(2)}, \theta_\lambda^{(2)}, f^{(2)})$ exist because $Q_J(\cdot)$ is continuous in $(\theta, \theta_\lambda, f)$

and the parameter space $\Theta \times \Theta_\lambda \times \mathcal{F}_J$ is compact (Assumption 16). It follows that

$$
\begin{aligned}
o_p(1) &= Q_J(\delta^*(\cdot, \theta_\lambda^{(1)}, s^n, P^R), z, p, \hat{\mathbf{V}}; \theta^{(1)}, f^{(1)}) - Q_J(\delta^*(\cdot, \theta_\lambda^{(1)}, s^0, P^0), z, p, \hat{\mathbf{V}}; \theta^{(1)}, f^{(1)}) \\
&\leq Q_J(\delta^*(\cdot, \theta_\lambda^{(1)}, s^n, P^R), z, p, \hat{\mathbf{V}}; \theta^{(1)}, f^{(1)}) - Q_J(\delta^*(\cdot, \theta_\lambda^{(2)}, s^0, P^0), z, p, \hat{\mathbf{V}}; \theta^{(2)}, f^{(2)}) \\
&\leq Q_J(\delta^*(\cdot, \theta_\lambda^{(2)}, s^n, P^R), z, p, \hat{\mathbf{V}}; \theta^{(2)}, f^{(2)}) - Q_J(\delta^*(\cdot, \theta_\lambda^{(2)}, s^0, P^0), z, p, \hat{\mathbf{V}}; \theta^{(2)}, f^{(2)}) = o_p(1)
\end{aligned}
$$

where the first and the last equality hold by (25) and (26). The first inequality holds because $(\theta^{(2)}, \theta_\lambda^{(2)}, f^{(2)})$ minimizes $Q_J(\delta^*(\cdot, \theta_\lambda, s^0, P^0), z, p, \hat{\mathbf{V}}; \theta, f)$ over $\Theta \times \Theta_\lambda \times \mathcal{F}_J$ and the second inequality holds because $(\theta^{(1)}, \theta_\lambda^{(1)}, f^{(1)})$ minimizes $Q_J(\delta^*(\cdot, \theta_\lambda, s^n, P^R), z, p, \hat{\mathbf{V}}; \theta, f)$ over $\Theta \times \Theta_\lambda \times \mathcal{F}_J$. This proves (24).

Finally note that (26) is $o_p(1)$ by the essentially same proof (page 647-648) in the proof of Theorem 1 of Berry, Linton, and Pakes (2004) under Assumption 11 and Assumption 17 because (i) all arguments there in terms of ξ also hold in terms of our δ^* and (ii) Assumption 11 replaces their Assumption A3 and Assumption 17 replaces their Assumption A5.

This completes the proof.

References

Bajari, P., and L. Benkard (2005): "Demand Estimation with Heterogeneous Consumers and Unobserved Product Characteristics: A Hedonic Approach," *Journal of Political Economy*, 113, 1239–1276.

Berry, S. (1994): "Estimating Discrete-Choice Models of Product Differentiation," *The RAND Journal of Economics*, 25(2), 242–262.

Berry, S., and P. Haile (2010): "Identification in Differentiated Products Markets Using Market Level Data," *NBER working paper*.

Berry, S., J. Levinsohn, and A. Pakes (1995): "Automobile Prices in Market Equilibrium," *Econometrica*, 60(4), 889–917.

Berry, S., O. Linton, and A. Pakes (2004): "Limit Theorems for Estimating the Parameters of Differentiated Product Demand Systems," *Review of Economic Studies*, 71, 613–654.

Blundell, R., and J. Powell (2003): "Endogeneity in Nonparametric and Semiparametric Regression Models," *Advances in Economics and Econometrics:Theory and Applications, Eighth World Congress*.

Chen, X. (2006): "Large Sample Sieve Estimation of Semi-Nonparametric Models," *forthcoming in Handbook of Econometrics, Volume 6*.

Chen, X., O. Linton, and I. van Keilegom (2003): "Estimation of Semiparametric Models When the Criterion Function is Not Smooth," *Econometrica*, 71, 1591–1608.

Chesher, A., A. Rosen, and K. Smolinski (2011): "An Instrumental Variable Model of Multiple Discrete Choice," Working paper.

Dieudonne, J. (1969): *Foundations of Modern Analysis*, New York: Academic Press.

Florens, J., J. Heckman, C. Meghir, and E. Vytlacil (2008): "Identification of Treatment Effects Using Control Functions in Models with Continuous Endogenous Treatment and Heterogeneous Effects," *Econometrica*, 76, 1191–1206.

Gandhi, A. (2009): "Inverting Demand Systems," Working paper.

Hahn, J., and G. Ridder (2008): "Conditional Moment Restrictions and Triangular Simultaneous Equations," Working Paper.

Imbens, G., and W. Newey (2003): "Identification and Estimation of Triangular Simultaneous Equations Models without Additivity," Working Paper.

Imbens, G., and W. Newey (2009): "Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity," *Econometrica*, 77, 1481–1512.

KIM, K., AND A. PETRIN (2010a): "Control Function Corrections for Unobserved Factors in Differentiated Product Models," Working Paper.

——— (2010b): "Control Functions for a Class of Non-Separable Models without Independence and Monotonicity Conditions," Working paper.

——— (2010c): "Multiplicative Heterogeneity Models with Endogeneity," Working paper.

——— (2010d): "Revisiting Instrumental Variables and the Classic Control Function Approach, with Implications for Parametric and Non-Parametric Regressions," Working paper.

LEWBEL, A. (2000): "Semiparametric Qualitative Response Model Estimation With Instrumental Variables and Unknown Heteroscedasticity," *Journal of Econometrics*, 97, 145–177.

MATZKIN, R. (2003): "Nonparametric Estimation of Nonadditive Random Functions," *Econometrica*, 71, 1339–1375.

MCFADDEN, D. (1981): "Structural discrete probability models derived from theories of choice," *Structural Analysis of Discrete Data and Econometric Applications*, pp. 115–178, Edited by Charles Manski and Daniel McFadden, The MIT press.

NEWEY, W., AND J. POWELL (2003): "Instrumental Variable Estimation of Nonparametric Models," *Econometrica*, 71, 1565–1578.

PAKES, A. (1996): "Dynamic structural models, problems and prospects: mixed continuous discrete controls and market interaction," in *Advances in Econometrics, Sixth World Congress, Volume II*, ed. by C. Sims, pp. 171–259, New York. Cambridge.

ROTHENBERG, T. (1971): "Identification in Parametric Models," *Econometrica*, 39, 577–591.