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 October 2012

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 "generalized" control functions and we show together they are sufficient for identification of all of the demand parameters. Our estimator extends (Berry, Linton, and Pakes, 2004) to the case where there are estimated regressors. 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.

*We thank Arthur Lewbel, Rosa Matzkin, and numerous seminar participants for many helpful suggestions. This paper previously circulated under the title "The Interaction of Observed and Unobserved Demand Factors in Discrete Choice Demand Models."

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 variable 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 *generalized* control functions 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 out the correlation between the unobserved demand factor and price. We then construct the new

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

moments conditions based on a specification that includes the control function as an additional 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.

In Section 6 we provide a proof of consistency of a sieve estimator with estimated regressors which have error that goes to zero as the sample size increases. The proof covers the case when the asymptotics are the number of products, and a special case is when the asymptotics are in the number of markets, as in Goolsbee and Petrin (2004). When the asymptotics are in the number of products Berry, Linton, and Pakes (2004) argue against maintaining uniform convergence of the objective function because shares and prices are equilibrium outcomes of strategically interacting firms. This interdependence generates conditional dependence in the estimate of the demand error when the parameter value is different from the truth, making it difficult to determine how the objective function behaves away from the true parameter value.

Berry, Linton, and Pakes (2004) show how to achieve identification without maintaining uniform convergence and we show how to extend the Berry, Linton, and Pakes (2004) consistency theorem to the case of our estimator. Our estimator must allow for the new approximation errors arising from pre-step estimators in addition to the sampling and simulation error present in Berry, Linton, and Pakes (2004). A strength of our approach is that it does not require us to find more instruments than are necessary in the separable setting. Just as in Berry/BLP, if price is the only endogenous variable then we only require one variable that moves price around and is excluded from utility. The cost of our approach is that we must be able to estimate the new control functions consistently.

An important difference between our approach and Berry, Linton, and Pakes (2004) is that we can somewhat weaken the invertibility assumption. When the demand error is additively separable inverting the market shares to recover mean utility is isomorphic to inverting the shares to recover demand error. When it is not separable these inversions are no longer isomorphic. In our case we only require invertibility of the vector of market shares in mean utility and not in the stronger requirement of invertibility in demand error. One implication is that we only require monotonicity in the own mean utility term for each product and not in the demand error, which means that we do not need to place restrictions on the signs of the utility parameters related to the interaction terms between the regressors and the demand error to ensure invertibility in the demand error.

Our estimation approach is straightforward. In a setting without random coefficients our estimator inverts market shares to recover mean utility and then reduces to three simple steps, which are at the least repeated least squares. 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 simple steps where in the last step we use a minimum distance estimation instead of least squares.[2]

In Section 8 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

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

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 9). 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 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

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

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

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

where $d_i = (d_{i1}, \ldots, d_{iR})$ 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. $\tau_{rk} d_{ir}$ is then the marginal utility derived from a unit of the kth characteristic for a consumer with demographic d_{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 distribution function and assuming ϵ_{ij} is independent and identically distributed extreme value, the market share of product j is

$$s_j(\delta) = \int \frac{e^{\delta_j + \mu_{ij}}}{\sum_{k=0}^{J} e^{\delta_k + \mu_{ik}}} dF(\mu).$$

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, $\overline{y}$ 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 $\sum_{k=1}^{K}\gamma_k x_{jk}\xi_j$ 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 θ 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.

[4]If we allow the interaction term with residual income - $(y_i - p_j)$ instead of $(\overline{y}-p_j)$ - 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.

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

$$E[\xi_j|z_j] = E\left[\frac{\delta_j - c_0 - \beta_0' X_j + \alpha_0 p_j}{1 + \gamma_0' X_j + \gamma_{p0}(\bar{y} - p_j)}|z_j\right] = 0. \tag{5}$$

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. Objective functions constructed based on these moment conditions (e.g. GMM) will violate the properness condition introduced by Palais (1959) for identification. The properness condition requires the objective function should diverge to infinity when each parameter tends to infinity. Objective functions based on (5) will tend to zero (e.g.) when any γ_{k0} is sent to infinity.

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 Removing Endogeneity with Control Functions

We add a control functions condition 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 $\mathbf{v}_j = p_j$ 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 $\mathbf{v}_j \neq p_j$ 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. We formalize this argument in our proof of identification for the logit case in Section 5.

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

□

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

$$v_j \quad = \quad p_j - E[p_j | z_j] = p_j - \Pi(z_j), j = 1, \ldots, J, \tag{6}$$

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) (also see Florens, Heckman, Meghir, and Vytlacil (2008)) as an alternative way to generate v_j. The controls for good j are then given by $\mathbf{v}_j = \mathfrak{V}_j(v_1, \ldots, v_J)$, for some known (vector) function $\mathfrak{V}_j(\cdot)$ 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 can be 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 $\mathbf{v}_j = \mathfrak{V}_j(v_1, \ldots, v_J)$, we can then exploit the moment condition:

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

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 (7) thus allows us to circumvent the problem of specifying the exact relationship between p_j and ξ_j.

In the logit case without random coefficients the structural parameters would all be identified from (7) 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. This is our key point of departure from conventional control function approaches (e.g. Newey, Powell, and Vella (1999) and Florens, Heckman, Meghir, and Vytlacil (2008)) that further assume

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

and therefore as long as $\mathbf{v}_j$ is measurably separated from p_j and z_j, then identification holds (i.e. there does not exist linear functional relationship between 1, x_j, p_j, $f(\mathbf{v}_j)$, $f(\mathbf{v}_j)x_j$, and $f(\mathbf{v}_j)(\bar{y} - p_j))$. However, we note that the assumption $E[\xi_j|z_j, \mathbf{v}_j] = E[\xi_j|\mathbf{v}_j]$ can become very restrictive because it essentially assumes we can recover the demand errors from pricing equations, which means we should know the true pricing functions. Instead we resort to the instrumental variables condition below for our identification while allowing $E[\xi_j|z_j, \mathbf{v}_j]$ can depend on z_j, i.e. $\mathbf{v}_j$ may not be a *perfect* control as in conventional control function methods.

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 a 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 $(x_j,\ p_j)$. Thus the generalized control function 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 in the logit case.

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

$$0 = E\left[\delta_j - \{c_0 + \beta_0' X_j - \alpha_0 p_j + f(Z_j, \mathbf{V}_j)(1 + \gamma_0' X_j + \gamma_{p0}(\bar{y} - p_j))\} | z_j, \mathbf{v}_j\right] \tag{8}$$

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

$$E[f(Z_j, \mathbf{V}_j)|z_j] = 0. \tag{9}$$

We use a multi-step least squares estimator (for the logit case) or minimum distance estimator (for the random coefficient logit case, see Section 6) based on the moment conditions from (8) and (9) 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 $\mathbf{v}_j = \mathfrak{V}_j(v_1, \ldots, v_J)$ 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 (9). 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 either non-linear least squares or minimum distance estimation.[5] We formally develop this estimator in Section 6 (also see Appendix B for the logit case).

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

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

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

$$\begin{aligned} f(z_j, \mathbf{v}_j) &= f(z_j, \mathbf{v}_j) - E[f(Z_j, \mathbf{V}_j)|z_j] \\ &= (\pi_0 + \pi_1' z_j + \pi_2 v_j + \pi_3' z_j\, v_j) - (\pi_0 + \pi_1' z_j + \pi_2 E[V_j|z_j] + \pi_3' z_j E[V_j|z_j]) \\ &= \pi_2 v_j + \pi_3' z_j\, v_j, \end{aligned} \tag{10}$$

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

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

z_j, but conditional on these terms is not an additive function of p_j nor z_j alone.

Identification in the logit case follows from plugging (10) in (8) 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 from 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 the demand parameters except the interaction parameters γ's because the CMR implies

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

Given the identified demand parameters, 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

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

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 $p_j = F^{-1}_{p_j|z_j}(\tilde{v}_j|z_j) \equiv \tilde{h}(z_j, \tilde{v}_j)$. One can then proceed as described above constructing $\tilde{\mathbf{v}}_j = \tilde{\mathfrak{V}}_j(\tilde{V}_1, \ldots, \tilde{V}_J)$. Identification also holds for this $\tilde{\mathbf{v}}_j$ (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 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

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

[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).

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

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

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

5 Identification in the Simple Logit Model

In this section we show global identification for the model with the simple logit error only to convey the intuition for this base case. An Appendix provides the proof of consistency for the sieve estimator in the simple logit case. In Section 6 we consider identification and estimation for the random coefficients setup, providing conditions under which our sieve estimator is consistent.

We study identification using the moment conditions (8) and (9). 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 $\mathbf{v}_j = \mathfrak{V}_j(p_1 - \Pi(z_1), \ldots, p_J - \Pi(z_J)) = \mathfrak{V}_j(v_1, \ldots, v_J)$ and also write $\mathbf{v}_j = \mathfrak{V}_j(v_j, v_{-j})$ where v_{-j} denotes a vector obtained by removing v_j from $(v_1, \ldots, v_J)$.[8] $\mathbf{v}_j$ is identified from the first step regression of (6), and we treat $\Pi(z_j)$ and $\mathbf{v}_j$ as known throughout the discussion.[9]

Indeed our identification argument below is not specific to a particular functional form like (2) but still holds when the linear utility term $c_0 + \beta_0' x_j - \alpha_0 p_j$ is replaced with a nonparametric function of x_j and p_j, say $\phi_0(x_j, p_j)$. Below we present our identification result for this nonparametric case for possible generalizations.

For our identification result the key assumption we require is that a control function exists, which satisfies the property that even conditional on this functional, price still has variation. Let $z_j = (x_j', z_{2j}')'$.

Assumption 1. *[CF2] $E[\xi_j \mid z_j, \mathbf{v}_j] = f(z_j, \mathbf{v}_j) \equiv f(\tilde{v}_j(z_i, \mathbf{v}_j))$ where the control function variables $\tilde{v}_j$ (assumed to be smooth) satisfies the property that for any (z_j^*, v_j^*, v_{-j}^*) in the support of (z_j, v_j, v_{-j}), (assume the support is an open set) there exists an implicit function $v_{-j}^*(z_{2j})$ such that $\tilde{v}_j\left(x_j^*, z_{2j}, \mathfrak{V}_j(v_j^*, v_{-j}^*(z_{2j}))\right) = \tilde{v}_j\left(x_j^*, z_{2j}^*, \mathfrak{V}_j(v_j^*, v_{-j}^*)\right)$ for all z_{2j} in a neighborhood of z_{2j}^*.*

Note that this assumption (that strengthens CF condition) on the existence of a control function strictly generalizes the standard control function condition required by Newey, Powell, and Vella (1999) for the case of separable models and Florens, Heckman, Meghir, and Vytlacil (2008) for non-separable models, which is that $E[\xi_j \mid x_j, p_j, \mathbf{v}_j] = E[\xi_j \mid z_j, \mathbf{v}_j] = f(\mathbf{v}_j)$. Observe that the standard control function restriction that z_j "drops out" of f conditional on $\mathbf{v}_j$ would trivially satisfy our more general condition (i.e. we take $\tilde{v}_j = \mathbf{v}_j$). We only require that the control function depend on a subset of the full information contained in (z_j, v_j, v_{-j}).

[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)).

The next assumption we make, which closes the gap caused by our generalization, is the instrumental variable assumption (CMR) used for identification of nonparametric separable models with endogeneity (see e.g. Newey and Powell (2003) and Hall and Horowitz (2005) among many others).

Assumption 2. *[CMR2] $E[\xi_j \mid z_j] = 0$ and (p_j, z_j) satisfy the completeness condition that 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.*

Below we can use these assumptions to prove identification of the parameters $\theta_0 = (\phi_0, \gamma_0, \gamma_{p0})$ and f_0. Note that if θ_0 and f_0 are identified they must be the unique solution to

$$0 = E\left[\delta_j - \{\phi_0(X_j, p_j) + f(Z_j, \mathbf{V}_j)(1 + \gamma_0' X_j + \gamma_{p0}(\bar{y} - p_j))\}|z_j, \mathbf{v}_j\right] \tag{11}$$

and (9). The conditional expectation $E[\delta_j|z_j, \mathbf{v}_j]$ is unique with probability one, which implies if there exists any other function $\bar{\theta}$ and $\bar{f}$ that satisfies (11) and (9) it must be that

$$\Pr\{\phi_0(x_j,p_j) + f_0(z_j,\mathbf{v}_j)(1+\gamma_0' x_j + \gamma_{p0}(\bar{y}-p_j)) = \bar{\phi}(x_j,p_j) + \bar{f}(z_j,\mathbf{v}_j)(1+\bar{\gamma}' x_j + \bar{\gamma}_p(\bar{y}-p_j))\} = 1. \tag{12}$$

Therefore, identification means we must have $\phi_0 = \bar{\phi}$, $\gamma_0 = \bar{\gamma}$, $\gamma_{p0} = \bar{\gamma}_p$, $f_0 = \bar{f}$ with probability one whenever (12) holds. Then working with differences $\psi(x_j, p_j) = \bar{\phi}(x_j, p_j) - \phi_0(x_j, p_j)$, $\kappa(z_j, \mathbf{v}_j) = \bar{f}(z_j, \mathbf{v}_j) - f_0(z_j, \mathbf{v}_j)$, $\kappa_x(z_j, \mathbf{v}_j) = \bar{\gamma}\bar{f}(z_j, \mathbf{v}_j) - \gamma_0 f_0(z_j, \mathbf{v}_j)$, and $\kappa_p(z_j, \mathbf{v}_j) = \bar{\gamma}_p\bar{f}(z_j, \mathbf{v}_j) - \gamma_{p0} f_0(z_j, \mathbf{v}_j)$ we can write (12) as

$$\Pr\{\psi(x_j, p_j) + \kappa(z_j, \mathbf{v}_j) + \kappa_x'(z_j, \mathbf{v}_j)x_j + \kappa_p(z_j, \mathbf{v}_j)(\bar{y} - p_j) = 0\} = 1. \tag{13}$$

If (13) holds, for identification we must have $\psi(x_j, p_j) = 0$, $\kappa(z_j, \mathbf{v}_j) = 0$, $\kappa_x(z_j, \mathbf{v}_j) = 0$, and $\kappa_p(z_j, \mathbf{v}_j) = 0$ with probability one. We formalize this identification statement in Theorem 2.

Theorem 2 (Identification)**.** *Let*

$$\Psi(\psi, \kappa, \kappa_x, \kappa_p) = \psi(x_j, p_j) + \kappa(z_j, \mathbf{v}_j) + \kappa_x'(z_j, \mathbf{v}_j)x_j + \kappa_p(z_j, \mathbf{v}_j)(\bar{y} - p_j),$$

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

$$\Pr\{\Psi(\psi, \kappa, \kappa_x, \kappa_p) = 0\} = 1 \tag{14}$$

then the structural parameters $\theta_0 = (\phi_0, \gamma_0, \gamma_{p0})$ and f_0 are identified.

Proof. The CF condition allows one to have the moment condition (11) (and thus equation (14)). If there exists an additive functional relationship between $\psi(x_j, p_j)$, $\kappa(z_j, \mathbf{v}_j)$, $x_{j1}\kappa_{x_1}(z_j, \mathbf{v}_j)$, $\ldots$, $x_{jK}\kappa_{x_K}(z_j, \mathbf{v}_j)$, and $(\bar{y} - p_j)\kappa_p(z_j, \mathbf{v}_j)$ then (14) must be satisfied. The contrapositive argument proves the statement. $\square$

We now use Theorem 2 to show identification under Assumptions CF2 and CMR2 when $\Pi(z_j)$ and $f(z_j, \mathbf{v}_j)$ are differentiable.[10]

[10] We also maintain that the one-sided derivatives of $\Psi(\psi, \kappa, \kappa_x, \kappa_p)$ 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).

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 CF2 and CMR2 conditions hold. Then θ_0 and f_0 are identified.*

Proof. Suppose two sets of parameters $(\phi_0, \gamma_0, \gamma_{p0}, f_0)$ and $(\phi, \gamma, \gamma_p, f)$ both explain the same conditional expectation (11). Then we must have that

$$\psi(x_j, p_j) + \kappa(z_j, \mathbf{v}_j) + \kappa_x'(z_j, \mathbf{v}_j)x_j + \kappa_p(z_j, \mathbf{v}_j)(\bar{y} - p_j) = 0 \tag{15}$$

a.e., (z_j, v_j, v_{-j}). For identification we need to show only $\psi(x_j, p_j) = 0$, $\gamma = \gamma^0$, $\gamma_p = \gamma_p^0$ and $\kappa(z_j, \mathbf{v}_j) = 0$ satisfies (15).

In particular then from (15) we must have that for each (z_j^*, v_j^*, v_{-j}^*) in the support,

$$\frac{\partial}{\partial z_{2j}} \left[\phi\left(x_j^*, p_j^*\right) + f(x_j^*, z_{2j}, \mathfrak{V}_j(v_j^*, v_{-j}^*(z_{2j})))\{1 + \gamma' x_j^* + \gamma_p(\bar{y} - p_j^*)\}\right]\Big|_{z_{2j}=z_{2j}^*} =$$
$$\frac{\partial}{\partial z_{2j}} \left[\phi_0\left(x_j^*, p_j^*\right) + f_0(x_j^*, z_{2j}, \mathfrak{V}_j(v_j^*, v_{-j}^*(z_{2j})))\{1 + \gamma_0' x_j^* + \gamma_{p0}(\bar{y} - p_j^*)\}\right]\Big|_{z_{2j}=z_{2j}^*},$$

where v_{-j}^* is the implicit function satisfying the property stated in Assumption CF2. Taking the derivative through the expression gives (by the chain rule)

$$\frac{\partial \Pi(z_j^*)}{\partial z_{2j}}\{\frac{\partial \psi(x_j^*, p_j^*)}{\partial p_j} - \kappa_p(z_j^*, \mathfrak{V}_j(v_j^*, v_{-j}^*))\} = 0$$

because $\frac{\partial f(x_j^*, z_{2j}, \mathfrak{V}_j(v_j^*, v_{-j}^*(z_{2j})))}{\partial z_{2j}} = 0$ in the neighborhood of z_{2j}^* by Assumption CF2 (i.e. we can fix $\tilde{v}_j$ while z_{2j} varies around z_{2j}^*) and because $p_j = \Pi(z_j) + v_j$. Then because $\frac{\partial \Pi(z_j^*)}{\partial z_{2j}} \neq 0$ by the completeness condition (i.e. instruments should satisfy the rank condition), we also have $\frac{\partial \psi(x_j^*, p_j^*)}{\partial p_j} - \kappa_p(z_j^*, \mathfrak{V}_j(v_j^*, v_{-j}^*)) = 0$. Then since the above equality holds for all (z_j^*, v_j^*, v_{-j}^*), we thus have for all (z_j, v_j, v_{-j}) that

$$\frac{\partial}{\partial p_j}\psi\left(x_j, p_j\right) - \kappa_p\left(z_j, \mathfrak{V}_j(v_j, v_{-j})\right) = 0. \tag{16}$$

Note that $E[\kappa_p\left(z_j, \mathbf{v}_j\right)|z_j] = 0$ because we can restrict our attention to candidate functions f and f_0 that satisfies $E[f(z_j, \mathbf{v}_j)|z_j] = 0$ and $E[f_0(z_j, \mathbf{v}_j)|z_j] = 0$ due to the law of iterated expectation and Assumption CMR2:

$$0 = E[\xi_j|z_j] = E[E[\xi_j|Z_j, V_j, V_{-j}]|z_j] = E[f\left(Z_j, \mathbf{V}_j\right)|z_j].$$

Then taking the conditional expectation to (16) observe that we can exploit the CMR to transform the equality as

$$E\left[\frac{\partial}{\partial p_j}\psi\left(x_j, p_j\right)|z_j\right] = 0$$

and thus by the completeness condition we have $\frac{\partial}{\partial p_j}\psi\left(x_j, p_j\right) = 0$. This implies by (16) that

$\kappa_p = \gamma_p f - \gamma_{p0} f_0 = 0$. This in turn implies by (15) that

$$\psi\left(x_j, p_j\right) + \kappa(z_j, \mathbf{v}_j) + \kappa_x'(z_j, \mathbf{v}_j) x_j = 0. \tag{17}$$

Then taking the conditional expectation to the above and applying the CMR condition, we find

$$E[\psi(x_j, p_j)|z_j] = 0$$

and thus by the completeness condition now we have $\phi(x_j, p_j) = \phi_0(x_j, p_j)$. This in turn implies by (17) $\kappa(z_j, \mathbf{v}_j) + \kappa_x'(z_j, \mathbf{v}_j)x_j = 0$, which implies (after dividing it by f_0 and multiplying it by γ_p) that $\gamma_{p0} - \gamma_p + (\gamma_{p0}\gamma - \gamma_p\gamma_0)' x_j = 0$ a.e. in x_j. Then it must be that $\gamma_{p0} = \gamma_p$ and $\gamma_0 = \gamma$, which also implies with $\kappa_p(z_j, \mathbf{v}_j) = 0$, $\kappa(z_j, \mathbf{v}_j) = 0$ and $\kappa_x(z_j, \mathbf{v}_j) = 0$.

We therefore have shown there do not exist two distinct tuples of $(\phi, \gamma, \gamma_p, f)$ that solves (11), hence identification. $\square$

6 Identification and Estimation in the Random Coefficients Model

In this section we formally develop our estimator as a sieve estimator and provide a proof of its consistency. The proof covers the case when the asymptotics are the number of products, as in Berry, Linton, and Pakes (2004) and the automobile data from Berry, Levinsohn, and Pakes (1995).[11] A special case is when the asymptotics are in the number of markets, as in Goolsbee and Petrin (2004) or Chintagunta, Dube, and Goh (2005).[12]

When the asymptotics are in the number of products Berry, Linton, and Pakes (2004) argue against maintaining uniform convergence of the objective function. The issue is that shares and prices are equilibrium outcomes of strategically interacting firms who observe the characteristics of all products in the market. This interdependence generates conditional dependence in the estimate of ξ when the parameter value is different from the truth, making it difficult to determine how the objective function behaves away from the true parameter value.

Berry, Linton, and Pakes (2004) show how to achieve identification without maintaining uniform convergence and we show how to extend the Berry, Linton, and Pakes (2004) consistency theorem to the case of our estimator. Our estimator must allow for the new approximation errors arising from pre-step estimators in addition to the sampling and simulation error present in Berry, Linton, and Pakes (2004). With the asymptotics in products it is no longer possible to allow the control function to vary by product, although it can vary by (e.g.) product type (stylish or not, large or small) or any other observed factor that is fixed as the number of products increases.

When the asymptotics are in the number of markets our consistency proof extends Chen (2007) (Section 3.1) to a setting with pre-step estimators. Under standard regularity conditions the sample objective function converges uniformly to its population counterpart making consistency straightforward to establish. The control functions can vary by product, or by product-season (e.g.), or by

[11] In the BLP data the number of products per market is over 100 and the number of markets is 20.

[12] In the former paper there are four television viewing options in every market and over 300 television markets determining by the cable providers. In the latter there are a small number of orange juices or margarines for whom sales are observed at a particular supermarket over 100 weeks, with the week being the market.

other observed factors as long as the number of control functions is not increasing as the sample size increases.

An important difference between our approach and Berry, Linton, and Pakes (2004) is that we can weaken the invertibility assumption. When the demand error is additively separable inverting the market shares to recover δ is isomorphic to inverting the shares to recover ξ. When it is not separable these inversions are no longer isomorphic. In our case we only require invertibility of the vector of market shares in δ and not in the stronger requirement of invertibility in ξ. One implication is that we only require monotonicity in the own mean utility term δ_j and not in demand error ξ_j, which means that we do not need to place restrictions on the signs of the utility parameters related to the interaction terms between the regressors and ξ_j to ensure ξ is unique and thus invertible.

6.1 Setup and Estimation

For transparency we assume the data is from a single market $M = 1$. We let $\nu(x, p, \xi, \lambda, \theta, \theta_\lambda)$ be a $J \times 1$ share function specific to a household type λ and let $P(\lambda)$ be the distribution of λ where θ denotes the mean utility parameters and θ_λ denotes the distribution parameters. Given a choice set with characteristics (x,p, ξ) the vector of aggregate market shares at values of (θ, θ_λ) is given by

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

where ξ appears only in mean utility because it does not have a random coefficient.[13] 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 population market shares s^0 are given by evaluating $\sigma(\delta(\cdot, \theta), \theta_\lambda, P)$ at (θ_0,$\theta_{\lambda 0}$,P^0), the true values of θ,θ_λ, and P. Also under Assumption 4 below the share equation is invertible, so there exists unique $\delta^* = \delta^*(\theta_{\lambda 0}, s^0, P^0)$ ($J \times 1$ vector) that solves the share equation

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

Berry, Linton, and Pakes (2004) treat two sources of error and we follow their approach. One source of error arises because of the use of simulation to approximate P^0 with P^R, the empirical measure of some i.i.d. sample $\lambda_1, \ldots, \lambda_R$ from $P(\lambda)$:

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

The second source of error is the sampling error in observed market shares s^n which are constructed from n i.i.d. draws from the population of consumers.

[13] Allowing for a random coefficient on ξ is an unresolved problem to date.

Our estimation approach is as follows. 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 $\hat{\mathbf{v}}_j = \mathfrak{V}_j(\hat{v}_1, \ldots, \hat{v}_J)$ where how to construct control variates, $\mathfrak{V}_j(\cdot)$ is up to researchers (so treated as known). 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 (9). In the final step we estimate $(\theta_0, \theta_{\lambda 0})$ and $f_0(\cdot)$ using a sieve estimation.

We incorporate the pre-step estimation error by letting $\mathcal{F}$ denote a space of functions that includes the true function f_0, endowed with $\|\cdot\|_{\mathcal{F}}$ a pseudo-metric on $\mathcal{F}$. We write the basis functions for $f(\cdot)$ as

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

where $\bar{\varphi}_l(z_j) = E[\varphi_l(Z_j, \mathbf{V}_j)|z_j]$ and $\{\varphi_l(z_j, \mathbf{v}_j),\ l = 1, 2, \ldots\}$ denotes a sequence of approximating basis functions of $(\mathbf{v}_j, z_j)$ such as power series or splines. Subtracting out the conditional means from the underlying basis functions ensures that any function $f(\cdot)$ in the sieve space satisfies the conditional moment restrictions from (9).

Define the (infeasible) sieve space $\mathcal{F}_J$ as the collection of functions

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

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

$$\hat{\mathcal{F}}_J = \{f : f = \sum_{l \le L(J)} a_l \hat{\tilde{\varphi}}_l(\cdot, \cdot), \|f\|_{\mathcal{F}} < \bar{C}\}. \tag{18}$$

Note that under mild regularity conditions with specific series estimations considered below, $\hat{\mathcal{F}}_J$ well approximates $\mathcal{F}_J$ (in a metric defined on the metric space $(\mathcal{F}, \|\cdot\|_{\mathcal{F}})$) in the sense that for any $f \in \mathcal{F}_J$ we can find a sequence $\hat{f} \in \hat{\mathcal{F}}_J$ such that $\left\|\hat{f} - f\right\|_{\mathcal{F}} \to 0$ as $\hat{\Pi}(\cdot) \to \Pi(\cdot)$ and $\hat{\bar{\varphi}}_l(\cdot) \to \bar{\varphi}_l(\cdot)$ (in a pseudo-metric $\|\cdot\|_s$).

To provide details in estimation suppose triangular array of data of the tuple $\{p_j, x_j, z_j\}_{j=1}^J$ are available. Let $\{\varphi_l(Z), l = 1, 2, \ldots\}$ denote a sequence of approximating basis functions (e.g. orthonormal polynomials or splines) of Z . Let $\varphi^{k(J)}(Z) = (\varphi_1(Z), \ldots, \varphi_{k(J)}(Z))'$, $\mathbf{P} = (\varphi^{k(J)}(Z_1), \ldots, \varphi^{k(J)}(Z_J))'$ and $(\mathbf{P}'\mathbf{P})^-$ denote the Moore-Penrose generalized inverse where $k(J)$ tends to infinity but $k(J)/J \to 0$. In our asymptotics later we assume $\varphi^{k(J)}(Z)$ is orthonormalized (see Lemma L1 in Appendix) and hence assume $\mathbf{P}'\mathbf{P}/J$ is nonsingular with probability approaching to one (w.p.a.1).

Then in the first stage we estimate the controls

$$\hat{\Pi}(z) = \varphi^{k(J)}(z)'(\mathbf{P}'\mathbf{P})^- \sum\nolimits_{j=1}^J \varphi^{k(J)}(z_j)p_j,\ \hat{v}_j = p_j - \hat{\Pi}(z_j),\ \text{and } \hat{\mathbf{v}}_j = V_j(\hat{v}_1, \ldots, \hat{v}_J)$$

and in the second stage we obtain the approximation of $f(z,\mathbf{v})$ as

$$\begin{aligned}\hat{f}_{L(J)}(z_j,\hat{\mathbf{v}}_j) &= \sum\nolimits_{l=1}^{L(J)} a_l\{\varphi_l(z_j,\hat{\mathbf{v}}_j)-\hat{\bar{\varphi}}_l(z_j)\} = \sum\nolimits_{l=1}^{L(J)} a_l\{\varphi_l(z_j,\hat{\mathbf{v}}_j)-\hat{E}[\varphi_l(Z_j,\hat{\mathbf{V}}_j)|Z=z_j]\} \\ &= \sum\nolimits_{l=1}^{L(J)} a_l\{\varphi_l(z_j,\hat{\mathbf{v}}_j)-p^{k(J)}(z_j)'(\mathbf{P}'\mathbf{P})^{-}\sum\nolimits_{j'=1}^{J} p^{k(J)}(z_{j'})\varphi_l(z_{j'},\hat{\mathbf{v}}_{j'})\}\end{aligned}$$

where $\{\varphi_l(z,\mathbf{v}), l=1,2,\ldots\}$ denote a sequence of approximating basis functions generated using $(z,\mathbf{v})$. We can use different sieves (e.g., power series, splines of different lengths) to approximate $\bar{\varphi}_l(z_j)=E[\varphi_l(Z_j,\mathbf{V}_j)|Z_j=z_j]$ and $\Pi(z_j)$ depending on their smoothness but we assume one uses the same sieves for ease of notation.

Let $\varphi^L(z_j,\mathbf{v}_j)=(\varphi_1(z_j,\mathbf{v}_j),\ldots,\varphi_L(z_i,\mathbf{v}_j))'$, $\varphi^L(z_j,\hat{\mathbf{v}}_j)=\varphi^L(z_j,\mathbf{v}_j)|_{\mathbf{v}_j=\hat{\mathbf{v}}_j}$, and define for some d_j its empirical conditional mean on $(z_j,\mathbf{v}_j)$ as

$$\hat{E}[d_j|z_j,\mathbf{v}_j]=\varphi^L(z_j,\mathbf{v}_j)'(\sum_{j=1}^{J}\varphi^L(z_j,\mathbf{v}_j)\varphi^L(z_j,\mathbf{v}_j)')^{-1}\sum_{j=1}^{J}\varphi^L(z_j,\mathbf{v}_j)d_j$$

and similarly $\hat{E}[d_j|z_j,\hat{\mathbf{v}}_j]$ where we replace $\mathbf{v}_j$ with $\hat{\mathbf{v}}_j$.

Then based on the moment condition like (7, in the case of fixed coefficients)

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

in the last stage we can estimate the demand parameters using (e.g.) a sieve MD-estimation:[14]

$$\begin{aligned}(\hat{\theta},\hat{\theta}_\lambda,\hat{f}) &= \arg\inf_{(\theta,\theta_\lambda,\hat{f}_{L(J)})\in\Theta\times\Theta_\lambda\times\hat{\mathcal{F}}_J}\frac{1}{J}\sum_{j=1}^{J}\{\hat{E}[\delta_j^*(\theta_\lambda,s^n,P^R)|z_j,\hat{\mathbf{v}}_j] \\ &\quad -(c+\beta' x_j-\alpha p_j+\hat{f}_{L(J)}(z_j,\hat{\mathbf{v}}_j)(1+\gamma' x_j+\gamma_p(\bar{y}-p_j)))\}^2\end{aligned} \tag{19}$$

where $\delta_j^*(\theta_\lambda,s^n,P^n)$ denotes the mean utility of the product j, which is the j-th element of δ^* that solves the share equation

$$s^n=\sigma(\delta^*,\theta_\lambda,P^R).$$

Being abstract from the simulation and the sampling error to approximate the true δ^*, for the consistency of this sieve estimation we need to promise $k(J)$,$L(J)\to\infty$ as $J\to\infty$, so as the sample size gets larger, we need to use more flexible specifications for $\hat{\Pi}(\cdot)$ and $\hat{f}_{L(J)}(\cdot)$. In practice, one can proceed estimation and inference with fixed $k=k(J)$ and $L=L(J)$.

Even though the asymptotics will be different under two different scenarios: parametric model (fixed $k(J)$ and $L(J)$) and semiparametric model (increasing $k(J)$ and $L(J)$), the computed standard errors under two different cases can be numerically identical or equivalent. This equivalence has been established in Ackerberg, Chen, and Hahn (2009) for a class of sieve multi-stage estimators. This suggests that we can ignore the semiparametric nature of the model and proceed both estimation and inference (e.g. calculating standard errors) as if the parametric model is the true model. Therefore one can calculate standard errors using the standard formula for the parametric multi-

[14]One can easily add a weighting function in the objective function to gain efficiency (see e.g. Ai and Chen 2003) but we are abstract from it for ease of notation.

step estimation (e.g. Murphy and Topel (1985) and Newey (1984)) when the simulation and the sampling error are negligible.

In the following sections we establish the consistency of the sieve estimation and the asymptotic normality of the demand parameter estimates in the presence of the simulation and the sampling errors. We now turn to the assumptions.

6.2 Assumptions

Our approach closely follows Berry, Linton, and Pakes (2004). Their Assumption A1 regulates the simulation and sampling errors and we rewrite it replacing ξ with δ throughout.

Assumption 3. *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 $(\delta(x,p,\xi,\theta),\theta_\lambda)$,*

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

where $\varepsilon_{j,r}(\delta(\cdot,\theta),\theta_\lambda)$ is bounded, continuous, and differentiable in $\delta(\cdot)$ and θ_λ. Define the $J\times J$ matrices $V_2 = nE_*[(s^n - s^0)(s^n - s^0)'] = diag[s^0] - s^0 s^{0\prime}$ and $V_3 = RE_*[\{\sigma(\delta(\cdot,\xi,\theta^0),\theta_{\lambda 0},P^R) - \sigma(\delta(\cdot,\xi,\theta_0),\theta_{\lambda 0},P^0)\}\{\sigma(\delta(\cdot,\xi,\theta^0),\theta_{\lambda 0},P^R) - \sigma(\delta(\cdot,\xi,\theta_0),\theta_{\lambda 0},P^0)\}']$, where ξ here are the true values.

Here we let diag$[s]$ denote a diagonal matrix with s on the principal diagonal and E_* denotes expectations w.r.t. the sampling and/or simulation disturbances conditional on product characteristics (x,p,ξ). Their Assumption A2 puts regularity conditions on the market share function that ensure its invertibility in ξ. Our market share function is written in terms of $\delta(\cdot,\theta)$ as $\sigma(\delta(\cdot,\theta),\theta_\lambda,P)$ so our Assumption 2 requires that similar conditions hold in terms of the mean utility .[15]

Assumption 4. *(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, $\frac{\partial\sigma_j(\delta,\theta_\lambda,P)}{\partial\delta_k}$ exists, and is continuously differentiable in both δ and θ_λ, with $\frac{\partial\sigma_j(\delta,\theta_\lambda,P)}{\partial\delta_j} > 0$, and for $k\neq j$, $\frac{\partial\sigma_j(\delta,\theta_\lambda,P)}{\partial\delta_k}\leq 0$. The matrix $\frac{\partial\sigma(\delta,\theta_\lambda,P)}{\partial\delta'}$ 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.*

Under Assumption 2 the mean utility $\delta^* = \delta^*(\theta_\lambda,s,P)$ that solves

$$s - \sigma(\delta^*,\theta_\lambda,P) = 0 \tag{20}$$

is unique so s and δ^* are one-to-one for any θ_λ and P. The true value of δ^* is given as $\delta^{*0} = \delta^*(\theta_{\lambda 0},s^0,P^0)$. By the implicit function theorem, Dieudonne (1969)(Theorem 10.2.1), and Assumption 3 the mapping $\delta^*(\theta_\lambda,s,P)$ is continuously differentiable in (θ_λ,s,P) in some neighborhood. Here note that $\delta^*(\theta_\lambda,s,P)$ denotes the mean utility inverted from the share equations, which depends on the parameter θ_λ but not on θ while we use $\delta(\cdot,\theta)$ to denote the specification of the mean utility as a function of (x,p,ξ) with θ the parameter vector such as in our estimation $\delta_j(\cdot,\theta) = c_0 + \beta_0' x_j - \alpha_0 p_j + \xi_j(1 + \gamma_0' x_j + \gamma_{p0}(\bar{y} - p_j))$.

[15]Note that we only consider the random coefficients logit model while Berry, Linton, and Pakes (2004) is applicable to other models like (e.g.) the vertical model.

As in Berry, Linton, and Pakes (2004) we use Assumption 4 to expand the inverse mapping from (θ_λ, s^n, P) to $\delta^*(\cdot)$ around s^0 to control the sampling error (they do the expansion around ξ^*). We then add a condition that restricts the rate at which s_j^0 approaches zero. It is identical to Condition S in Berry, Linton, and Pakes (2004):

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

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

We turn to developing an analog to Assumption A3 in Berry, Linton, and Pakes (2004). This amounts to controlling the way s^n approaches s^0 and $\sigma(\delta^*(\cdot), \theta_\lambda, P^R)$ approaches to $\sigma(\delta^*(\cdot), \theta_\lambda, P^0)$. We work on the parameter 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. All metrics are in terms of δ instead of ξ. We use the metric $\rho_\delta(\delta^*, \tilde{\delta}^*) = J^{-1} \sum_{j=1}^J (\delta_j^* - \tilde{\delta}_j^*)^2$ and define for each $\epsilon > 0$, the following neighborhoods of θ_0, $\theta_{\lambda 0}$, f_0, P^0, and s^0: $\mathcal{N}_{\theta_0}(\epsilon) = \{\theta : \rho_E(\theta, \theta_0) < \epsilon\}$, $\mathcal{N}_{\theta_{\lambda 0}}(\epsilon) = \{\theta_\lambda : \rho_E(\theta_\lambda, \theta_{\lambda 0}) < \epsilon\}$, $\mathcal{N}_{P^0}(\epsilon) = \{P : \rho_P(P, P^0) < \epsilon\}$, and $\mathcal{N}_{s^0}(\epsilon) = \{s : \rho_{s^0}(s, s^0) < \epsilon\}$. Also for each θ_λ and $\epsilon > 0$, define $\mathcal{N}_{\delta^{*0}}(\theta_\lambda, \epsilon) = \{\delta^* : \rho_\delta(\delta^*, \delta^*(\theta_\lambda, s^0, P^0)) < \epsilon\}$. Assumption 5 is then given as

Assumption 5. *The random sequences* s^n *and* $\sigma^R(\theta_\lambda)$ *are consistent with respect to the corresponding metrics,*

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

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

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

where $\zeta_\varphi(L) = \sup_{z, \mathbf{v}} ||\varphi^L(z, \mathbf{v})||$.

Note that here "L" refers to a size of sieve, so in parametric models (where L is finite) the term $\zeta_\varphi(L)$ in the condition (c) and (d) does not play any role but in semi-nonparametric models like ours (where L grows) the condition (c) and (d) controls the growth of the size of sieve relative to the size of sampling and simulation draws.

For general use of our consistency results that can be applied to other estimation methods, we define our estimator $(\hat{\theta}, \hat{\theta}_\lambda, \hat{f}(\cdot))$ as the value of parameters that minimize a generic sample criterion function

$$(\hat{\theta},\hat{\theta}_\lambda,\hat{f}) = \text{arginf}_{(\theta,\theta_\lambda,f)\in\Theta\times\Theta_\lambda\times\hat{\mathcal{F}}_J} Q_J(\delta^*(\theta_\lambda, s^n, P^R), z, p, \hat{\mathbf{v}}; \theta, f) + o_p(1) \tag{21}$$

and develop consistency results under this generic criterion function.

For the case of the MD estimation as our leading case it is given as

$$\begin{aligned} &Q_J(\delta^*(\theta_\lambda, s, P), \cdot, \mathbf{v}; \theta, f) \equiv \\ &\frac{1}{J}\sum_{j=1}^{J}\{\hat{E}[\delta_j^*(\theta_\lambda, s, P)|z_j, \mathbf{v}_j] - (c + \beta' x_j - \alpha p_j + f(z_j, \mathbf{v}_j)(1 + \gamma' x_j + \gamma_p(\bar{y} - p_j)))\}^2 \end{aligned} \tag{22}$$

although we emphasize that the approach can be applied to more flexible utility specifications. Also define the population criterion function

$$\begin{aligned} & Q_J^0(\delta^*(\theta_\lambda, s, P), \cdot, \mathbf{v}; \theta, f) \\ = \;\; & E[\frac{1}{J}\sum_{j=1}^{J}\{E[\delta_j^*(\theta_\lambda, s, P)|z_j, \mathbf{v}_j] - (c + \beta' x_j - \alpha p_j + f(z_j, \mathbf{v}_j)(1 + \gamma' x_j + \gamma_p(\bar{y} - p_j)))\}^2]. \end{aligned}$$

By construction our estimator is an extremum estimator that satisfies

Assumption 6. *$Q_J(\delta^*(\hat{\theta}_\lambda, s^n, P^R), \cdot, \hat{\mathbf{v}}; \hat{\theta}, \hat{f}) \leq inf_{(\theta,\theta_\lambda,f)\in\Theta\times\Theta_\lambda\times\hat{\mathcal{F}}_J} Q_J(\delta^*(\theta_\lambda, s^n, P^R), \cdot, \hat{\mathbf{v}}; \theta, f) + o_p(1)$.*

Up to this point we have extended several assumptions from Berry, Linton, and Pakes (2004) to our setting but we have not yet added assumptions which ensure consistency in the presence of pre-step estimators. We 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 $\|\cdot\|_s$. The next two assumptions are sufficient for the pre-step estimators to be consistent.

Assumption 7. *$\left\|\hat{\Pi}(\cdot) - \Pi_0(\cdot)\right\|_s = o_p(1)$ and $\left\|\hat{\bar{\varphi}}_l(\cdot) - \bar{\varphi}_{0l}(\cdot)\right\|_s = o_p(1)$ for all l.*

Assumption 7 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 spline approximations 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). We also provide primitive conditions for Assumption 7: consistency and convergence rates of the pre-step estimators in the appendix for both power series and spline approximations (see Assumptions L1).

Assumption 8. *(i) $E[|\delta_j^*(\theta_{\lambda 0}, s, P)|^2|z_j, \mathbf{v}_j]$ is bounded and $\delta_j^*(\theta_\lambda, s, P)$ satisfies a Lipschitz condition such that for a constant $\kappa_\delta \in (0, 1]$ and a measurable function $c(s, P)$ with a bounded second moment $E[c(s, P)^2|z_j, \mathbf{v}_j] < \infty$,*

$$E[|\delta_j^*(\theta_\lambda^1, s, P) - \delta_j^*(\theta_\lambda^2, s, P)|] \leq c(s, P)||\theta_\lambda^1 - \theta_\lambda^2||^{\kappa_\delta}$$

for all s, P and $\theta_\lambda^1, \theta_\lambda^2 \in \Theta_\lambda$ and (ii) $\varphi^L(z_j, \mathbf{v}_j)$ is orthonormalized such that there exists a $C(\epsilon)$ such that $\Pr(||\sum_{j=1}^J \varphi^L(z_j, \mathbf{v}_j)\varphi^L(z_j, \mathbf{v}_j)'/J - I|| > C(\epsilon)) < \epsilon$.

Under Assumption 8 the conditional mean function of $\delta_j^*(\theta_\lambda, s, P)$ on $(z_j, \mathbf{v}_j)$ is well approximated by the sieves and a similar condition is imposed in Ai and Chen (2003) and Newey and Powell (2003). Therefore under Assumptions 7 and 8 we can verify

$$
\begin{aligned}
& \frac{1}{J}\sum_{j=1}^{J}\{\hat{E}[\delta_j^*(\theta_\lambda, s^0, P^0)|z_j, \hat{\mathbf{v}}_j] - E[\delta_j^*(\theta_\lambda, s^0, P^0)|z_j, \mathbf{v}_j]\}^2 \\
\leq \quad & 2\frac{1}{J}\sum_{j=1}^{J}\{\hat{E}[\delta_j^*(\theta_\lambda, s^0, P^0)|z_j, \hat{\mathbf{v}}_j] - E[\delta_j^*(\theta_\lambda, s^0, P^0)|z_j, \hat{\mathbf{v}}_j]\}^2 \\
& +2\frac{1}{J}\sum_{j=1}^{J}\{E[\delta_j^*(\theta_\lambda, s^0, P^0)|z_j, \hat{\mathbf{v}}_j] - E[\delta_j^*(\theta_\lambda, s^0, P^0)|z_j, \mathbf{v}_j]\}^2 = o_p(1)
\end{aligned}
$$

and also $\frac{1}{J}\sum_{j=1}^{J}\{\hat{E}[\delta_j^*(\theta_\lambda, s^n, P^R)|z_j, \hat{\mathbf{v}}_j] - \hat{E}[\delta_j^*(\theta_\lambda, s^0, P^0)|z_j, \hat{\mathbf{v}}_j]\}^2 = o_p(1)$ under Assumptions 5, 7, 8, and 14 below (see Appendix A.1). Therefore it follows that under Assumptions 5, 7, 8, and 14, $E[\delta_j^*(\theta_\lambda, s^0, P^0)|z_j, \mathbf{v}_j]$ is well approximated by $\hat{E}[\delta_j^*(\theta_\lambda, s^n, P^R)|z_j, \hat{\mathbf{v}}_j]$, which is necessary to make the distance between the sample objective $Q_J(\delta^*(\theta_\lambda, s^n, P^R), z, p, \hat{\mathbf{v}}; \theta, f)$ and the population objective $Q_J^0(\delta^*(\theta_\lambda, s^0, P^0), \cdot, \mathbf{v}; \theta, f)$ small enough when J, n, and R are large enough.

Assumption 9. *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 $J \to \infty$.*

Assumption 9 says any f in $\mathcal{F}$ is well approximated by the sieves and this assumption is also known to hold if $\mathcal{F}$ is a set of a class of smooth functions such as Hölder class.

The next assumption ensures 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. For this we need to define the neighborhoods $\mathcal{N}_{f_0,J}(\epsilon) = \{f : ||f - f_0||_{\mathcal{F}} < \epsilon, f \in \mathcal{F}_J\}, \mathcal{N}_{\Pi_0}(\epsilon) = \{\Pi : \|\Pi - \Pi_0\|_s < \epsilon\}$ 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 10. *For any $C > 0$ there exists $\epsilon > 0$ such that*

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

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

The assumptions we have made so far allow us to focus on the behavior of the population objective function on $(\theta, \theta_\lambda, f) \in \Theta \times \Theta_\lambda \times \mathcal{F}$. Our last set of assumptions establish identification. Berry, Linton, and Pakes (2004) Assumption A6 - their identification assumption - requires the objective function evaluated at the true parameter value to be less than the objective function value evaluated at any other parameter value. They show for the Simple Logit model identification reduces to a standard rank condition on the matrix of instrument-regressor moments. We have an analogous result for our setting. For the Random Coefficients Logit case they simply maintain identification as they note further intuition into when identification holds is complicated by the equilibrium nature of the data generating process. We do not have anything to add on the intuition dimension. We do provide a set of conditions under which our estimator satisfies our analogue of their high-level identification condition and is thus consistent as long as Assumptions 3-10 hold.

We start with two assumptions on continuity which are often easy to verify in specific examples.

Assumption 11. *$Q_J^0(\delta^*(\theta_\lambda, s, P), \cdot, \mathbf{v}; \theta, f)$ is continuous in $(\theta, \theta_\lambda, f) \in \Theta \times \Theta_\lambda \times \mathcal{F}$.*

In our example (22) above Assumption 11 is satisfied because $Q_J^0(\delta^*(\theta_\lambda, s, P), \cdot, \mathbf{V}; \theta, f)$ is evidently continuous in (θ, f). Assumption 3 coupled with the implicit function theorem (Dieudonne (1969) Theorem 10.2.1) implies that the mapping $\delta^*(\theta_\lambda, s, P)$ is continuous in θ_λ, and by inspection $Q_J^0(\delta^*(\theta_\lambda, s, P), \cdot, \mathbf{v}; \theta, f)$ is continuous in $\delta^*(\theta_\lambda, s, P)$ so the objective function is also continuous in θ_λ.

Assumption 12. *$Q_J^0(\delta^*(\theta_\lambda, s, P), \cdot, \mathbf{v}; \theta, f_J)$ is continuous in $\Pi(\cdot)$ and $\bar{\varphi}_l(\cdot)$ uniformly for all $(\theta, \theta_\lambda, f_J) \in \Theta \times \Theta_\lambda \times \mathcal{F}_J$.*

Assumption 12 is also satisfied in our example because any $f_J \in \mathcal{F}_J$ is continuous in $\Pi(\cdot)$ and $\bar{\varphi}_l(\cdot)$ by construction of $\mathcal{F}_J$ and because $\Pi(\cdot)$ and $\bar{\varphi}_l(\cdot)$ enter $Q_J^0(\delta^*(\theta_\lambda, s, P), \cdot, \mathbf{v}; \theta, f_J)$ through f_J.

We also maintain the standard regularity condition that our parameter space is compact and we add an assumption that the sieve space for the control function is also compact.

Assumption 13. *The parameter space $\Theta \times \Theta_\lambda$ is compact and the sieve space, $\mathcal{F}_J$, is compact under the pseudo-metric $||\cdot||_{\mathcal{F}}$.*

A sufficient condition for compactness is that the sieve space is based on power series or splines as in our construction (see Chen (2007)).

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 14 below corresponds to Assumption A5 in Berry, Linton, and Pakes (2004) for the logit like case. Therefore this condition combined with Assumption 5 also ensures that $\delta^*(\theta_\lambda, s^0, P^0)$ is well approximated by $\delta^*(\theta_\lambda, s^n, P^R)$ (see discussion in Berry, Linton, and Pakes (2004) p.647 for their proof of their A.2).

Assumption 14. *For all ϵ, there exists $C(\epsilon) > 0$ such that*

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

The last assumption is our version of their identification assumption and it regulates the behavior of the population criterion function as a function of $(\theta, \theta_\lambda, f)$ outside a neighborhood of $(\theta_0, \theta_{\lambda 0}, f_0)$, stating the values must differ by a positive amount in the limit. Note that Assumption 15 below does not require the limit to exist.

Assumption 15. *(i)* $Q_J^0(\delta^*(\theta_{\lambda 0}, s^0, P^0), \cdot, \mathbf{v}; \theta_0, f_0) < \infty$*; (ii) For all* $\epsilon > 0$*, there exists* $C(\epsilon) > 0$ *such that for all* $J \geq \exists J_0$ *large enough*

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

We now state our consistency theorem.

Theorem 4. *Suppose Condition S and Assumptions 3-15 hold for some* $n(J), R(J) \to \infty$*. Then* $\hat{\theta} \to_p \theta_0$ *and* $\hat{\theta}_\lambda \to_p \theta_{\lambda 0}$*.*

7 Asymptotic Normality

We turn to the asymptotic normality. The variance of the estimator can be obtained as the sum of two variance components. One is the variance in the absence of the sampling error in observed shares and simulation error in predicted shares. The second is the variances due to the sampling and simulation error that affect the estimator through the inverted mean utility.

To analyze effects of the sampling error and the simulation error on the variance of the estimator, consider

$$\begin{aligned} \delta^*(\theta_\lambda, s^n, P^R) &= \delta^*(\theta_\lambda, s^0, P^0) \\ &\quad + \{\delta^*(\theta_\lambda, s^n, P^R) - \delta^*(\theta_\lambda, s^0, P^R)\} + \{\delta^*(\theta_\lambda, s^0, P^R) - \delta^*(\theta_\lambda, s^0, P^0)\} \end{aligned} \tag{23}$$

and will find expressions for the last two terms in terms of the sampling and simulation errors. Define the sampling and simulation errors by the $J \times 1$ vectors

$$\varepsilon^n = s^n - s^0 \text{ and } \varepsilon^R(\theta_\lambda) = \sigma^R(\theta_\lambda) - \sigma(\theta_\lambda)$$

where $\sigma^R(\theta_\lambda) = \sigma(\delta^*(\theta_\lambda, s^0, P^0), \theta_\lambda, P^R)$ and $\sigma(\theta_\lambda) = \sigma(\delta^*(\theta_\lambda, s^0, P^0), \theta_\lambda, P^0)$. By Assumption 3 both ε^n and $\varepsilon^R(\theta_\lambda)$ are sums of i.i.d. mean zero random vectors with known covariance matrix.

By the definition of ε^n and $\varepsilon^R(\theta_\lambda)$ and from (20), we have

$$s^0 + \varepsilon^n - \varepsilon^R(\theta_\lambda) = \sigma(\delta^*(\theta_\lambda, s^n, P^R), \theta_\lambda, P^0)$$

and therefore we can expand the inverse map from (θ_λ, s^n, P) to $\delta^*(\theta_\lambda, s^n, P)$ around s^0. Assumption 4 ensures that for each J, almost every P, almost all δ^*, and every $\theta_\lambda \in \Theta_\lambda$, the function $\sigma(\delta^*, \theta_\lambda, P)$ is differentiable in δ^*, and its derivative has an inverse

$$H_\delta^{-1}(\delta^*, \theta_\lambda, P) = \left\{ \frac{\partial \sigma(\delta^*, \theta_\lambda, P)}{\partial \delta^{*\prime}} \right\}^{-1}.$$

To save notation define $\sigma(\theta_\lambda, s, P) = \sigma(\delta^*(\theta_\lambda, s, P), \theta_\lambda, P)$, $H_\delta(\theta_\lambda, s, P) = H_\delta(\delta^*(\theta_\lambda, s, P), \theta_\lambda, P)$, and $H_{\delta 0} = H_\delta(\theta_{\lambda 0}, s^0, P^0)$. Then applying Taylor expansions to the last two terms in (23) we can obtain

$$\delta^*(\theta_\lambda, s^n, P^R) \simeq \delta^*(\theta_\lambda, s^0, P^0) + H_{\delta 0}^{-1}\{\varepsilon^n - \varepsilon^R(\theta_{\lambda 0})\} \tag{24}$$

for θ_λ approaching to $\theta_{\lambda 0}$ and the last term enters the influence function for the sampling and simulation errors in the asymptotic expansion to obtain the asymptotic normality. The asymptotic variance in the absence of the sampling and simulation errors are obtained as the variance at $\delta^*(\theta_\lambda, s^0, P^0)$.

In deriving the asymptotic distribution for a specific estimator we focus on the sieve MD estimator in (19). We obtain the convergence rate and the asymptotic normality of the parameter estimates $(\hat{\theta}, \hat{\theta}_\lambda)$ building on Newey, Powell, and Vella (1999) and Chen (2007). But we have a few added complications to their problem. First we have additional nonparametric estimation in the middle step of estimation and second our estimator is a sieve MD estimator. Because we estimate $(\theta_0, \theta_{\lambda 0})$ and f_0 simultaneously at the main estimation and also because of the middle step estimation, we cannot directly apply Chen, Linton, and van Keilegom (2003) to our problem either. Finally we also need to account for the sampling and simulation error in the asymptotic distribution.

Our inference will focus on the finite dimensional parameter $(\theta_0, \theta_{\lambda 0})$ and we view f_0 as a nuisance parameter.

7.1 Asymptotic Normality that Does Not Account for the Sampling and Simulation Errors

We first derive the asymptotic normality of $(\hat{\theta}, \hat{\theta}_\lambda)$ and a consistent estimator for the variance term when the contribution of the sampling and the simulation errors to the variance is negligible and will add variances due to these errors later. The asymptotic variance that does not account for the sampling and the simulation errors is obtained by ignoring the term

$$\{\delta^*(\theta_\lambda, s^n, P^R) - \delta^*(\theta_\lambda, s^0, P^R)\} + \{\delta^*(\theta_\lambda, s^0, P^R) - \delta^*(\theta_\lambda, s^0, P^0)\}$$

in (23) or equivalently ignoring $H_{\delta 0}^{-1}\{\varepsilon^n - \varepsilon^R(\theta_{\lambda 0})\}$ in (24) in the stochastic expansion.

Our asymptotics builds on results from the asymptotic normality of series estimators with generated regressors and that of sieve estimators (see e.g., Newey, Powell, and Vella (1999) and Chen (2007)). Define

$$g(z_j, \mathbf{v}_j; \theta, f) = c + \beta' x_j - \alpha p_j + f(z_j, \mathbf{v}_j)(1 + \gamma' x_j + \gamma_p(\bar{y} - p_j)),\ g_{0j} = g(z_j, \mathbf{v}_j; \theta_0, f_0),$$

$$\Psi_\theta(z_j, \mathbf{v}_j) = \frac{\partial g(\cdot)}{\partial \theta} = (1, x_j', -p_j, x_j' f(z_j, \mathbf{v}_j), (\bar{y} - p_j) f(z_j, \mathbf{v}_j))',\ \Psi_{\theta_0, j} = \frac{\partial g(\cdot)}{\partial \theta}|_{f=f_0},$$

and let

$$\Delta_{\theta_\lambda, j}(s, P) = \frac{\partial \delta_j^*(\theta_\lambda, s, P)}{\partial \theta_\lambda},\ \Delta_{\theta_{\lambda 0}, j} = \left.\frac{\partial \delta_j^*(\theta_\lambda, s^0, P^0)}{\partial \theta_\lambda}\right|_{\theta_\lambda = \theta_{\lambda 0}}.$$

Below with possible abuse of notation $\frac{\partial g(z,\mathbf{v})}{\partial f} f$ will denote the pathwise (functional) derivative $\frac{dg(z,\mathbf{v})}{df}[f]$ as defined in Chen (2007). We use this notation because this derivative is well-defined as the usual derivative in our problem. We will use this notation to denote $\frac{\partial g(z,\mathbf{v})}{\partial f} = (1+\gamma' x + \gamma_p(\bar{y} - p))$ and similar notation is used for others.

The $\sqrt{J}$-consistency and the asymptotic normality in the form of

$$\sqrt{J}((\hat{\theta}_\lambda', \hat{\theta}')' - (\theta_{\lambda 0}', \theta_0')') \to_d N(0, \Omega)$$

depends on the existence of the representation such that for a functional $b(\theta_\lambda, \theta, f)$, we have

$$\begin{aligned} & \sqrt{J}(\theta_\lambda' - \theta_{\lambda 0}', \theta' - \theta_0')' = \sqrt{J} b(\theta_\lambda - \theta_{\lambda 0}, \theta - \theta_0, f) \\ \simeq\ & \sqrt{J} E[\omega^{*J}(Z_j, \mathbf{V}_j)\{E[\Delta'_{\theta_{\lambda 0}, j}|Z_j, \mathbf{V}_j](\theta_\lambda - \theta_{\lambda 0}) - \Psi_\theta(Z_j, \mathbf{V}_j)'(\theta - \theta_0) + C\frac{\partial g(Z_j, \mathbf{V}_j)}{\partial f} f(Z_j, \mathbf{V}_j)\}] \end{aligned}$$

for some constant C and the second moment of the Riesz representer like term $\omega^*(Z_j, \mathbf{V}_j)$ is bounded.

In this case $\sqrt{J}((\hat{\theta}_\lambda', \hat{\theta}')' - (\theta_{\lambda 0}', \theta_0')')$ is asymptotically normal and $\omega^{*J}(z_j, \mathbf{v}_j)$ has the form of

$$\omega^{*J}(z_j, \mathbf{v}_j) = \left(\sum_{j=1}^{J} E[r_0(Z_j, \mathbf{V}_j) r_0(Z_j, \mathbf{V}_j)']/J \right)^{-1} r_0(z_j, \mathbf{v}_j)$$

where $r_0(z_j, \mathbf{v}_j)$ is the mean-squared projection residual of $E[(\Psi'_{\theta_{\lambda 0}, j}, -\Psi'_{\theta_0, j})'|Z_j, \mathbf{V}_j]$ on the functions of the form $\frac{\partial g(z_j, \mathbf{v}_j)}{\partial f} f(z_j, \mathbf{v}_j)$ that satisfies $E[f(Z_j, \mathbf{V}_j)|Z_j] = 0$.[16]

Moreover the existence of the above representation implies that the asymptotic variance Ω has an explicit form. To obtain the explicit form of the asymptotic variance. Let

$$\Sigma_{0j}(z_j, \mathbf{v}_j) = \mathrm{Var}(\delta_j^*(\theta_{\lambda 0}, s^0, P^0) - g_{0j}|z_j, \mathbf{v}_j)$$

and let $\rho_v(z_j) = E[\omega^{*J}(Z_j, \mathbf{V}_j)\frac{\partial g_{0j}}{\partial f_0}(\frac{\partial f_0(Z_j, \mathbf{V}_j)}{\partial V_j} - E[\frac{\partial f_0(Z_j, \mathbf{V}_j)}{\partial V_j}|Z_j])|z_j]$ and

$$\rho_{\bar{\varphi}_l}(z_j) = E[a_l \omega^{*J}(Z_j, \mathbf{V}_j)\frac{\partial g_{0j}}{\partial f_0}|z_j].$$

Then the asymptotic variance of the estimator $(\hat{\theta}_\lambda, \hat{\theta})$ is given by

$$\Omega = \lim_{J \to \infty} \sum_{j=1}^{J} \Omega_j / J$$

where

$$\begin{aligned} \Omega_j \ =\ & E[\omega^{*J}(Z_j, \mathbf{V}_j)\Sigma_{0j}(Z_j, \mathbf{V}_j)\omega^{*J}(Z_j, \mathbf{V}_j)'] + E[\rho_v(Z_j)\mathrm{var}(p_j|Z_j)\rho_v(Z_j)'] \\ & + \sum_l E[\rho_{\bar{\varphi}_l}(Z_j)\mathrm{var}(\varphi_l(Z_j, \mathbf{V}_j)|Z)\rho_{\bar{\varphi}_l}(Z_j)']. \end{aligned} \quad (25)$$

The first term in the variance accounts for the main estimation, the second term accounts for the estimation of the control (V), and the last term accounts for the middle step estimation.

Next we focus on obtaining correct standard errors for $(\hat{\theta}_\lambda, \hat{\theta})$ and providing a consistent estimator for the standard errors. To derive a consistent estimator of Ω we introduce additional notation. From here we let x_j be one dimensional for notational simplicity but without loss of gener-

[16] Let $(\Delta'_{\theta_{\lambda 0}, j}, -\Psi'_{\theta_0, j})'_l$ denote the l-th element of $(\Delta'_{\theta_{\lambda 0}, j}, -\Psi'_{\theta_0, j})'$ and define

$$f_l{}^* = \mathrm{argmin}_{f_l \in \mathcal{F}} \frac{1}{J}\sum_{j=1}^{J} E\left[\left\{E[(\Delta'_{\theta_{\lambda 0}, j}, -\Psi'_{\theta_0, j})'_l|Z_j, \mathbf{V}_j] - \frac{\partial g_0(Z_j, \mathbf{V}_j)}{\partial f} f_l\right\}^2\right].$$

Then we obtain $r_0(z_j, \mathbf{v}_j) = E[(\Delta'_{\theta_{\lambda 0}, j}, -\Psi'_{\theta_0, j})'|z_j, \mathbf{v}_j] - \frac{\partial g_0(z_j, \mathbf{v}_j)}{\partial f}(f_1^*, \cdots, f^*_{\dim(\theta_\lambda) + \dim(\theta)})'$.

ality. Define $\Psi^L_{\theta_0,j} = (1, x_j, -p_j, x_j f_0(z_j,\mathbf{v}_j), (\bar{y}-p_j)f_0(z_j,\mathbf{v}_j), \frac{\partial g_{0j}}{\partial f_0}\tilde{\varphi}^L(z_j,\mathbf{v}_j)')'$ where $\tilde{\varphi}^L(z_j,\mathbf{v}_j) = (\tilde{\varphi}_1(z_j,\mathbf{v}_j),\ldots,\tilde{\varphi}_L(z_j,\mathbf{v}_j))'$ and $\frac{\partial g_{0j}}{\partial f_0} \equiv (1+\gamma_0 x_j + \gamma_{p0}(\bar{y}-p_j))$ and let

$$A = (\sum_{j=1}^{J} E[r_{0j}r_{0j}']/J)^{-1}\sum_{j=1}^{J} E[r_{0j}E[(\Delta'_{\theta_{\lambda 0},j}, \Psi^{L\prime}_{\theta_0,j})|Z_j,\mathbf{V}_j]/J]$$

where we abbreviate $r_{0j} \equiv r_0(z_j,\mathbf{v}_j)$.

Then note that we have

$$(\theta'_{\lambda 0},\theta'_0)' = A\vartheta_0,\ (\hat{\theta}'_\lambda,\hat{\theta}')' = A\hat{\vartheta}$$

where $\vartheta = (\theta'_\lambda, c, \beta, \alpha, \gamma, \gamma_p, a'_L)'$ and we let $a_L = (a_1,\ldots,a_L)'$ with abuse of notation. Moreover observe that $A = (I_{\dim(\theta_\lambda,\theta)}, \mathbf{0}_{\dim(\theta_\lambda,\theta)\times L})$ where $I_{\dim(\theta_\lambda,\theta)}$ is the $\dim(\theta_\lambda,\theta)\times\dim(\theta_\lambda,\theta)$ identity matrix and $\mathbf{0}_{\dim(\theta_\lambda,\theta)\times L}$ is the $\dim(\theta_\lambda,\theta)\times L$ zero matrix because the projection of the variable being projected on the mean-squared projection residual is the residual itself and the projection of the projection variable on the mean-squared projection residual is equal to zero. We can practically view A as a selection matrix that selects the parameter of interest $(\theta'_{\lambda 0},\theta'_0)$ from the whole parameter vector ϑ. Other linear combinations of θ_0 or $\theta_{\lambda 0}$ is also easily obtained by choosing different A's (see Newey, Powell, and Vella (1999) and Chen (2007)).

To obtain a consistent variance estimator, further let $\hat{g}(z_j,\hat{\mathbf{v}}_j) = \hat{c}+\hat{\beta}x_j-\hat{\alpha}p_j+\hat{f}(z_j,\hat{\mathbf{v}}_j)(1+\hat{\gamma}x_j+\hat{\gamma}_p(\bar{y}-p_j))$ and $\hat{\hat{g}}_j = \hat{g}(z_j,\hat{\mathbf{v}}_j)$. Let $\hat{\Psi}_{\hat{\theta}_\lambda,j} = \hat{E}[\frac{\partial\delta^*_j(\theta_\lambda,s^n,P^R)}{\partial\theta_\lambda}|z_j,\hat{\mathbf{v}}_j]|_{\theta_\lambda=\hat{\theta}_\lambda}$ and $\hat{\Psi}^L_{\theta,j} = (1, x_j, -p_j, x_j\hat{f}(z_j,\hat{\mathbf{v}}_j), (\bar{y}-p_j)\hat{f}(z_j,\hat{\mathbf{v}}_j), \frac{\partial\hat{\hat{g}}_j}{\partial\hat{f}}\hat{\tilde{\varphi}}^L(z_j,\hat{\mathbf{v}}_j)')'$ where $\hat{\tilde{\varphi}}^L(z_j,\mathbf{v}_j) = (\hat{\tilde{\varphi}}_1(z_j,\mathbf{v}_j),\ldots,\hat{\tilde{\varphi}}_L(z_j,\mathbf{v}_j))'$ and let $\hat{\Psi}^L_j = (\hat{\Psi}'_{\hat{\theta}_\lambda,j}, -\hat{\Psi}^{L\prime}_{\theta,j})'$. Then define the followings:

$$\begin{aligned}
\hat{\mathcal{T}} &= \sum\nolimits_{j=1}^{J}\hat{\Psi}^L_j\hat{\Psi}^{L\prime}_j/J,\ \hat{\Sigma} = \sum\nolimits_{j=1}^{J}(\delta^*_j(\hat{\theta}_\lambda,s^n,P^R)-\hat{g}(z_j,\hat{\mathbf{v}}_j))^2\hat{\Psi}^L_j\hat{\Psi}^{L\prime}_j/J \qquad (26)\\
\hat{\mathcal{T}}_1 &= \mathbf{P}'\mathbf{P}/J,\ \hat{\Sigma}_1 = \sum_{j=1}^{J}\hat{v}_j^2\varphi^k(z_j)\varphi^k(z_j)'/J,\ \hat{\Sigma}_{2,l} = \sum_{j=1}^{J}(\varphi_l(z_j,\hat{\mathbf{v}}_j)-\hat{\tilde{\varphi}}_l(z_j))^2\varphi^k(z_j)\varphi^k(z_j)'/J\\
\hat{H}_{11} &= \sum_{j=1}^{J}\frac{\partial\hat{\hat{g}}_j}{\partial\hat{f}}\sum_{l=1}^{L}\hat{a}_l\frac{\partial\varphi_l(z_j,\hat{\mathbf{v}}_j)}{\partial v_j}\hat{\Psi}^L_j\varphi^k(z_j)'/J,\\
\hat{H}_{12} &= \sum_{j=1}^{J}\frac{\partial\hat{\hat{g}}_j}{\partial\hat{f}}\varphi^k(z_j)'\left((\mathbf{P}'\mathbf{P})^-\sum_{j'=1}^{J}\varphi^k(z_{j'})\frac{\partial\sum_{l=1}^{L}\hat{a}_l\varphi_l(z_{j'},\hat{\mathbf{v}}_{j'})}{\partial v_{j'}}\right)\hat{\Psi}^L_j\varphi^k(z_j)'/J,\\
\hat{H}_{2,l} &= \sum_{j=1}^{J}\hat{a}_l\frac{\partial\hat{\hat{g}}_j}{\partial\hat{f}}\hat{\Psi}^L_j\varphi^k(z_j)'/J,\ \hat{H}_1 = \hat{H}_{11}-\hat{H}_{12}.
\end{aligned}$$

Then, we can estimate Ω consistently with

$$\hat{\Omega} = A\hat{\mathcal{T}}^{-1}\left[\hat{\Sigma}+\hat{H}_1\hat{\mathcal{T}}_1^{-1}\hat{\Sigma}_1\hat{\mathcal{T}}_1^{-1}\hat{H}_1' + \sum\nolimits_{l=1}^{L}\hat{H}_{2,l}\hat{\mathcal{T}}_1^{-1}\hat{\Sigma}_{2,l}\hat{\mathcal{T}}_1^{-1}\hat{H}'_{2,l}\right]\hat{\mathcal{T}}^{-1}A'. \qquad (27)$$

This is the heteroskedasticity robust variance estimator that accounts for the first and the middle step estimations. The first variance term $A\hat{\mathcal{T}}^{-1}\hat{\Sigma}\hat{\mathcal{T}}^{-1}A'$ corresponds to the variance estimator without pre-step estimations. The second variance term (that accounts for the estimation of V)

corresponds to the second term in (25) and the third variance term (that accounts for the estimation of $\bar{\varphi}_l(\cdot)$'s) corresponds to the third term in (25). If we view our model as a parametric one with fixed $k(J)$ and $L(J)$, the same variance estimator $\hat{\Omega}$ can be used as the estimator of the variance for the parametric model (e.g, Newey (1984), Murphy and Topel (1985)).

To present the theorem, we need additional notation and assumptions. For any differentiable function $c(w)$, let $|\mu| = \sum_{l=1}^{\dim(w)} \mu_l$ and define $\partial^\mu c(w) = \partial^{|\mu|} c(w)/\partial w_1 \cdots \partial w_{\dim(w)}$. Also define $|c(w)|_\iota = \max_{|\mu|\leq\iota} \sup_{w\in\mathcal{W}} ||\partial^\mu c(w)||$ and others are defined similarly.

Assumption 16 (C1)**.** *(i) $\{\delta_j^*(\theta_\lambda, s, P), p_j, z_j : j \leq J, J \geq 1\}$ is a triangular array of random variables on a probability space for all (θ_λ, s, P) in a small neighborhood of $(\theta_{\lambda 0}, s^0, P^0)$; var$(p_j|z_j)$ and var$(\delta_j^*(\theta_\lambda, s, P)|z_j, \boldsymbol{v}_j)$ (for all (θ_λ, s, P) in a small neighborhood of $(\theta_{\lambda 0}, P^0, s^0)$) are bounded for all j, and var$(\varphi_l(Z_j, \mathbf{V}_j)|z_j)$ are bounded for all l and all j; (ii) (p_j, z_j) are continuously distributed with densities that are bounded away from zero on their supports, respectively and their supports are compact; (iii) $\Pi_0(z)$ is continuously differentiable of order s_Π and all the derivatives of order s_Π are bounded on the support of Z; (iv) $\bar{\varphi}_{0l}(Z)$ is continuously differentiable of order s_φ and all the derivatives of order s_φ are bounded for all l on the support of Z; (v) $f_0(z_j, \mathbf{v}_j)$ is Lipschitz in $\mathbf{v}_j$ and is continuously differentiable of order s_f and all the derivatives of order s_f are bounded on the support of $(z_j, \mathbf{v}_j)$; (vi) $\varphi_l(z_j, \mathbf{v}_j)$ is Lipschitz and is twice continuously differentiable in $\mathbf{v}_j$ and its first and second derivatives are bounded for all l; (vii) $\frac{\partial\delta_j^*(\theta_\lambda,s,P)}{\partial\theta_\lambda}$ is continuous at $(\theta_{\lambda 0}, s^0, P^0)$ and $||\frac{\partial\delta_j^*(\theta_\lambda,s,P)}{\partial\theta_\lambda}|| < C$ for some $C < \infty$ for all j in a small neighborhood of $(\theta_{\lambda 0}, s^0, P^0)$; (viii) $E[\frac{\partial\delta_j^*(\theta_\lambda,s,P)}{\partial\theta_\lambda}|z_j, \mathbf{v}_j]$ is Lipschitz in $\mathbf{v}_j$ and is continuously differentiable of order s_δ and all the derivatives of order s_δ are bounded on the support of $(z_j, \mathbf{v}_j)$ in the neighborhood of $(\theta_{\lambda 0}, s^0, P^0)$; (ix) Let a metric be $\bar{\rho}_\delta(\delta^*, \tilde{\delta}^*) = \max\{J^{-1}\sum_{j=1}^J(\delta_j^* - \tilde{\delta}_j^*)^2, J^{-1}\sum_{j=1}^J(\frac{\partial\delta_j^*}{\partial\theta_\lambda} - \frac{\partial\tilde{\delta}_j^*}{\partial\theta_\lambda})^2\}$ and let $\bar{\mathcal{N}}_{\delta^{*0}}(\theta_\lambda, \epsilon) = \{\delta^* : \bar{\rho}_\delta(\delta^*, \delta^*(\theta_\lambda, s^0, P^0)) < \epsilon\}$. Then For all ϵ, there exists $C(\epsilon) > 0$ such that*

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

where $\Theta_{\lambda 0}$ denotes a neighborhood of $\theta_{\lambda 0}$.

Assumption C1 (i) is about nature of the data and other conditions in Assumption C1 are standards in sieve estimations. Assumptions C1 (iii), (iv), and (v) let the unknown functions $\Pi_0(z)$, $\bar{\varphi}_{0l}(z)$, and $f_0(z, \mathbf{v})$ belong to a Hölder class of functions, respectively and they can be approximated up to the orders of $O(k(J)^{-s_\Pi/\dim(z)})$, $O(k(J)^{-s_\varphi/\dim(z)})$, and $O(L(J)^{-s_f/\dim(z,\mathbf{v})})$, respectively when we approximate them using polynomials or splines (see Timan (1963), Schumaker (1981), Newey (1997), and Chen (2007)) where $\dim(z)$ and $\dim(z, \mathbf{v})$ denote the dimension of Z and $(Z, \mathbf{V})$, respectively. Assumption C1 (viii) implies the conditional expectation $E[\frac{\partial\delta_j^*(\theta_\lambda,s,P)}{\partial\theta_\lambda}|z_j, \mathbf{v}_j]$ is well approximated up to the orders of $O(L(J)^{-s_\delta/\dim(z,\mathbf{v})})$ as well. We focus on polynomials (i.e., power series) and spline approximations in this paper. Assumption C1 (vi) is satisfied for the approximating polynomials and splines with appropriate orders. The assumption that Z is continuous is not essential when a subset of Z is discrete, we can condition on those discrete variables and the model becomes parametric in regard to those variables. Assumption C1 (vii) enables us to apply the mean value expansion of $\delta_j^*(\theta_\lambda, \cdot)$ w.r.t. θ_λ in a small neighborhood of $(\theta_{\lambda 0}, s^0, P^0)$. Assumption C1 (ix) strengthens Assumption 14 but only in a neighborhood of $\theta_{\lambda 0}$.

This condition ensures that at least asymptotically we can distinguish the δ^* as a function of θ_λ that sets the models predictions for shares equal to the actual shares from other function $\delta \neq \delta^*$ as a function of θ_λ, at least up to its first derivative. This condition ensures that $\frac{\partial \delta^*(\theta_\lambda, s^0, P^0)}{\partial \theta_\lambda}$ is also well approximated by $\frac{\partial \delta^*(\theta_\lambda, s^n, P^R)}{\partial \theta_\lambda}$ as well as $\delta^*(\theta_\lambda, s^n, P^R)$ approximates $\delta^*(\theta_\lambda, s^0, P^0)$.

Assumption 17 (N1)**.** *(i) (a) $r_0(z_j, \mathbf{v}_j)$ is continuously differentiable with order s_r and $E[||r_0(Z_j, \mathbf{V}_j)||^4]$ is bounded; (b) $\sum_{j=1}^J E[r_0(Z_j, \mathbf{V}_j) r_0(Z_j, \mathbf{V}_j)']/J$ has smallest eigenvalues that are bounded away from zero for all J large enough; (ii) there exist ι, ϱ, and a_L such that $|f_0(z, \mathbf{v}) - a_L' \tilde{\varphi}^L(z, \mathbf{v})|_\iota \leq CL^{-\varrho}$; (iii) $\Sigma_{0j}(z_j, \mathbf{v}_j)$ is bounded away from zero, $E[(\delta_j^*(\theta_{\lambda 0}) - g_{0j})^4 | z_j, \mathbf{v}_j]$ and $E[V_j^4 | z_j]$ are bounded for all j and $E[\tilde{\varphi}_l(Z_j, \mathbf{V}_j)^4 | z_j]$ is bounded for all l and j.*

Next we impose the rate conditions that restrict the growth of $k(J)$ and $L(J)$ as J tends to infinity.

Assumption 18 (N2)**.** *Let $\triangle_{J,1} = k(J)^{1/2}/\sqrt{J} + k(J)^{-s_\Pi/\dim(z)}$, $\triangle_{J,2} = k(J)^{1/2}/\sqrt{J} + k(J)^{-s_\varphi/\dim(z)}$, and $\Delta_J = \max\{\Delta_{J,1}, \Delta_{J,2}\} \to 0$ and $\triangle_\delta = L(J)^{1/2}/\sqrt{J} + L(J)^{-s_\delta/\dim(z,\mathbf{v})} \to 0$.
Let $\sqrt{J} k(J)^{-s_\Pi/\dim(z)}$, $\sqrt{J} k(J)^{-s_\varrho/\dim(z)}$, $\sqrt{J} k(J)^{1/2} L(J)^{-s_f/\dim(z,\mathbf{v})} \to 0$ and they are sufficiently small. For the polynomial approximations $\frac{L(J)^5 k(J)^{1/2} + L(J)^{9/2} k(J)^{3/2}}{\sqrt{J}} \to 0$ and for the spline approximations $\frac{L(J)^{7/2} k(J)^{1/2} + L(J)^3 k(J) + L(J)^2 k(J)^{3/2}}{\sqrt{J}} \to 0$.*

Theorem 5 (AN1)**.** *Suppose Assumptions 3-6, 9-10, and 13-15 hold. Suppose Condition S, Assumption C1, N1-N2 are satisfied. Then*

$$\sqrt{J}((\hat{\theta}_\lambda', \hat{\theta}')' - (\theta_{\lambda 0}', \theta_0')') \to_d N(0, \Omega) \text{ and } \hat{\Omega} \to_p \Omega.$$

7.2 Accounting for the Sampling and the Simulation Errors

We derive variance terms due to the sampling and the simulation errors. As in Berry, Linton, and Pakes (2004) the challenge here is to control the behavior of $J \times J$ matrix $H_\delta^{-1}(\delta^*(\theta_\lambda, s, P), \theta_\lambda, P)$ when the number of products J grows. $H_\delta^{-1}(\cdot)$ is the inverse of $\partial\sigma(\cdot)/\partial\delta$, so when the model implies diffuse substitution patterns such as the random coefficient logit models, the partial $\partial\sigma(\cdot)/\partial\delta$ tends to zero as J grows and it makes the inverse $H_\delta^{-1}(\cdot)$ grow large. This means when J is large the inverted δ^* (so ξ) becomes very sensitive to even small sampling or simulation error.

In this section following Berry, Linton, and Pakes (2004) we obtain relevant variance terms for our estimation problem. Let $\mathbf{r}_0(\mathbf{z}, \mathbf{v})' = (r_0(z_1, \mathbf{v}_1), \ldots, r_0(z_J, \mathbf{v}_J))$ and define the stochastic process in $(\delta^*, \theta_\lambda, P)$

$$\upsilon_J(\delta^*, \theta_\lambda, P) = \frac{1}{\sqrt{J}} \mathbf{r}_0(\mathbf{z}, \mathbf{v})' H_\delta^{-1}(\delta^*, \theta_\lambda, P)(\varepsilon^n - \varepsilon^R(\theta_\lambda)). \tag{28}$$

We obtain the influence functions due to the sampling and the simulations errors (see Appendix D) as

$$\frac{1}{\sqrt{J}} \omega^{*JJ\prime} H_{\delta 0}{}^{-1}(\varepsilon^n - \varepsilon^R(\theta_{\lambda 0})) = (\Xi^J)^{-1} \upsilon_J(\delta^{*0}, \theta_{\lambda 0}, P^0)$$

where $\omega^{*JJ\prime} = (\omega_J^*(z_1, \mathbf{v}_1), \ldots, \omega_J^*(z_J, \mathbf{v}_J))$ and $\Xi^J \equiv \sum_{j=1}^J E[r_0(Z_j, \mathbf{V}_j) r_0(Z_j, \mathbf{V}_j)']/J$. Therefore analyzing the stochastic process $\upsilon_J(\delta^*, \theta_\lambda, P)$ is necessary to derive variance terms. Write $\mathbf{r}_0(\mathbf{z}, \mathbf{v})' H_\delta^{-1}(\delta^*, \theta_\lambda, P) \equiv (c_1(\delta^*, \theta_\lambda, P), \ldots, c_J(\delta^*, \theta_\lambda, P))$.

Then we can rewrite $\upsilon_J(\delta^*, \theta_\lambda, P)$ as two sums of independent random variables from a triangular array

$$\upsilon_J(\delta^*, \theta_\lambda, P) = \sum_{i=1}^{n} Y_{Ji}(\delta^*, \theta_\lambda, P) - \sum_{r=1}^{R} Y^*_{Jr}(\delta^*, \theta_\lambda, P)$$

where

$$\begin{aligned} Y_{Ji}(\delta^*, \theta_\lambda, P) &= \frac{1}{n\sqrt{J}} \sum_{j=1}^{J} c_j(\delta^*, \theta_\lambda, P)\varepsilon_{ji} \\ Y^*_{Jr}(\delta^*, \theta_\lambda, P) &= \frac{1}{R\sqrt{J}} \sum_{j=1}^{J} c_j(\delta^*, \theta_\lambda, P)\varepsilon_{jr}(\theta_\lambda). \end{aligned}$$

Note that Y_{Ji} and Y^*_{Jr} are i.i.d across i and r respectively and their distributions depend on J. We then provide conditions that the process $\upsilon_J(\delta^*, \theta_\lambda, P)$ has limit distribution at $(\theta_{\lambda 0}, s^0, P^0)$ and we add the resulting asymptotic variance terms due to the sampling and the simulation errors to the asymptotic variance Ω to obtain the full asymptotic variance of $\sqrt{J}((\hat{\theta}'_\lambda, \hat{\theta}') - (\theta'_{\lambda 0}, \theta'_0))$.

Assumption N3 below replace Assumptions B4 in Berry, Linton, and Pakes (2004) and we argue below our model specifications with and without random coefficients satisfy this Assumption N3.

Assumption 19 (N3)**.** *Let $Y_{Ji} = Y_{Ji}(\delta^*(\theta^0_\lambda, s^0, P^0), \theta_{\lambda 0}, P^0)$ and $Y^*_{Jr} = Y^*_{Jr}(\delta^*(\theta^0_\lambda, s^0, P^0), \theta_{\lambda 0}, P^0)$. With probability one (i) $\lim_{J\to\infty} nE_*[Y_{Ji}Y'_{Ji}] = \Phi_2$ and (ii) $\lim_{J\to\infty} RE_*[Y^*_{Jr}Y^{*\prime}_{Jr}] = \Phi_3$ for finite positive definite non-random matrices Φ_2 and Φ_3. Also for some $\tau > 0$ with probability one (iii) $nE_*[||Y_{Ji}||^{2+\tau}] = o(1)$ and (iv) $RE_*[||Y^*_{Rr}||^{2+\tau}] = o(1)$.*

In the original BLP specification ξ is additive in δ. Therefore our $H_\delta(\cdot)$ is equivalent to their $H(\cdot)$, derivative of $\sigma(\cdot)$ with respect to ξ. In the logit case without random coefficients we have

$$H_\delta(\cdot, s, \cdot) = S - ss' \text{ and } H_\delta^{-1}(\cdot, s, \cdot) = S^{-1} + \mathbf{i}\mathbf{i}'/s_0,$$

where $S = \text{diag}[s]$ and $\mathbf{i} = (1, \ldots, 1)'$. Then by the essentially same argument in Berry, Linton, and Pakes (2004) (page 636-637) when the model is logit without random coefficients we obtain

$$\begin{aligned} \Phi_2(J) &= \frac{1}{nJ}\mathbf{r}_0(\mathbf{z}, \mathbf{v})' H^{-1}_{\delta 0}\mathbf{r}_0(\mathbf{z}, \mathbf{v}) = \frac{J}{n} \times \left[\frac{1}{J}\sum_{j=1}^{J} E[r_0(z_j, \mathbf{v}_j) r_0(z_j, \mathbf{v}_j)'](Js_j)^{-1}\right] \\ &+ \frac{J^2}{n} \times \left[\frac{\frac{1}{J}\sum_{j=1}^{J} E[r_0(z_j, \mathbf{v}_j)]\frac{1}{J}\sum_{j=1}^{J} E[r_0(z_j, \mathbf{v}_j)']}{(Js_0)}\right] \\ &= O_p(J/n) + O_p(J^2/n) = O_p(J^2/n) \end{aligned}$$

by Condition S and Assumption N1 (i) and then we have

$$\Phi_2 = \lim_{J\to\infty} \frac{J^2}{n} \times \left[\frac{\lim_{J\to\infty}\frac{1}{J}\sum_{j=1}^{J} E[r_0(z_j, \mathbf{v}_j)] \lim_{J\to\infty}\frac{1}{J}\sum_{j=1}^{J} E[r_0(z_j, \mathbf{v}_j)']}{\lim_{J\to\infty}(Js_0)}\right].$$

Therefore the logit model with our mean utility specification allowing for interactions satisfies Assumption N3 (i).

Also in the case of the random coefficient model by the essentially same argument in Berry, Linton, and Pakes (2004) (page 637-638), we have

$$H_{\delta}^{-1} = [E[\mathcal{H}_{\delta}(\lambda)]]^{-1} \leq E[\mathcal{H}_{\delta}(\lambda)^{-1}]$$

where $\mathcal{H}_{\delta}(\lambda) = S(\lambda) - s(\lambda)s(\lambda)'$, $s(\lambda) = (s_1(\lambda), \ldots, s_J(\lambda))'$, and $S(\lambda) = \text{diag}(s(\lambda))$. If we assume $s_j(\lambda) \geq \underline{s}_j$ for all $\theta_\lambda \in \Theta_\lambda$ and $j = 0, 1, \ldots, J$ for some non-random sequence of constants $\underline{s}_j$ that satisfy condition S we obtain

$$H_{\delta}^{-1} \leq \underline{S}^{-1} + \frac{\mathbf{ii}'}{\underline{s}_0} \equiv \overline{H_{\delta}^{-1}} \text{ and } H_{\delta}^{-1}V_2H_{\delta}^{-1\prime} \leq \overline{H_{\delta}^{-1}}V_2\overline{H_{\delta}^{-1}}' \text{ and } H_{\delta}^{-1}V_3H_{\delta}^{-1\prime} \leq \overline{H_{\delta}^{-1}}V_3\overline{H_{\delta}^{-1}}' \tag{29}$$

where $\underline{S} = \text{diag}(\underline{s}_1, \ldots, \underline{s}_J)$. We then obtain under Condition S and Assumption N1 (i),

$$\begin{aligned} \Phi_2 &= \lim_{J\to\infty} \frac{1}{nJ} \mathbf{r}_0(\mathbf{z}, \mathbf{v})' H_{\delta 0}^{-1} V_2 H_{\delta 0}^{-1\prime} \mathbf{r}_0(\mathbf{z}, \mathbf{v}) && (30) \\ &= \lim_{J\to\infty} \frac{J^2}{n} \times \frac{\lim_{J\to\infty} \frac{1}{J}\sum_{j=1}^{J} E[r_0(z_j, \mathbf{v}_j)] \frac{1}{J}\sum_{j=1}^{J} E[r_0(z_j, \mathbf{v}_j)']}{\lim_{J\to\infty}(J\int s_0(\lambda)dP(\lambda))} \\ \Phi_3 &= \lim_{J\to\infty} \frac{1}{RJ} \mathbf{r}_0(\mathbf{z}, \mathbf{v})' H_{\delta 0}^{-1} V_3 H_{\delta 0}^{-1\prime} \mathbf{r}_0(\mathbf{z}, \mathbf{v}) && (31) \\ &\leq \lim_{J\to\infty} \frac{J^2}{R} \times \frac{\lim_{J\to\infty} \frac{1}{J}\sum_{j=1}^{J} E[r_0(z_j, \mathbf{v}_j)] \frac{1}{J}\sum_{j=1}^{J} E[r_0(z_j, \mathbf{v}_j)']}{\lim_{J\to\infty}(J\int s_0(\lambda)dP(\lambda))} \end{aligned}$$

and therefore Assumption N3 is also satisfied in this case too. Note that (30) and (31) respectively correspond to (38) and (39) in Berry, Linton, and Pakes (2004) (page 638). The proof only requires to replace their $H(\cdot)$ with $H_\delta(\cdot)$ and also their z with $\mathbf{r}_0(\mathbf{z}, \mathbf{v})$, so is essentially identical. By the same token Assumption N3 (iii) and (iv) are also satisfied in our model too.

We then assume the stochastic process in (28) is stochastic equicontinuous in a small neighborhood of $(\delta^{*0}, \theta_{\lambda 0}, P^0)$ such that the process $\upsilon_J(\delta^*, \theta_\lambda, P)$ becomes arbitrarily close to $\upsilon_J(\delta^{*0}, \theta_{\lambda 0}, P^0)$ as $(\delta^*, \theta_\lambda, P) \to (\delta^{*0}, \theta_{\lambda 0}, P^0)$. This ensures the remainder terms do not affect the asymptotic distribution when we replace $\upsilon_J(\delta^*(\hat{\theta}_\lambda, s^n, P^R), \hat{\theta}_\lambda, P^R)$ with $\upsilon_J(\delta^{*0}, \theta_{\lambda 0}, P^0)$.

Assumption 20 (N4)**.** *The process $\upsilon_J(\delta^*, \theta_\lambda, P)$ is stochastically equicontinuous in $(\delta^*, \theta_\lambda, P)$ at $(\delta^*(\theta_{\lambda 0}, s^0, P^0), \theta_{\lambda 0}, P^0)$ such that for all sequences of positive numbers $\epsilon_J \to 0$,*

$$\lim_{J\to\infty} \Pr\{ \sup_{\theta_\lambda \in \mathcal{N}_{\theta_{\lambda 0}}(\epsilon_J)} \sup_{(\delta^*, P) \in \mathcal{N}_{\delta^{*0}}(\theta_\lambda, \epsilon_J) \times \mathcal{N}_{P^0}(\epsilon_J)} ||\upsilon_J(\delta^*, \theta_\lambda, P) - \upsilon_J(\delta^*(\theta_{\lambda 0}, s^0, P^0), \theta_{\lambda 0}, P^0)||\} = o_p(1).$$

This stochastic equicontinuity holds for the logit model and the random coefficient logit model as shown in Berry, Linton, and Pakes (2004). Again we only replace their $H(\cdot)$ with our $H_\delta(\cdot)$ and replace their z with $\mathbf{r}_0(\mathbf{z}, \mathbf{v})$ and the same arguments hold.

We then obtain the variance contribution due to the sampling and simulations errors as

$$(\Xi^J)^{-1} \upsilon_J(\delta^{*0}, \theta_{\lambda 0}, P^0) \to_d N(0, \Omega_2 + \Omega_3) \text{ and } \Omega_2 + \Omega_3 \equiv \Xi^{-1}(\Phi_2 + \Phi_3)(\Xi^{-1})'$$

where $\lim_{J\to\infty} \Xi^J = \Xi$.

We then can consistently estimate Ω_2 and Ω_3 respectively with

$$\begin{aligned}\hat{\Omega}_2 &= \frac{1}{nJ}A\hat{\mathcal{T}}^{-1}((\hat{\Psi}^{L,J})'\hat{H}_\delta^{-1}\hat{V}_2\hat{H}_\delta^{-1\prime}\hat{\Psi}^{L,J})\hat{\mathcal{T}}^{-1}A' \\ \hat{\Omega}_3 &= \frac{1}{RJ}A\hat{\mathcal{T}}^{-1}((\hat{\Psi}^{L,J})'\hat{H}_\delta^{-1}\hat{V}_3\hat{H}_\delta^{-1\prime}\hat{\Psi}^{L,J})\hat{\mathcal{T}}^{-1}A'\end{aligned} \tag{32}$$

where $\hat{H}_\delta = H_\delta(\hat{\theta}_\lambda, s^n, P^R)$, $\hat{V}_2 = S^n - s^n s^{n\prime}$, and $\hat{V}_3 = \frac{1}{R}\sum_{r=1}^R \varepsilon_r(\hat{\theta}_\lambda)\varepsilon_r(\hat{\theta}_\lambda)'$.

Theorem 6 (AN2)**.** *Suppose Assumptions 3-6, 9-10, and 13-15 hold. Suppose Condition S, Assumption C1, N1-N4 are satisfied. Then*

$$\sqrt{J}((\hat{\theta}_\lambda', \hat{\theta}')' - (\theta_{\lambda 0}', \theta_0')') \to_d N(0, \Omega + \Omega_2 + \Omega_3) \text{ and } (\hat{\Omega}, \hat{\Omega}_2, \hat{\Omega}_3) \to_p (\Omega, \Omega_2, \Omega_3).$$

Based on this asymptotic distribution, one can construct the confidence intervals of individual parameters and calculate standard errors straightforwardly using (26), (27), and (32).

8 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

$$\begin{aligned}[1]\ p &= 2 + Z + (5 + Z^2 + 5Z)\xi + \varsigma \\ [2]\ p &= Z + (5 + 5Z + \varsigma)\xi.\end{aligned}$$

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.

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:

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

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 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.

[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.

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

9 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.

9.1 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

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

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:

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

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.

9.2 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 sieve MD objective function given as

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

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 $\hat{\sigma} = \text{argmin}_\sigma \, Q_J(\theta(\sigma), \pi(\sigma); \sigma)$.

9.3 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 1% as the p-value for H_0 :$(\gamma_0, \gamma_{p0}) = 0$ is 0.0001, 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 intercept term have significant $\sigma_k' s$, but with the exception of the point estimate for β_k on

[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	-0.234
		(0.004)	(0.011)	(0.018)	(0.019)
Mean	Constant	-10.071	-9.915	-9.668	-10.435
Parameters		(0.252)	(0.263)	(0.290)	(0.680)
	HP/Weight	-0.122	1.226	2.815	1.079
		(0.277)	(0.404)	(0.527)	(1.189)
	Air	-0.034	0.486	1.379	1.383
		(0.072)	(0.133)	(0.179)	(0.189)
	MP$	0.265	0.172	0.103	0.146
		(0.043)	(0.049)	(0.054)	(0.068)
	Size	2.342	2.292	2.361	2.486
		(0.125)	(0.129)	(0.140)	(0.175)
Interaction	($\bar{y}$-price)$\cdot\xi$			-0.028	-0.045
Parameters				(0.019)	(0.060)
	HP/Weight$\cdot\xi$			0.975	1.212
				(1.238)	(2.443)
	Air$\cdot\xi$			0.414	0.637
				(0.429	(0.982)
	MP\$$\cdot\xi$			-0.045	-0.160
				(0.093)	(0.192)
	Size$\cdot\xi$			0.224	0.395
				(0.502)	(1.229)
Std. Deviations	Constant				1.783 (0.713)
	HP/Weight				2.454 (0.718)
	Air				0.249 (0.309)
	MP$				0.002 (0.008)
	Size				0.108 (0.089)
Control Ftns	V_1			3.597	2.395
				(2.988)	(3.382)
	V_2			-1.00	-0.835
				(1.949)	(1.754)
	V_3			0.041	0.158
				(0.998)	(0.790)
	V_4			-0.736	-0.429
				(0.589)	(0.572)
	V_5			0.100	0.077
				(0.183)	(0.166)
	V_6			1.120	0.666
				(0.898)	(0.928)
	V_7			-0.081	-0.073
				(0.126)	(0.123)
	V_8			0.302	0.208
				(0.272)	(0.283)
	V_9			-0.120	-0.035
				(0.263)	(0.175)

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.

Horsepower/weight, all of the other point estimates from Column 3 are largely the same. The 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 0.028, and the coefficient on the price coefficient changes from 0.232 to 0.234 and the price interaction term from -0.028 to -0.045.

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 -2.02, and the mean elasticity increases from -1.60 to -2.63. 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	-2.02	-2.08
Mean	-1.04	-1.60	-2.63	-2.68
Standard Deviation	0.77	1.17	1.69	1.71
No. of Inelastic Demands	68%	21%	4%	5%
Elasticities from 1990				
Median	-0.94	-1.43	-2.76	-2.84
Mean	-1.24	-1.90	-3.21	-3.31
Standard Deviation	0.84	1.28	1.86	1.87
No. of Inelastic Demands	53%	12%	2%	2%
1990 Models (from BLP, Table VI):				
Mazda 323	-0.45	-0.69	-1.55	-1.77
Honda Accord	-0.82	-1.26	-1.47	-1.42
Acura Legend	-1.69	-2.57	-4.17	-4.24
BMW 735i	-3.32	-5.09	-7.21	-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.

10 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 that adds estimated regressors to the setting of Berry, Linton, and Pakes (2004). 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.

Appendix

A Consistency Theorem for Random Coefficients Logit Models

A.1 Proof of General Consistency (Theorem 4)

In proving Theorem 4 we use a strategy 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:

$$Q_J(\delta^*(\tilde{\theta}_\lambda, s^0, P^0), z, p, \hat{\mathbf{v}}; \tilde{\theta}, \tilde{f}) = \inf_{(\theta,\theta_\lambda,f)\in\Theta\times\Theta_\lambda\times\mathcal{F}_J} Q_J(\delta^*(\theta_\lambda, s^0, P^0), z, p, \hat{\mathbf{v}}; \theta, f) + o_p(1). \tag{33}$$

Let $\varepsilon > 0$ be any small real numbers. Note that any estimator $(\tilde{\theta}, \tilde{\theta}_\lambda, \tilde{f})$ satisfying (33) also satisfies that with probability approaching to one (w.p.a.1), $Q_J(\delta^*(\tilde{\theta}_\lambda, s^0, P^0), \cdot, \hat{\mathbf{v}}; \tilde{\theta}, \tilde{f}) < Q_J(\delta^*(\theta_\lambda, s^0, P^0), \cdot, \hat{\mathbf{v}}; \theta, f_J) + \frac{\varepsilon}{6}$ 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

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

Then by Assumption 10, the consistency of the pre-stage estimators (Assumption 7) and Assumption 8, we have w.p.a.1, $Q_J^0(\delta^*(\tilde{\theta}_\lambda, s^0, P^0), \cdot, \mathbf{v}; \tilde{\theta}, \tilde{f}) - Q_J(\delta^*(\tilde{\theta}_\lambda, s^0, P^0), \cdot, \hat{\mathbf{v}}; \tilde{\theta}, \tilde{f}) < \frac{\varepsilon}{6}$ and

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

It follows that w.p.a.1,

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

Next note by the continuity assumption (Assumption 12) and the consistency of the pre-stage estimators (Assumption 7), we have w.p.a.1, $Q_J^0(\delta^*(\tilde{\theta}_\lambda, s^0, P^0), \cdot, \mathbf{v}; \tilde{\theta}, \tilde{f}) - Q_J^0(\delta^*(\tilde{\theta}_\lambda, s^0, P^0), \cdot, \hat{\mathbf{v}}; \tilde{\theta}, \tilde{f}) < \frac{\varepsilon}{6}$ and $Q_J^0(\delta^*(\theta_{\lambda 0}, s^0, P^0), \cdot, \mathbf{v}; \theta_0, \pi_J f_0) - Q_J^0(\delta^*(\theta_{\lambda 0}, s^0, P^0), \cdot, \hat{\mathbf{v}}; \theta_0, \pi_J f_0) > -\frac{\varepsilon}{6}$. It follows that w.p.a.1,

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

Then by Assumption 15 and Assumption 11 (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

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

It follows that

$$Q_J^0(\delta^*(\tilde{\theta}_\lambda, s^0, P^0), z, p, \mathbf{v}; \tilde{\theta}, \tilde{f}) < Q_J^0(\delta^*(\theta_{\lambda 0}, s^0, P^0), z, p, \mathbf{v}; \theta_0, f_0) + \varepsilon. \tag{34}$$

Next note that for any $\epsilon > 0$, by Assumption 9, Assumption 11 (continuity), 13 (compactness),

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

exists (it can vary by J). Then by Assumption 15 and the fact that $\mathcal{F}_J \subset \mathcal{F}$, it must be that

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

Take ε small enough that

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

Then from (34) it follows that w.p.a.1,

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

Then by Assumption 11 (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 (33) 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 (33) :

$$\begin{aligned}
& Q_J(\delta^*(\hat{\theta}_\lambda, s^0, P^0), z, p, \hat{\mathbf{v}}; \hat{\theta}, \hat{f}) = Q_J(\delta^*(\hat{\theta}_\lambda, s^n, P^R), z, p, \hat{\mathbf{v}}; \hat{\theta}, \hat{f}) + o_p(1) && (35)\\
\leq\ & \inf_{(\theta,\theta_\lambda,f)\in\Theta\times\Theta_\lambda\times\hat{\mathcal{F}}_J} Q_J(\delta^*(\theta_\lambda, s^n, P^R), z, p, \hat{\mathbf{v}}; \theta, f) + o_p(1) && (36)\\
=\ & \inf_{(\theta,\theta_\lambda,f)\in\Theta\times\Theta_\lambda\times\mathcal{F}_J} Q_J(\delta^*(\theta_\lambda, s^n, P^R), z, p, \hat{\mathbf{v}}; \theta, f) + o_p(1) && (37)\\
=\ & \inf_{(\theta,\theta_\lambda,f)\in\Theta\times\Theta_\lambda\times\mathcal{F}_J} Q_J(\delta^*(\theta_\lambda, s^0, P^0), z, p, \hat{\mathbf{v}}; \theta, f) + o_p(1) && (38)
\end{aligned}$$

where (36) (the first inequality) holds because $(\hat{\theta}, \hat{\theta}_\lambda, \hat{f})$ is an extremum estimator satisfying (21) and (37) (the second equality) holds because $Q_J(\cdot, \hat{\mathbf{v}}; \theta, f)$ is continuous in f and because for any $f \in \mathcal{F}_J$ we can find a sequence $\hat{f} \in \hat{\mathcal{F}}_J$ such that $\left\|\hat{f} - f\right\|_{\mathcal{F}} \to 0$ as $\hat{\Pi}(\cdot) \to \Pi(\cdot)$ and $\hat{\bar{\varphi}}_l(\cdot) \to \bar{\varphi}_l(\cdot)$ (in a pseudo-metric $\|\cdot\|_s$) by Assumption 7. We focus on (35) (the first equality) and (38) (the last equality). Consider that by applying the Cauchy-Schwarz inequality twice we obtain

$$\begin{aligned}
& \sup{}_{(\theta,\theta_\lambda,f)\in\Theta\times\Theta_\lambda\times(\hat{\mathcal{F}}_J\cup\mathcal{F}_J)} |Q_J(\delta^*(\theta_\lambda, s^n, P^R), \cdot, \hat{\mathbf{v}}; \theta, f) - Q_J(\delta^*(\theta_\lambda, s^0, P^0), \cdot, \hat{\mathbf{v}}; \theta, f)| && (39)\\
\leq\ & \sup{}_{(\theta,\theta_\lambda,f)\in\Theta\times\Theta_\lambda\times(\hat{\mathcal{F}}_J\cup\mathcal{F}_J)} J^{-1}\sum_{j=1}^{J}(\hat{E}[\delta_j^*(\theta_\lambda, s^n, P^R)|z_j, \hat{\mathbf{v}}_j] - \hat{E}[\delta_j^*(\theta_\lambda, s^0, P^0)|z_j, \hat{\mathbf{v}}_j])^2 \times\\
& \times 2\sup{}_{(\theta,\theta_\lambda,f)\in\Theta\times\Theta_\lambda\times(\hat{\mathcal{F}}_J\cup\mathcal{F}_J)} (Q_J(\delta^*(\theta_\lambda, s^n, P^R), \cdot, \hat{\mathbf{V}}; \theta, f) + Q_J(\delta^*(\theta_\lambda, s^0, P^0), \cdot, \hat{\mathbf{V}}; \theta, f))\\
\leq\ & C \cdot \sup{}_{\theta_\lambda\in\Theta_\lambda} J^{-1}\sum_{j=1}^{J}(\hat{E}[\delta_j^*(\theta_\lambda, s^n, P^R)|z_j, \hat{\mathbf{v}}_j] - \hat{E}[\delta_j^*(\theta_\lambda, s^0, P^0)|z_j, \hat{\mathbf{v}}_j])^2
\end{aligned}$$

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 13), 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 $\delta^*(\theta_\lambda,\cdot)$ does not depend on (θ,f).

Therefore (39) is $o_p(1)$ if

$$\sup_{\theta_\lambda\in\Theta_\lambda} J^{-1}\sum_{j=1}^{J}(\hat{E}[\delta_j^*(\theta_\lambda,s^n,P^R)|z_j,\hat{\mathbf{v}}_j]-\hat{E}[\delta_j^*(\theta_\lambda,s^0,P^0)|z_j,\hat{\mathbf{v}}_j])^2=o_p(1). \tag{40}$$

This in turn implies (35) immediately and also implies (38) by the triangle inequality as we argue below. Let $Q_J(\delta^*(\theta_\lambda^{(1)},s^n,P^R),z,p,\hat{\mathbf{v}};\theta^{(1)},f^{(1)}) = \inf_{(\theta,\theta_\lambda,f)\in\Theta\times\Theta_\lambda\times\mathcal{F}_J} Q_J(\delta^*(\theta_\lambda,s^n,P^R),z,p,\hat{\mathbf{v}};\theta,f)$ and $Q_J(\delta^*(\theta_\lambda^{(2)},s^0,P^0),z,p,\hat{\mathbf{v}};\theta^{(2)},f^{(2)}) = \inf_{(\theta,\theta_\lambda,f)\in\Theta\times\Theta_\lambda\times\mathcal{F}_J} Q_J(\delta^*(\theta_\lambda,s^0,P^0),z,p,\hat{\mathbf{v}};\theta,f)$. 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 13). It follows that

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

where the first and the last equality hold by (39) and (40). Above the first inequality holds because $(\theta^{(2)},\theta_\lambda^{(2)},f^{(2)})$ minimizes $Q_J(\delta^*(\theta_\lambda,s^0,P^0),z,p,\hat{\mathbf{v}};\theta,f)$ over $\Theta\times\Theta_\lambda\times\mathcal{F}_J$ and the second inequality holds because $(\theta^{(1)},\theta_\lambda^{(1)},f^{(1)})$ minimizes $Q_J(\delta^*(\theta_\lambda,s^n,P^R),z,p,\hat{\mathbf{v}};\theta,f)$ over $\Theta\times\Theta_\lambda\times\mathcal{F}_J$. This proves (38).

Finally we verify (40) is $o_p(1)$. Note that

$$\begin{aligned}
J^{-1}\textstyle\sum_{j=1}^{J}(&\hat{E}[\delta_j^*(\theta_\lambda,s^n,P^R)|z_j,\hat{\mathbf{v}}_j]-\hat{E}[\delta_j^*(\theta_\lambda,s^0,P^0)|z_j,\hat{\mathbf{v}}_j])^2 \qquad (41)\\
= \quad & J^{-1}\sum_{j=1}^{J}\{\delta_j^*(\theta_\lambda,s^n,P^R)-\delta_j^*(\theta_\lambda,s^0,P^0)\}\varphi^L(z_j,\hat{\mathbf{v}}_j)'(\sum_{j=1}^{J}\varphi^L(z_j,\hat{\mathbf{v}}_j)\varphi^L(z_j,\hat{\mathbf{v}}_j)'/J)^{-1}\\
&\times\sum_{j=1}^{J}\varphi^L(z_j,\hat{\mathbf{v}}_j)\{\delta_j^*(\theta_\lambda,s^n,P^R)-\delta_j^*(\theta_\lambda,s^0,P^0)\}/J\\
\le \quad & O_p(1)||\sum_{j=1}^{J}\varphi^L(z_j,\hat{\mathbf{v}}_j)\{\delta_j^*(\theta_\lambda,s^n,P^R)-\delta_j^*(\theta_\lambda,s^0,P^0)\}/J||^2\\
\le \quad & O_p(1)\sum_{j=1}^{J}||\varphi^L(z_j,\hat{\mathbf{v}}_j)||^2/J\sum_{j=1}^{J}\{\delta_j^*(\theta_\lambda,s^n,P^R)-\delta_j^*(\theta_\lambda,s^0,P^0)\}^2/J\\
\le \quad & O_p(1)\zeta_\varphi(L)^2\cdot\sum_{j=1}^{J}\{\delta_j^*(\theta_\lambda,s^n,P^R)-\delta_j^*(\theta_\lambda,s^0,P^0)\}^2/J=o_p(1)
\end{aligned}$$

where the first inequality holds because $\sum_{j=1}^{J}\varphi^L(z_j,\hat{\mathbf{v}}_j)\varphi^L(z_j,\hat{\mathbf{v}}_j)'/J$ becomes nonsingular w.p.a.1 (Assumption 7 and 8 (ii)) and the last result holds by the essentially same proof of A.2 (page 647-648) in the proof of Theorem 1 of Berry, Linton, and Pakes (2004) under Assumption 5 and

Assumption 14 because (i) all arguments there in terms of ξ also hold in terms of our δ^* and (ii) Assumption 5 replaces their Assumption A3 and Assumption 14 replaces their Assumption A5.

This completes the proof.

B Consistency Theorem for the Simple Logit

We show consistency of our multi-step sieve estimator for the simple logit case.

Denote a sample objective function $Q_J(\delta, z, p, \hat{\mathbf{v}}; \theta, f)$ for estimation based on the moment condition of (8). If we use nonlinear sieve least squares estimation, then the objective function for estimation 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 $(\theta, f) \in \Theta \times \hat{\mathcal{F}}_J$. Our estimator is minimizing the sample objective function

$$(\hat{\theta}, \hat{f}) = \operatorname{arginf}_{(\theta,f)\in\Theta\times\hat{\mathcal{F}}_J} Q_J(\delta, z, p, \hat{\mathbf{v}}; \theta, f) + o_p(1). \tag{42}$$

We also define the corresponding population objective function as

$$Q_J^0(\delta, z, p, \mathbf{v}; \theta, f) = \frac{1}{J}\sum_{m=1}^{M}\sum_{j=1}^{J_m} E[\{\delta_{mj} - (c + \beta' x_{mj} - \alpha p_{mj} + f(z_j, \mathbf{v}_j)(1 + \gamma' x_{mj} + \gamma_p(\bar{y}_m - p_{mj})))\}^2]. \tag{43}$$

The consistency theorem below holds either when the asymptotics is in the number of products ($J_m \to \infty$) or in the number of markets ($M \to \infty$). Note that we do not require $Q_J^0(\delta, z, p, \mathbf{v}; \theta, f)$ converges when the asymptotics is in the number of products while the convergence typically holds when the asymptotics is in the number of markets. In the latter case requirements for the consistency can be further simplified.

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 (2007).[22] Here we abstract from the sampling error in the market shares to save notation. We have allowed it for the random coefficients logit case. The contribution of this sampling error to the variance of the estimator will be negligible when the market size (number of consumers) is large. The following assumptions are commonly imposed and standard in the sieve estimation literature and we have already discussed conditions related to them for the random coefficient logit case, so we minimize our discussion. We can show or have shown most of assumptions below are satisfied for the logit case. We list the following assumptions for transparency or possible application of our theorem to other estimation problems.

As we have shown in Section 5, first we require identification

[22] Our problem is different from Newey and Powell (2003)'s Theorem 4.1 because we use estimated regressors (functions, $\hat{\Pi}(\cdot)$ and $\hat{\hat{\varphi}}_l(\cdot)$) 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.

Assumption 21 (B1)**.** *$(\theta_0, f_0) \in \Theta \times \mathcal{F}$ is the only $(\theta, f) \in \Theta \times \mathcal{F}^{23}$ satisfying the moment condition (8) and (9) and $Q_J^0(\delta, z, p, \mathbf{v}; \theta_0, f_0) < \infty$.*

Next we note our estimator is an extremum estimator solving (42), so satisfies

Assumption 22 (B2)**.** $Q_J(\delta, z, p, \hat{\mathbf{v}}; \hat{\theta}, \hat{f}) \leq inf_{(\theta,f)\in\Theta\times\hat{\mathcal{F}}_J} Q_J(\delta, z, p, \hat{\mathbf{v}}; \theta, f) + o_p(1)$

Assumption B3 below says that both $\Pi_0(\cdot)$ and $\bar{\varphi}_{0l}(\cdot)$ can be approximated by the first stage and the middle stage series approximations. Again this is well 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 23 (B3)**.** $\left\|\hat{\Pi}(\cdot) - \Pi_0(\cdot)\right\|_s = o_p(1)$ *and* $\left\|\hat{\bar{\varphi}}_l(\cdot) - \bar{\varphi}_{0l}(\cdot)\right\|_s = o_p(1)$ *for all l.*

Assumption 24 (B4)**.** *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 $J \to \infty$.*

Assumption B4 is also known to hold if $\mathcal{F}$ is a set of a class of smooth functions such as Hölder class.

The following continuity conditions obviously hold for our objective function.

Assumption 25 (B5)**.** *$Q_J^0(\delta, z, p, \mathbf{v}; \theta, f)$ is continuous in $(\theta, f) \in \Theta \times \mathcal{F}$.*

Note that Assumption B5 is trivial from observing the construction of $Q_J^0(\delta, z, p, \mathbf{v}; \theta, f)$ in (43).

Assumption 26 (B6)**.** *$Q_J^0(\delta, z, p, \mathbf{v}; \theta, f_J)$ is continuous in $\Pi(\cdot)$ and $\bar{\varphi}_l(\cdot)$ uniformly for all $(\theta, f_J) \in \Theta \times \mathcal{F}_J$.*

Assumption B6 is also trivially satisfied because any $f_J \in \mathcal{F}_J$ is continuous in $\Pi(\cdot)$ and $\bar{\varphi}_l(\cdot)$ by construction of $\mathcal{F}_J$ and because $\Pi(\cdot)$ and $\bar{\varphi}_l(\cdot)$ enter $Q_J^0(\delta, z, p, \mathbf{v}; \theta, f_J)$ only by f_J and $Q_J^0(\delta, z, p, \mathbf{v}; \theta, f_J)$ is continuous in f_J.

Next we impose compactness on the sieve space.

Assumption 27 (B7)**.** *The parameter space Θ is compact and the sieve space, $\mathcal{F}_J$, is compact under the pseudo-metric $||\cdot||_{\mathcal{F}}$.*

This compactness condition holds when the sieve space is based on power series or splines as in our construction.

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.

Assumption 28 (B8)**.** *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 $\mathbf{v}_{mj} = g_j(p_{m1} - \Pi(z_{m1}), \ldots, p_{mJ_m} - \Pi(z_{mJ_m}))$.

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

Note that Assumption B8 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 Assumption B8 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 $\|\bar{\varphi}_l - \bar{\varphi}_{0l}\|_s \leq \epsilon_J$.

Theorem 7. *Suppose Assumptions B1-B8 are satisfied. Then* $\hat{\theta} \to_p \theta_0$.

B.1 Proof of Theorem 7

We prove the consistency by extending Chen (2007)'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:

$$Q_J(\delta, z, p, \hat{\mathbf{v}}; \tilde{\theta}, \tilde{f}) \leq \inf_{(\theta,f)\in\Theta\times\mathcal{F}_J} Q_J(\delta, z, p, \hat{\mathbf{v}}; \theta, f) + o_p(1). \tag{44}$$

Let $\varepsilon > 0$ be any small real numbers. Any estimator $(\tilde{\theta}, \tilde{f})$ that satisfies (44) 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 B8 and the consistency of the pre-stage estimators (B3), 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,

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

Next we note that by the continuity assumption (B6) and the consistency of the pre-stage estimators (B3), 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 B1 and B5 (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

$$Q_J^0(\delta, z, p, \mathbf{v}; \tilde{\theta}, \tilde{f}) < Q_J^0(\delta, z, p, \mathbf{v}; \theta_0, f_0) + \varepsilon. \tag{45}$$

Next note that for any $\epsilon > 0$, by B4, B5(continuity), B7 (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 B1 (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 (45) 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 B5 (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 (44) 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 B2 (extremum estimator) and the second equality holds because $Q_J(\delta, z, p, \hat{\mathbf{v}}; \theta, f)$ is continuous in f and because for any $f \in \mathcal{F}_J$ we can find a sequence $\hat{f} \in \hat{\mathcal{F}}_J$ such that $\left\|\hat{f} - f\right\|_{\mathcal{F}} \to 0$ as $\hat{\Pi}(\cdot) \to \Pi(\cdot)$ and $\hat{\bar{\varphi}}_l(\cdot) \to \bar{\varphi}_l(\cdot)$ (in a pseudo-metric $\|\cdot\|_s$) by Assumption B3.

C Convergence Rate of the Estimator

Following up the consistency, we derive the mean-squared error convergence rates of the estimator $\hat{f}(\cdot)$, which will be useful to obtain the $\sqrt{J}$-consistency and the asymptotic normality of the estimate of interest, $(\hat{\theta}, \hat{\theta}_\lambda)$.

Regularity conditions we impose here are standard in the sieve estimation literature. We first introduce some notation. Let

$$g_0(z_j, \mathbf{v}_j) = c_0 + \beta_0' x_j - \alpha p_j + f_0(z_j, \mathbf{v}_j)(1 + \gamma_0' x_j + \gamma_p(\bar{y} - p_j))$$

and define $\varsigma_j(\theta_\lambda, s^n, p^R) = \delta_j^*(\theta_\lambda, s^n, P^R) - g_0(z_j, \mathbf{v}_j)$ and $\varsigma_j(\theta_\lambda) = \delta_j^*(\theta_\lambda, s^0, P^0) - g_0(z_j, \mathbf{v}_j)$. Here we let $g_0(z_j, \mathbf{v}_j)$ be a function of $(z_j, \mathbf{v}_j)$ because x_j is included in z_j. For a matrix D, let $\|D\| = (\mathrm{tr}(D'D))^{1/2}$, for a random matrix D, we let $\|D\|_\infty$ be the infimum of constants C such that $\Pr(||D|| < C) = 1$. We also assume that the supports of distributions of p, $\mathbf{V}$, and Z are compact to avoid other complications but this can be relaxed with additional complexity (e.g., trimming devices).

In addition to Assumption C1 we impose the rate conditions that restrict the growth of $k(J)$ and $L(J)$ as J tends to infinity.

Assumption 29 (C2). *Let* $\triangle_{J,1} = k(J)^{1/2}/\sqrt{J} + k(J)^{-s_\Pi/\dim(z)}$, $\triangle_{J,2} = k(J)^{1/2}/\sqrt{J} + k(J)^{-s_\varphi/\dim(z)}$, *and* $\Delta_J = \max\{\Delta_{J,1}, \Delta_{J,2}\} \to 0$. *Also* $\triangle_\delta = L(J)^{1/2}/\sqrt{J} + L(J)^{-s_\delta/\dim(z,\mathbf{v})} \to 0$. *For polynomial approximations* $k(J)^3/J \to 0$, $L(J)^3/J \to 0$, $\sqrt{L^3/J} + \triangle_J(L(J)^4 + L(J)^2 k(J)^{3/2}/\sqrt{J}) + L(J)^{1-s_f/\dim(z,\mathbf{v})} \to 0$ *and for the spline approximations* $k(J)^2/J \to 0$, $L(J)^2/J \to 0$, $\sqrt{L^3/J} + \triangle_J(L(J)^{5/2} + L(J)^2 k(J)/\sqrt{J}) + L(J)^{1-s_f/\dim(z,\mathbf{v})} \to 0$.

Then we obtain the mean-squared error convergence of $\hat{f}(\cdot)$:

Theorem 8. *Suppose Assumptions 3-6, 9-10, 13-15, Condition S, and Assumptions C1-C2 are satisfied. Suppose J^2/n and J^2/R are bounded. Then*

$$\left(\int(\hat{f}(z,\mathbf{v})-f_0(z,\mathbf{v}))^2 d\mu_0(z,\mathbf{v})\right)^{1/2} \;=\; O_p(\sqrt{L(J)/J}+L(J)\triangle_J+L(J)^{-s_f/\dim(z,\mathbf{v})})$$

where $\mu_0(z,\mathbf{v})$ denotes the distribution function of $(Z,\mathbf{V})$.

Note that one would obtain the convergence rate $O(\sqrt{L(J)/J}+L(J)^{-s_f/\dim(z,\mathbf{v})})$ (e.g., Newey (1997)) if the first and the second step estimations are not required.

C.1 Proof of Theorem 8

We introduce notation and prove Lemma L1 below that is useful to derive the convergence rate result.

Define $f_L(z,\mathbf{v})=a_L'\tilde{\varphi}^L(z,\mathbf{v})$ and $\hat{\tilde{f}}_L(z,\mathbf{v})=a_L'\hat{\tilde{\varphi}}^L(z,\mathbf{v})$ where a_L satisfies Assumption L1 (iv) below. Define $\psi^L_{\theta 0,j}\equiv(1,x_j',-p_j,x_j'f_0(z_j,\mathbf{v}_j),(\bar{y}-p_j)f_0(z_j,\mathbf{v}_j),(1+\gamma_0'x_j+\gamma_{p0}(\bar{y}-p_j))\tilde{\varphi}^L(z_j,\mathbf{v}_j)')'$, $\psi^L_\theta(z_j,\mathbf{v}_j)=(1,x_j',-p_j,x_j'f_L(z_j,\mathbf{v}_j),(\bar{y}-p_j)f_L(z_j,\mathbf{v}_j),(1+\gamma_0'x_j+\gamma_{p0}(\bar{y}-p_j))\tilde{\varphi}^L(z_j,\mathbf{v}_j)')'$, $\hat{\psi}^L_\theta(z_j,\mathbf{v}_j)=(1,x_j',-p_j,x_j'\hat{\tilde{f}}_L(z_j,\mathbf{v}_j),(\bar{y}-p_j)\hat{\tilde{f}}_L(z_j,\mathbf{v}_j),(1+\gamma_0'x_j+\gamma_{p0}(\bar{y}-p_j))\hat{\tilde{\varphi}}^L(z_j,\mathbf{v}_j)')'$, and $\hat{\Psi}^L_{\theta,j}=(1,x_j',-p_j,x_j'\hat{f}(z_j,\hat{\mathbf{v}}_j),(\bar{y}-p_j)\hat{f}(z_j,\hat{\mathbf{v}}_j),(1+\hat{\gamma}'x_j+\hat{\gamma}_p(\bar{y}-p_j))\hat{\tilde{\varphi}}^L(z_j,\hat{\mathbf{v}}_j)')'$. We further let $\hat{\hat{\psi}}^L_{\theta,j}=\hat{\psi}^L_\theta(z_j,\hat{\mathbf{v}}_j)$, $\psi^L_{\theta,j}=\psi^L_\theta(z_j,\mathbf{v}_j)$, and $\hat{\psi}^L_{\theta,j}=\hat{\psi}^L_\theta(z_j,\mathbf{v}_j)$. Let $\hat{\Psi}_{\theta_\lambda,j}=\hat{E}[\frac{\partial\delta_j^*(\theta_\lambda,s^n,P^R)}{\partial\theta_\lambda}|z_j,\hat{\mathbf{v}}_j]$ and $\Psi_{\theta_\lambda,j}=E\left[\frac{\partial\delta_j^*(\theta_\lambda,s^0,P^0)}{\partial\theta_\lambda}|z_j,\mathbf{v}_j\right]$. Let $\hat{\Psi}^L_j=(\hat{\Psi}'_{\hat{\theta}_\lambda,j},-\hat{\Psi}^{L\prime}_{\theta,j})'$ and $\hat{\Psi}^{L,J}=(\hat{\Psi}^L_1,\ldots,\hat{\Psi}^L_J)'$. Similarly we let $\psi^{L,J}=(\psi^L_1,\ldots,\psi^L_J)'$ with $\psi^L_j=(\Psi'_{\theta_{\lambda 0},j},-\psi^{L\prime}_{\theta,j})'$, $\hat{\psi}^{L,J}=(\hat{\psi}^L_1,\ldots,\hat{\psi}^L_J)'$ with $\hat{\psi}^L_j=(\Psi'_{\theta_{\lambda 0},j},-\hat{\psi}^{L\prime}_{\theta,j})'$, and $\hat{\hat{\psi}}^{L,J}=(\hat{\hat{\psi}}^L_1,\ldots,\hat{\hat{\psi}}^L_J)'$ with $\hat{\hat{\psi}}^L_j=(\Psi'_{\theta_{\lambda 0},j},-\hat{\hat{\psi}}^{L\prime}_{\theta,j})'$.

Let C (also C_1,C_2, and others) denote a generic positive constant and let $C(p,x)$ and $C(z,\mathbf{v})$ (also $C_1(\cdot)$, $C_2(\cdot)$, and others) denote generic bounded positive function of (p,x) and $(z,\mathbf{v})$ respectively. We often write $C_j=C(p_j,x_j)$ or $C_j=C(z_j,\mathbf{v}_j)$. We let $\mathcal{W}=\mathcal{Z}\times\mathcal{V}$ (the support of $(Z,\mathbf{V})$).

Assumption 30 (L1)**.** *(i) $(p_j,Z_j,\mathbf{V}_j)$ is continuously distributed with bounded density; (ii) (a) For each k and L there are nonsingular matrices B, $\tilde{B}$, and B_1 such that for $\varphi^k_{B_1}(z)=B_1\varphi^k(z)$, $\varphi^L_B(z,\mathbf{v})=B\varphi^L(z,\mathbf{v})$ and $\tilde{\varphi}^L_{\tilde{B}}(z,\mathbf{v})=\tilde{B}\tilde{\varphi}^L(z,\mathbf{v})$, $\sum_{j=1}^J E[\varphi^L_B(Z_j,\mathbf{V}_j)\varphi^L_B(Z_j,\mathbf{V}_j)']/J$, $\sum_{j=1}^J E[\tilde{\varphi}^L_{\tilde{B}}(Z_j,\mathbf{V}_j)\tilde{\varphi}^L_{\tilde{B}}(Z_j,\mathbf{V}_j)']/J$ and $\sum_{j=1}^J E[\varphi^k_{B_1}(Z_j)\varphi^k_{B_1}(Z_j)']/J$ have smallest eigenvalues that are bounded away from zero for all J large enough, uniformly in k and L; (ii) (b) Let $\Psi^L_{0,j}=(\Psi'_{\theta_{\lambda 0},j},-\psi^{L\prime}_{\theta 0,j})'$. Then for each k and L, $\sum_{j=1}^J E[\Psi^L_{0,j}\Psi^{L\prime}_{0,j}]/J$ has a smallest eigenvalue that is bounded away from zero for all J large enough, uniformly in k and L; (iii) For each integer $\iota>0$, there are $\zeta_\iota(L)$ and $\zeta_\iota(k)$ with $|\varphi^L(z,\mathbf{v})|_\iota\le\zeta_\iota(L)$, $|\tilde{\varphi}^L(z,\mathbf{v})|_\iota\le\zeta_\iota(L)$, and $|\varphi^k(z)|_\iota\le\zeta_\iota(k)$; (iv) There exist ι, ϱ_1, ϱ_2, ϱ, $\varrho_\delta>0$ and a_L, b_L, B_L, λ^1_k, and $\lambda^2_{l,k}$ such that $|\Pi_0(z)-\lambda_k^{1\prime}\varphi^k(z)|_\iota=Ck^{-\varrho_1}$, $|\bar{\varphi}_{0l}(z)-\lambda^{2\prime}_{l,k}\varphi^k(z)|_\iota=Ck^{-\varrho_2}$ for all l, $|f_0(z,\mathbf{v})-a_L'\tilde{\varphi}^L(z,\mathbf{v})|_\iota=CL^{-\varrho}$, $\sum_j|E[\delta_j^*(\theta_\lambda,s^0,P^0)|z_j,\mathbf{v}_j]-b_L'\varphi^L(z_j,\mathbf{v}_j)|_\iota/J=C(L^{-\varrho})$ and $\sum_j|E[\frac{\partial\delta_j^*(\theta_\lambda,s^0,P^0)}{\partial\theta_\lambda}|z_j,\mathbf{v}_j]-B_L\varphi^L(z_j,\mathbf{v}_j)|_\iota/J=C(L^{-\varrho})$; (v) both $\mathcal{Z}$ and the support of p are compact.*

Let $\triangle_{J,1}=k(J)^{1/2}/\sqrt{J}+k(J)^{-\varrho_1}$, $\triangle_{J,2}=k(J)^{1/2}/\sqrt{J}+k(J)^{-\varrho_2}$, and $\Delta_J=\max\{\Delta_{J,1},\Delta_{J,2}\}$.

Lemma 1 (L1)**.** *Suppose Assumptions 3-6, 9-10, 13-15, Condition S, Assumptions L1, and Assumptions C1 (i), (iv)-(ix) hold. Suppose J^2/n and J^2/R are bounded. Further suppose (i) $\zeta_0(k)\sqrt{k/J} \to 0$, (ii) $\zeta_0(L)\sqrt{L/J} \to 0$, and (iii)*

$$\begin{aligned} L\triangle_J^{\varphi} &= (L\zeta_1(L) + L^{3/2}\zeta_0(k)\sqrt{k/J} + L^{3/2})\triangle_J \to 0 \\ L\triangle_{J,\vartheta} &= \sqrt{L^3/J} + L^2\zeta_0(k)\triangle_{J,1}\sqrt{k/J} + L^2\triangle_{J,2} + L^{1-\varrho} \to 0. \end{aligned}$$

Then,

$$(\sum\nolimits_{j=1}^{J}\left(\hat{f}(z_j,\mathbf{v}_j) - f_0(z_j,\mathbf{v}_j)\right)^2/J) = O_p(\sqrt{L/J} + L\zeta_0(k)\triangle_{J,1}\sqrt{k/J} + L\triangle_{J,2} + L^{-\varrho}).$$

C.1.1 Proof of Lemma L1

Without loss of generality, we will let $\varphi^k(z) = \varphi^k_{B_1}(z)$, $\varphi^L(z,\mathbf{v}) = \varphi^L_B(z,\mathbf{v})$ and $\tilde{\varphi}^L(z,\mathbf{v}) = \tilde{\varphi}^L_{\tilde{B}}(z,\mathbf{v})$. Let $\hat{\Pi}_j = \hat{\Pi}(z_j)$ and $\Pi_j = \Pi_0(z_j)$. Let $\hat{\bar{\varphi}}_{l,j} = \hat{\bar{\varphi}}_l(z_j)$ and $\bar{\varphi}_{l,j} = \bar{\varphi}_{0l}(z_j)$. Let $\hat{\tilde{\varphi}}_{l,j} = \hat{\tilde{\varphi}}_l(z_j,\hat{\mathbf{v}}_j)$ and $\tilde{\varphi}_{l,j} = \tilde{\varphi}_l(z_j,\mathbf{v}_j)$. Also let $\hat{\tilde{\varphi}}^L_j = \hat{\tilde{\varphi}}^L(z_j,\hat{\mathbf{v}}_j)$ and $\tilde{\varphi}^L_j = \tilde{\varphi}^L(z_j,\mathbf{v}_j)$. Further define $\dot{\bar{\varphi}}_l(z) = \varphi^k(z)'(\mathbf{P}'\mathbf{P})^- \sum_{j=1}^J \varphi^k(z_j)\varphi_l(z_j,\mathbf{v}_j)$ where we have $\hat{\bar{\varphi}}_l(z) = \varphi^k(z)'(\mathbf{P}'\mathbf{P})^- \sum_{j=1}^J \varphi^k(z_j)\varphi_l(z_j,\hat{\mathbf{v}}_j)$. Let $\dot{\bar{\varphi}}^L(z) = (\dot{\bar{\varphi}}_1(z),\ldots,\dot{\bar{\varphi}}_L(z))'$ and $\bar{\varphi}^L(z) = (\bar{\varphi}_1(z),\ldots,\bar{\varphi}_L(z))'$. We also let $\varphi^L(z_j,\hat{\mathbf{v}}_j) = (\varphi_1(z_j,\hat{\mathbf{v}}_j),\ldots,\varphi_L(z_j,\hat{\mathbf{v}}_j))'$ and $\varphi^L(z_j,\mathbf{v}_j) = (\varphi_1(z_j,\mathbf{v}_j),\ldots,\varphi_L(z_i,\mathbf{v}_j))'$.

First note $(\mathbf{P}'\mathbf{P})/J$ becomes nonsingular w.p.a.1[24] as $\zeta_0(k)^2k/J \to 0$ by Assumption L1 (ii) and by the essentially same proof in Theorem 1 of Newey 1997. Then by the essentially same proof (A.3) of Lemma A1 in Newey, Powell, and Vella (1999)), we obtain

$$\sum\nolimits_{j=1}^{J}||\hat{\Pi}_j - \Pi_j||^2/J = O_p(\triangle_{J,1}^2) \text{ and } \sum\nolimits_{j=1}^{J}||\dot{\bar{\varphi}}_{l,j} - \bar{\varphi}_{l,j}||^2/J = O_p(\triangle_{J,2}^2) \text{ for all } l. \quad (46)$$

Also by a similar argument to Theorem 1 of Newey (1997), it follows that

$$\begin{aligned} \max_{j\le J}||\hat{\Pi}_j - \Pi_j|| &= O_p(\zeta_0(k)\triangle_{J,1}) && (47) \\ \max_{j\le J}||\dot{\bar{\varphi}}_{l,j} - \bar{\varphi}_{l,j}|| &= O_p(\zeta_0(k)\triangle_{J,2}) \text{ for all } l. && (48) \end{aligned}$$

Again by the essentially same proof (A.3) of Lemma A1 in Newey, Powell, and Vella (1999)), we obtain

$$\begin{aligned} &\sum\nolimits_{j=1}^{J}||\hat{E}[\frac{\partial\delta_j^*(\theta_\lambda,s^0,P^0)}{\partial\theta_\lambda}|z_j,\mathbf{v}_j] - E[\frac{\partial\delta_j^*(\theta_\lambda,s^0,P^0)}{\partial\theta_\lambda}|z_j,\mathbf{v}_j]||^2/J \\ = &\ O_p(\triangle_\delta \equiv L(J)^{1/2}/\sqrt{J} + L(J)^{-\rho_\delta}) = o_p(1). \end{aligned}$$

[24] In the sense that there exists a $C(\epsilon)$ such that $\Pr(||(\mathbf{P}'\mathbf{P})/J - I|| > C(\epsilon)) < \epsilon$ for all J large enough. Others are similarly defined.

Then because $E[\frac{\partial\delta_j^*(\theta_\lambda,s^0,P^0)}{\partial\theta_\lambda}|z_j,\mathbf{v}_j]$ is Lipschitz in $\mathbf{v}_j$ and by Assumption C1 (viii) we further obtain

$$\begin{aligned}
&\sum\nolimits_{j=1}^J ||\hat{E}[\frac{\partial\delta_j^*(\theta_\lambda,s^0,P^0)}{\partial\theta_\lambda}|z_j,\hat{\mathbf{v}}_j]-E[\frac{\partial\delta_j^*(\theta_\lambda,s^0,P^0)}{\partial\theta_\lambda}|z_j,\mathbf{v}_j]||^2/J \qquad (49)\\
\leq\ & 2\sum\nolimits_{j=1}^J ||\hat{E}[\frac{\partial\delta_j^*(\theta_\lambda,s^0,P^0)}{\partial\theta_\lambda}|z_j,\hat{\mathbf{v}}_j]-E[\frac{\partial\delta_j^*(\theta_\lambda,s^0,P^0)}{\partial\theta_\lambda}|z_j,\hat{\mathbf{v}}_j]||^2/J\\
&+2\sum\nolimits_{j=1}^J ||E[\frac{\partial\delta_j^*(\theta_\lambda,s^0,P^0)}{\partial\theta_\lambda}|z_j,\hat{\mathbf{v}}_j]-E[\frac{\partial\delta_j^*(\theta_\lambda,s^0,P^0)}{\partial\theta_\lambda}|z_j,\mathbf{v}_j]||^2/J\\
=\ & \sum\nolimits_{j=1}^J ||\hat{\Pi}_j-\Pi_j||^2/J+o_p(1)=o_p(1)
\end{aligned}$$

by the essentially same proof in Newey, Powell, and Vella (1999) (p. 595). Also applying the similar argument to (41) we further find

$$\begin{aligned}
&\sum_{j=1}^J ||\hat{E}[\frac{\partial\delta_j^*(\theta_\lambda,s^n,P^R)}{\partial\theta_\lambda}|z_j,\hat{\mathbf{v}}_j]-\hat{E}[\frac{\partial\delta_j^*(\theta_\lambda,s^0,P^0)}{\partial\theta_\lambda}|z_j,\hat{\mathbf{v}}_j]||^2/J \qquad (50)\\
&\leq O_p(1)\zeta_\varphi(L)^2\cdot\sum_{j=1}^J\{\frac{\partial\delta_j^*(\theta_\lambda,s^n,P^R)}{\partial\theta_\lambda}-\frac{\partial\delta_j^*(\theta_\lambda,s^0,P^0)}{\partial\theta_\lambda}\}^2/J=o_p(1)
\end{aligned}$$

under Assumption 5 and Assumption C1 (ix).

Combining (49) and (50) by triangle inequality we obtain (in a neighborhood of $\theta_{\lambda 0}$)

$$J^{-1}\sum_{j=1}^J ||\hat{\Psi}_{\theta_\lambda,j}-\Psi_{\theta_\lambda,j}||^2=o_p(1). \qquad (51)$$

Define $\hat{\mathcal{T}}=(\hat{\Psi}^{L,J})'\hat{\Psi}^{L,J}/J$, $\hat{\dot{\mathcal{T}}}=(\hat{\dot{\psi}}^{L,J})'\hat{\dot{\psi}}^{L,J}/J$, and $\dot{T}=(\psi^{L,J})'\psi^{L,J}/J$. Our goal is to show that $\hat{\mathcal{T}}$ is nonsingular w.p.a.1. First note that $\dot{T}$ is nonsingular w.p.a.1 by Assumption L1 (ii) (b) because $|f_L-f_0|_\iota\leq CL^{-\varrho}\rightarrow 0$ by Assumption L1 (iv) and $\zeta_0(L)^2L/J\rightarrow 0$ (i.e. $||\dot{T}-\mathcal{T}^J||\rightarrow 0$ w.p.a.1. where $\mathcal{T}^J=\sum_{j=1}^J E[\Psi_{0j}^L\Psi_{0j}^{L\prime}]/J$). Assumption L1 (ii) (b) can hold as follows. Recall that our identification condition requires that $1,x_j',p_j,x_j'f(z_j,\mathbf{v}_j)$, and $(\bar{y}-p_j)f(z_j,\mathbf{v}_j)$ for any $f(z_j,\mathbf{v}_j)$ such that $E[f(Z_j,\mathbf{V}_j)|z_j]=0$ have no additive functional relationship, similarly it requires that $1,x_j',p_j,x_j'f(z_j,\mathbf{v}_j)$, $(\bar{y}-p_j)f(z_j,\mathbf{v}_j)$, and $(\gamma_0'x_j+\gamma_{p0}(\bar{y}-p_j)+1)\tilde{\varphi}_j^L$ have no additive functional relationship for any $f(z_j,\mathbf{v}_j)$ such that $E[f(Z_j,\mathbf{V}_j)|z_j]=0$ because $E[\tilde{\varphi}_j^L|z_j]=0$ by construction of $\tilde{\varphi}_j^L$. Moreover $E[(\gamma_0'X_j+(\bar{y}-p_j)\gamma_{p0}+1)^2\tilde{\varphi}_j^L\tilde{\varphi}_j^{L\prime}]$ is nonsingular by Assumption L1 (ii) (a), $\text{var}(\gamma_0'X_j+(\bar{y}-p_j)\gamma_{p0}+1)>0$ for all j, and by the essentially same proof in Lemma A1 of Newey, Powell, and Vella (1999). The same conclusion holds even when instead we take $\dot{T}=\sum_{j=1}^J C(z_j,\mathbf{v}_j)\psi_j^L\psi_j^{L\prime}/J$ for some positive bounded function $C(z_j,\mathbf{v}_j)$ and this helps to derive the consistency of the heteroskedasticity robust variance estimator later.

Next note that

$$\begin{aligned}
\left\|\hat{\tilde{\varphi}}_j^L-\tilde{\varphi}_j^L\right\| &\leq \left\|\varphi^L(z_j,\hat{\mathbf{v}}_j)-\varphi^L(z_j,\mathbf{v}_j)\right\|+\left\|\hat{\bar{\varphi}}^L(z_j)-\bar{\varphi}^L(z_j)\right\| \qquad (52)\\
&\leq \left\|\varphi^L(z_j,\hat{\mathbf{v}}_j)-\varphi^L(z_j,\mathbf{v}_j)\right\|+\left\|\hat{\bar{\varphi}}^L(z_j)-\dot{\bar{\varphi}}^L(z_j)\right\|+\left\|\dot{\bar{\varphi}}^L(z_j)-\bar{\varphi}^L(z_j)\right\|.
\end{aligned}$$

We find $\left\|\varphi^L(z_j,\hat{\mathbf{v}}_j)-\varphi^L(z_j,\mathbf{v}_j)\right\| \leq C\zeta_1(L)||\hat{\Pi}_j-\Pi_j||$ applying a mean value expansion because $\varphi_l(z_j,\mathbf{v}_j)$ is Lipschitz in Π_j for all l (Assumption C1 (vi)). Combined with (46), it implies that

$$\sum_{j=1}^J \left\|\varphi^L(z_j,\hat{\mathbf{v}}_j)-\varphi^L(z_j,\mathbf{v}_j)\right\|^2/J = O_p(\zeta_1(L)^2\triangle_{J,1}^2). \tag{53}$$

Next let $\hat{\omega}_l=(\varphi_l(z_1,\hat{\mathbf{v}}_1)-\varphi_l(z_1,\mathbf{v}_1),\ldots,\varphi_l(z_J,\hat{\mathbf{v}}_J)-\varphi_l(z_J,\mathbf{v}_J))'$. Then we can write

$$\begin{aligned}
\sum_{j=1}^J \left\|\hat{\dot{\varphi}}_l(z_j)-\dot{\varphi}_l(z_j)\right\|^2/J &= \operatorname{tr}\left\{\sum_{j=1}^J \varphi^k(z_j)'(\mathbf{P}'\mathbf{P})^-\mathbf{P}'\hat{\omega}_l\hat{\omega}_l'\mathbf{P}(\mathbf{P}'\mathbf{P})^-\varphi^k(z_j)\right\}/J \qquad (54)\\
&= \operatorname{tr}\left\{(\mathbf{P}'\mathbf{P})^-\mathbf{P}'\hat{\omega}_l\hat{\omega}_l'\mathbf{P}(\mathbf{P}'\mathbf{P})^-\sum_{j=1}^J\varphi^k(z_j)\varphi^k(z_j)'\right\}/J\\
&= \operatorname{tr}\left\{(\mathbf{P}'\mathbf{P})^-\mathbf{P}'\hat{\omega}_l\hat{\omega}_l'\mathbf{P}\right\}/J\\
&\leq C\max_{j\leq J}||\hat{\Pi}_j-\Pi_j||^2\operatorname{tr}\left\{(\mathbf{P}'\mathbf{P})^-\mathbf{P}'\mathbf{P}\right\}/J \leq C\zeta_0(k)^2\triangle_{J,1}^2k/J
\end{aligned}$$

where the first inequality is obtained by (47) and applying a mean value expansion to $\varphi_l(z_j,\mathbf{v}_j)$ which is Lipschitz in Π_j for all l (Assumption C1 (vi)). From (46), (52), (53), and (54), we conclude

$$\sum_{j=1}^J ||\hat{\bar{\varphi}}^L(z_j)-\bar{\varphi}^L(z_j)||^2/J = O_p(L\zeta_0(k)^2\triangle_{J,1}^2k/J)+O_p(L\triangle_{J,2}^2)=o_p(1) \tag{55}$$

and

$$\sum_{j=1}^J \left\|\hat{\tilde{\varphi}}_j^L-\tilde{\varphi}_j^L\right\|^2/J = O_p(\zeta_1(L)^2\triangle_{J,1}^2)+O_p(L\zeta_0(k)^2\triangle_{J,1}^2k/J)+O_p(L\triangle_{J,2}^2)=o_p(1). \tag{56}$$

This also implies that by the triangle inequality and the Markov inequality,

$$\sum_{j=1}^J ||\hat{\tilde{\varphi}}_j^L||^2/J \leq 2\sum_{j=1}^J ||\hat{\tilde{\varphi}}_j^L-\tilde{\varphi}_j^L||^2/J+2\sum_{j=1}^J||\tilde{\varphi}_j^L||^2/J = o_p(1)+O_p(L). \tag{57}$$

Let

$$\triangle_J^{\varphi}=(\zeta_1(L)+L^{1/2}\zeta_0(k)\sqrt{k/J}+L^{1/2})\triangle_J.$$

It also follows that

$$\sum_{j=1}^J \left\|\hat{\dot{\psi}}_j^L-\psi_j^L\right\|^2/J \leq \sum_{j=1}^J(C_j\left\|a_L\right\|^2+C_{1j})\left\|\hat{\tilde{\varphi}}_j^L-\tilde{\varphi}_j^L\right\|^2/J = O_p(L(\triangle_J^{\varphi})^2)=o_p(1). \tag{58}$$

Then applying (58) and applying the triangle inequality and Cauchy-Schwarz inequality and by Assumption L1 (iii) , we obtain

$$\begin{aligned}
||\hat{\dot{T}}-\dot{T}|| &\leq \sum_{j=1}^J\left\|\hat{\dot{\psi}}_j^L-\psi_j^L\right\|^2/J+2\sum_{j=1}^J\left\|\psi_j^L\right\|\left\|\hat{\dot{\psi}}_j^L-\psi_j^L\right\|/J \qquad (59)\\
&\leq O_p(L(\triangle_J^{\varphi})^2)+2\left(\sum_{j=1}^J\left\|\psi_j^L\right\|^2/J\right)^{1/2}\left(\sum_{j=1}^J\left\|\hat{\dot{\psi}}_j^L-\psi_j^L\right\|^2/J\right)^{1/2}\\
&= O_p(L(\triangle_J^{\varphi})^2)+O_p(L^{1/2}L^{1/2}\triangle_J^{\varphi})=o_p(1).
\end{aligned}$$

Therefore we conclude $\hat{\hat{T}}$ is also nonsingular w.p.a.1.

Next let $\hat{\vartheta}=(\hat{\theta}_\lambda',\hat{c},\hat{\beta}',\hat{\alpha},\hat{\gamma}',\hat{\gamma}_p,\hat{a}_L')'$ and $\vartheta_0=(\theta_{\lambda 0}',c_0,\beta_0',\alpha_0,\gamma_0',\gamma_{p0},a_L')'$ where a_L satisfies Assumption L1 (iv). Because $\hat{\Psi}_j^L$ depends on the estimates and we have shown the consistency (implying $||\hat{\vartheta}-\vartheta_0||=o_p(1)$), we derive the convergence rate by letting $||\hat{\vartheta}-\vartheta_0||=O_p(J^{-\tau})$ and then obtain conditions that the convergence rate τ should satisfy later.

Note that we can write

$$\begin{aligned}
\sum_{j=1}^J ||\hat{\Psi}_j^L-\hat{\hat{\psi}}_j^L||^2/J \quad &\leq \quad \sum_{j=1}^J C_{1j}\left\|\hat{\vartheta}-\vartheta_0\right\|^2 ||\hat{\hat{\varphi}}_j^L||^2/J+\sum_{j=1}^J ||\hat{\Psi}_{\hat{\theta}_\lambda,j}-\Psi_{\theta_{\lambda0},j}||^2/J \quad (60)\\
&\leq 2\quad \sum_{j=1}^J C_{1j}\left\|\hat{\vartheta}-\vartheta_0\right\|^2\left(||\hat{\hat{\varphi}}_j^L-\tilde{\varphi}_j^L||^2+||\tilde{\varphi}_j^L||^2\right)/J+ \quad (61)\\
&\quad +2\sum_{j=1}^J ||\hat{\Psi}_{\hat{\theta}_\lambda,j}-\Psi_{\hat{\theta}_\lambda,j}||^2/J+2\sum_{j=1}^J ||\Psi_{\hat{\theta}_\lambda,j}-\Psi_{\theta_{\lambda0},j}||^2/J
\end{aligned}$$

where the second term in (61) is $o_p(1)$ by (51) and the last term is also $o_p(1)$ by the Markov inequality because $\frac{\partial\delta_j^*(\theta_\lambda,s,P)}{\partial\theta_\lambda}$ is continuous in (θ_λ,s,P) at $(\theta_{\lambda0},s^0,P^0)$, $||\frac{\partial\delta_j^*(\theta_\lambda,s,P)}{\partial\theta_\lambda}||$ is bounded in the neighborhood of $(\theta_{\lambda0},s^0,P^0)$ (by Assumption C1 (vii)), and $\hat{\theta}_\lambda\to\theta_{\lambda0}$. Then from (51), $||\hat{\vartheta}-\vartheta_0||=O_p(J^{-\tau})$, (56), and $\sum_{j=1}^J C_{1j}\left\|\tilde{\varphi}_j^L\right\|^2/J=O_p(L)$ by the Markov inequality, we conclude

$$\sum_{j=1}^J ||\hat{\Psi}_j^L-\hat{\hat{\psi}}_j^L||^2/J=O(\sum_{j=1}^J ||\hat{\Psi}_{\theta,j}^L-\hat{\hat{\psi}}_{\theta,j}^L||^2/J+\sum_{j=1}^J ||\hat{\Psi}_{\hat{\theta}_\lambda,j}-\Psi_{\theta_{\lambda0},j}||^2/J)=O_p(J^{-2\tau}L)+o_p(1). \quad (62)$$

Then (58) and (62) implies

$$\sum_{j=1}^J \left\|\hat{\Psi}_j^L\right\|^2/J=O_p(L)$$

because $\sum_{j=1}^J\left\|\hat{\Psi}_j^L\right\|^2/J\leq 3\sum_{j=1}^J\left\|\hat{\Psi}_j^L-\hat{\hat{\psi}}_j^L\right\|^2/J+3\sum_{j=1}^J\left\|\hat{\hat{\psi}}_j^L-\psi_j^L\right\|^2/J+3\sum_{j=1}^J\left\|\psi_j^L\right\|^2/J=O_p(L)$. Also from (58) and (62) we conclude

$$\begin{aligned}
||\hat{\mathcal{T}}-\hat{\hat{T}}|| \quad &\leq \quad \sum_{j=1}^J ||\hat{\Psi}_j^L-\hat{\hat{\psi}}_j^L||^2/J+2\sum_{j=1}^J\left(||\hat{\hat{\psi}}_j^L-\psi_j^L||+||\psi_j^L||\right)||\hat{\Psi}_j^L-\hat{\hat{\psi}}_j^L||/J \quad (63)\\
&\leq \quad \sum_{j=1}^J ||\hat{\Psi}_j^L-\hat{\hat{\psi}}_j^L||^2/J\\
&\quad +C\left(\sum_{j=1}^J\left(||\hat{\hat{\psi}}_j^L-\psi_j^L||^2+||\psi_j^L||^2\right)/J\right)^{1/2}\left(\sum_{j=1}^J ||\hat{\Psi}_j^L-\hat{\hat{\psi}}_j^L||^2/J\right)^{1/2}\\
&= \quad O_p(J^{-2\tau}L+L^{1/2}J^{-\tau}L^{1/2})=O_p(J^{-\tau}L).
\end{aligned}$$

Therefore under the rate condition $J^{-\tau}L\to 0$, by (59), (63), and $\dot{T}$ is nonsingular w.p.a.1, we conclude $\hat{\mathcal{T}}$ is nonsingular w.p.a.1. The same conclusion holds even when we instead take $\hat{\mathcal{T}}=\sum_{j=1}^J C(z_j,\mathbf{v}_j)\hat{\Psi}_j^L\hat{\Psi}_j^{L\prime}/J$, $\hat{\hat{T}}=\sum_{j=1}^J C(z_j,\mathbf{v}_j)\hat{\hat{\psi}}_j^L\hat{\hat{\psi}}_j^{L\prime}/J$, and $\dot{T}=\sum_{i=1}^n C(z_j,\mathbf{v}_j)\psi_j^L\psi_j^{L\prime}/J$ for some positive bounded function $C(z_j,\mathbf{v}_j)$ and this helps to derive the consistency of the heteroskedasticity robust variance estimator later.

Let $\tilde{\varsigma}_j=\delta_j^*(\theta_{\lambda0},s^n,P^R)-g_0(z_j,\mathbf{v}_j)$ and $\varsigma_j=\delta_j^*(\theta_{\lambda0},s^0,P^0)-g_0(z_j,\mathbf{v}_j)$ and let $\tilde{\varsigma}=(\tilde{\varsigma}_j,\ldots,\tilde{\varsigma}_j)'$

and $\varsigma = (\varsigma_j, \ldots, \varsigma_j)'$. Then by the intermediate value theorem we have

$$\tilde{\varsigma} - \varsigma = H_\delta^{-1}(\bar{\delta^*}, \theta_{\lambda 0}, P^R)\varepsilon^n - H_\delta^{-1}(\tilde{\delta^*}, \theta_{\lambda 0}, P^R)\varepsilon^R(\theta_\lambda) \tag{64}$$

for some intermediate $\bar{\delta^*}$ between $\delta_j^*(\theta_{\lambda 0}, s^n, P^R)$ and $\delta_j^*(\theta_{\lambda 0}, s^0, P^R)$ and for some intermediate $\tilde{\delta^*}$ between $\delta_j^*(\theta_{\lambda 0}, s^0, P^R)$ and $\delta_j^*(\theta_{\lambda 0}, s^0, P^0)$, respectively. Consider by the essentially same proof for (A.9) in Berry, Linton, and Pakes (2004) we have for any positive sequence $\epsilon_J \to 0$,

$$\begin{aligned} & \sup_{(\bar{\delta^*}, \bar{P}) \in \mathcal{N}_{\delta^* 0}(\theta_{\lambda 0}, \epsilon_J) \times \mathcal{N}_{P^0}(\epsilon_J)} ||\frac{1}{\sqrt{J}}(\hat{\Psi}^{L,J} - \psi^{L,J})'\{H_\delta^{-1}(\bar{\delta^*}, \theta_{\lambda 0}, \bar{P}) - H_\delta^{-1}(\delta^{*0}, \theta_{\lambda 0}, P^0)\}\varepsilon^n|| \\ = \;& o_p(J^{-\tau}L^{1/2} + L^{1/2}(\triangle_J^\varphi)) = o_p(1) \end{aligned}$$

by the stochastic equicontinuity Assumption N4. Similarly we have for any positive sequence $\epsilon_J \to 0$,

$$\begin{aligned} & \sup_{(\tilde{\delta^*}, \tilde{P}) \in \mathcal{N}_{\delta^* 0}(\theta_{\lambda 0}, \epsilon_J) \times \mathcal{N}_{P^0}(\epsilon_J)} ||\frac{1}{\sqrt{J}}(\hat{\Psi}^{L,J} - \psi^{L,J})'\{H_\delta^{-1}(\tilde{\delta^*}, \theta_{\lambda 0}, \tilde{P}) - H_\delta^{-1}(\delta^{*0}, \theta_{\lambda 0}, P^0)\}\varepsilon^R(\theta_{\lambda 0})|| \\ = \;& o_p(J^{-\tau}L^{1/2} + L^{1/2}(\triangle_J^\varphi)) = o_p(1) \end{aligned}$$

by the stochastic equicontinuity Assumption N4.

Let $(\mathbf{Z}, \mathbf{V}) = ((Z_1, \mathbf{V}_1), \ldots, (Z_J, \mathbf{V}_J))$. Then we have $E[\varsigma_j|\mathbf{Z}, \mathbf{V}] = 0$ and by the independence assumption of the observations given $(\mathbf{Z}, \mathbf{V})$, we have $E[\varsigma_j\varsigma_{j'}|\mathbf{Z}, \mathbf{V}] = 0$ for $j \neq j'$. We also have $E[\varsigma_j^2|\mathbf{Z}, \mathbf{V}] < \infty$. Then by (64), (58), (62), and the triangle inequality, under $J^{-\tau}L \to 0$ we obtain

$$\begin{aligned} & E\big[||(\hat{\Psi}^{L,J} - \psi^{L,J})'\tilde{\varsigma}/J||^2|\mathbf{Z}, \mathbf{V}\big] \\ \leq \;& C_1 J^{-2}\sum\nolimits_{j=1}^J E[\varsigma_j^2|\mathbf{Z}, \mathbf{V}]||\hat{\Psi}_j^L - \psi_j^L||^2 \\ & + C_2 J^{-2}\mathrm{tr}\{(\hat{\Psi}^{L,J} - \psi^{L,J})'E[(\tilde{\varsigma} - \varsigma)(\tilde{\varsigma} - \varsigma)'|\mathbf{Z}, \mathbf{V}](\hat{\Psi}^{L,J} - \psi^{L,J})\} \\ \leq \;& J^{-1}O_p(J^{-2\iota}L + L(\triangle_n^\varphi)^2) \\ & + C_2 n^{-1}J^{-2}\mathrm{tr}\{(\hat{\Psi}^{L,J} - \psi^{L,J})'H^{-1}(\bar{\delta^*}, \theta_{\lambda 0}, P^R)nE_*[\varepsilon^n\varepsilon^{n\prime}]H^{-1}(\bar{\delta^*}, \theta_{\lambda 0}, P^R)'(\hat{\Psi}^{L,J} - \psi^{L,J})\} \\ & + C_2 R^{-1}J^{-2}\mathrm{tr}\{(\hat{\Psi}^{L,J} - \psi^{L,J})'H^{-1}(\tilde{\delta^*}, \theta_{\lambda 0}, P^R)RE_*[\varepsilon^R\varepsilon^{R\prime}]H^{-1}(\tilde{\delta^*}, \theta_\lambda, P^R)'(\hat{\Psi}^{L,J} - \psi^{L,J})\} \\ \leq \;& J^{-1}O_p(J^{-2\iota}L + L(\triangle_n^\varphi)^2) + C_2\frac{1}{J}\left[\frac{1}{nJ}\mathrm{tr}\{(\hat{\Psi}^{L,J} - \psi^{L,J})'H_{\delta 0}^{-1}V_2 H_{\delta 0}^{-1\prime}(\hat{\Psi}^{L,J} - \psi^{L,J})\}\right] \\ & + C_2\frac{1}{J}\left[\frac{1}{RJ}\mathrm{tr}\{(\hat{\Psi}^{L,J} - \psi^{L,J})'H_{\delta 0}^{-1}V_3 H_{\delta 0}^{-1\prime}(\hat{\Psi}^{L,J} - \psi^{L,J})\}\right] \\ \leq \;& J^{-1}O_p(J^{-2\tau}L + L(\triangle_J^\varphi)^2) + O_p(J^{-1}) \end{aligned}$$

where the bounds for the last two terms in the last inequality are obtained by the essentially same proofs for (38) and (39) in Berry, Linton, and Pakes (2004) for the random coefficient logit models (also for the logit without random coefficients) assuming $\frac{J^2}{n}$ and $\frac{J^2}{R}$ are bounded. Then from the standard result (see Newey (1997) or Newey, Powell, and Vella (1999)) that the bound of a term in the conditional mean implies the bound of the term itself, we obtain $||(\hat{\Psi}^{L,J} - \psi^{L,J})'\tilde{\varsigma}/J||^2 = O_p(J^{-1})$. Also note that $E[\big\|(\psi^{L,J})'\varsigma/J\big\|^2] = CL/J$ by the essentially same proof of Lemma A1 in Newey, Powell, and Vella (1999)) and that $E[\big\|(\psi^{L,J})'(\tilde{\varsigma} - \varsigma)/J\big\|^2] = CL/J$ by the similar proof as above.

Therefore, by the triangle inequality

$$\begin{aligned}
||(\hat{\Psi}^{L,J})'\tilde{\varsigma}/J||^2 &\leq 2||(\hat{\Psi}^{L,J}-\psi^{L,J})'\tilde{\varsigma}/J||^2+2||(\psi^{L,J})'\tilde{\varsigma}/J||^2 \\
&\leq 2||(\hat{\Psi}^{L,J}-\psi^{L,J})'\tilde{\varsigma}/J||^2+4||(\psi^{L,J})'\varsigma/J||^2+\left\|(\psi^{L,J})'(\tilde{\varsigma}-\varsigma)/J\right\|^2 \\
&= O_p(J^{-1})+O_p(L/J)=O_p(L/J).
\end{aligned} \tag{65}$$

Define

$$\begin{aligned}
\hat{\hat{g}}_j &= \hat{c}+x_j'\hat{\beta}-\hat{\alpha}p_j+(1+x_j'\hat{\gamma}+\hat{\gamma}_p(\bar{y}-p_j))\hat{f}(z_j,\hat{\mathbf{v}}_j), \\
\hat{g}_{Lj} &= c_0+x_j'\beta_0-\alpha_0p_j+(1+x_j'\gamma_0+\hat{\gamma}_{p0}(\bar{y}-p_j))\hat{f}_L(z_j,\hat{\mathbf{v}}_j), \\
\tilde{g}_{Lj} &= c_0+x_j'\beta_0-\alpha_0p_j+(1+x_j'\gamma_0+\hat{\gamma}_{p0}(\bar{y}-p_j))f_L(z_j,\hat{\mathbf{v}}_j), \\
\tilde{g}_{0j} &= c_0+x_j'\beta_0-\alpha_0p_j+(1+x_j'\gamma_0+\hat{\gamma}_{p0}(\bar{y}-p_j))f_L(z_j,\mathbf{v}_j),
\end{aligned}$$

where $\hat{f}(z_j,\hat{\mathbf{v}}_j)=\hat{a}_L'\hat{\tilde{\varphi}}(z_j,\hat{\mathbf{v}}_j)$, $\hat{f}_L(z_j,\hat{\mathbf{v}}_j)=a_L'\hat{\tilde{\varphi}}(z_j,\hat{\mathbf{v}}_j)$, $f_L(z_j,\mathbf{v}_j)=a_L'\tilde{\varphi}(z_j,\mathbf{v}_j)$, $f_L(z_j,\hat{\mathbf{v}}_j)=a_L'(\varphi(z_j,\hat{\mathbf{v}}_j)-\bar{\varphi}^L(z_j))$, and let $\hat{\hat{g}},\hat{g}_L$, $\tilde{g}_L$, $\tilde{g}_0$, and g_0 stack the J observations of $\hat{\hat{g}}_j,\hat{g}_{Lj}$, $\tilde{g}_{Lj}$, $\tilde{g}_{0j}$, and g_{0j}, respectively.

Then from the first order condition of the sieve M-estimation,[25] we obtain

$$\begin{aligned}
o_p(1) &= \sum_{j=1}^J\hat{\Psi}_j^L(\hat{E}[\delta_j^*(\hat{\theta}_\lambda,s^n,P^R)|z_j,\hat{\mathbf{v}}_j]-\hat{\hat{g}}_j)/J \\
&= \hat{\Psi}^{L,J\prime}(\delta^*(\hat{\theta}_\lambda,s^n,P^R)-\hat{\hat{g}})/J+\sum_{j=1}^J\hat{\Psi}_j^L\{\hat{E}[\delta_j^*(\hat{\theta}_\lambda,s^n,P^R)|z_j,\hat{\mathbf{v}}_j]-\delta_j^*(\hat{\theta}_\lambda,s^n,P^R)\}/J \\
&= \hat{\Psi}^{L,J\prime}(\tilde{\varsigma}+(\delta^*(\hat{\theta}_\lambda,s^n,P^R)-\delta^*(\theta_{\lambda0},s^n,P^R))-(\hat{\hat{g}}-\hat{g}_L)-(\hat{g}_L-g_L)-(g_L-g_0))/J+o_p(1) \\
&= \hat{\Psi}^{L,J\prime}(\tilde{\varsigma}-\hat{\Psi}^{L,J}(\hat{\vartheta}-\vartheta_0)-(\tilde{\Psi}^{L,J}-\hat{\Psi}^{L,J})(\hat{\vartheta}-\vartheta_0)) \\
&\quad -\hat{\Psi}^{L,J\prime}((\hat{g}_L-\tilde{g}_L)-(\tilde{g}_L-\tilde{g}_0)-(\tilde{g}_0-g_0))/J+o_p(1).
\end{aligned} \tag{66}$$

Above in the third equality we note for each element $\hat{\Psi}_j^{L,l}$ of $\hat{\Psi}_j^L$, $l=1,2,\ldots,\dim(\hat{\Psi}_j^L)$

$$\begin{aligned}
&\sum_{j=1}^J\hat{\Psi}_j^{L,l}\{\hat{E}[\delta_j^*(\hat{\theta}_\lambda,s^n,P^R)|z_j,\hat{\mathbf{v}}_j]-\delta_j^*(\hat{\theta}_\lambda,s^n,P^R)\}/J \\
=\ &\sum_{j=1}^J\{\hat{E}[\hat{\Psi}_j^{L,l}|z_j,\hat{\mathbf{v}}_j]-\hat{\Psi}_j^{L,l}\}\delta_j^*(\hat{\theta}_\lambda,s^n,P^R)/J=o_p(1)
\end{aligned}$$

because each element in $\hat{E}[\hat{\Psi}_j^L|z_j,\hat{\mathbf{v}}_j]-\hat{\Psi}_j^L$ is either zero or arbitrarily close to zero (this is because $\hat{E}[\hat{\Psi}_j^L|z_j,\hat{\mathbf{v}}_j]$ is a projection of $\hat{\Psi}_j^L$ on the space in which $\hat{\Psi}_j^L$ lies) and $\delta_j^*(\hat{\theta}_\lambda,s^n,P^R)$ is uniformly bounded. In the last equality of (66) we applied a mean value expansion to $-(\delta^*(\hat{\theta}_\lambda,s^n,P^R)-\delta^*(\theta_{\lambda0},s^n,P^R))+(\hat{\hat{g}}-\hat{g}_L)$ such that $\tilde{\Psi}^{L,J}=(\tilde{\Psi}_1^L,\ldots,\tilde{\Psi}_J^L)'$ is defined as

$$\tilde{\Psi}_j^L=-(-\hat{\Psi}_{\tilde{\theta}_\lambda,j}',1,x_j',-p_j,x_j'\tilde{a}_L'\hat{\tilde{\varphi}}^L(z_j,\hat{\mathbf{v}}_j),(\bar{y}-p_j)\tilde{a}_L'\hat{\tilde{\varphi}}^L(z_j,\hat{\mathbf{v}}_j),(1+x_j'\tilde{\gamma}+(\bar{y}-p_j)\tilde{\gamma}_p)\hat{\tilde{\varphi}}^L(z_j,\hat{\mathbf{v}}_j)')'$$

[25]Take the minimization error (tolerance) of estimation arbitrary small to justify this asymptotic expansion.

and $(\tilde{\theta}_\lambda, \tilde{\theta}, \tilde{a}_L)$ lies between $(\theta_{\lambda 0}, \theta_0, a_L)$ and $(\hat{\theta}_\lambda, \hat{\theta}, \hat{a}_L)$.

Next note that (similarly to (60))

$$||\tilde{\Psi}^{L,J} - \hat{\Psi}^{L,J}||^2/J \le ||\hat{\vartheta} - \tilde{\vartheta}_0||^2 (\sum_{j=1}^{J} C_j ||\hat{\hat{\varphi}}_j^L||^2/J + \sum_{j=1}^{J} C|| \frac{\partial \hat{\Psi}_{\tilde{\theta}_\lambda, j}}{\partial \theta_\lambda}||^2/J) = O_p(LJ^{-2\tau}).$$

It follows that by $\hat{\Psi}^{L,J}(\hat{\Psi}^{L,J\prime}\hat{\Psi}^{L,J})^{-1}\hat{\Psi}^{L,J\prime}$ idempotent, the triangle inequality, and the Cauchy-Schwarz inequality

$$\begin{aligned} & ||\hat{\mathcal{T}}^{-1}\hat{\Psi}^{L,J\prime}(\tilde{\Psi}^{L,J} - \hat{\Psi}^{L,J})(\hat{\vartheta} - \vartheta_0)/J|| \\ = \; & \{(\hat{\vartheta} - \vartheta_0)'(\tilde{\Psi}^{L,J} - \hat{\Psi}^{L,J})'\hat{\Psi}^{L,J}(\hat{\Psi}^{L,J\prime}\hat{\Psi}^{L,J})^{-1}\hat{\Psi}^{L,J\prime}(\tilde{\Psi}^{L,J} - \hat{\Psi}^{L,J})(\hat{\vartheta} - \vartheta_0)/J\}^{1/2} \\ \le \; & O_p(1)||\hat{\vartheta} - \vartheta_0||(\sum\nolimits_{j=1}^{J} ||\tilde{\Psi}_j^L - \hat{\Psi}_j^L||^2/J)^{1/2} \\ = \; & O_p(J^{-\tau}L^{1/2}J^{-\tau}) = O_p(L^{1/2}J^{-2\tau}). \end{aligned} \tag{67}$$

Similarly by $\hat{\Psi}^{L,J}(\hat{\Psi}^{L,J\prime}\hat{\Psi}^{L,J})^{-1}\hat{\Psi}^{L,J\prime}$ idempotent and Assumption L1 (iv),

$$||\hat{\mathcal{T}}^{-1}\hat{\Psi}^{L,J\prime}(\tilde{g}_L - \tilde{g}_0)/J|| \;=\; O_p(1)\{(\tilde{g}_L - \tilde{g}_0)'(\tilde{g}_L - \tilde{g}_0)/J\}^{1/2} = O_p(L^{-\varrho}). \tag{68}$$

Next note that by $\hat{\Psi}^{L,J}(\hat{\Psi}^{L,J\prime}\hat{\Psi}^{L,J})^{-1}\hat{\Psi}^{L,J\prime}$ idempotent, the Cauchy-Schwarz inequality and (55),

$$\begin{aligned} & ||\hat{\mathcal{T}}^{-1}\hat{\Psi}^{L,J\prime}(\hat{g}_L - \tilde{g}_L)/J|| = O_p(1)\{(\hat{g}_L - \tilde{g}_L)'(\hat{g}_L - \tilde{g}_L)/J\}^{1/2} \\ \le \; & O_p(1)(\sum_{j=1}^{J} C_j ||\hat{f}(z_j, \hat{\mathbf{v}}_j) - f_L(z_j, \hat{\mathbf{v}}_j)||^2/J)^{1/2} \\ \le \; & O_p(1)(\sum_{j=1}^{J} ||a_L||^2 ||\hat{\bar{\varphi}}^L(z_j) - \bar{\varphi}^L(z_j)||^2/J)^{1/2} = O_p(L\zeta_0(k)\triangle_{J,1}\sqrt{k/J} + L\triangle_{J,2}). \end{aligned} \tag{69}$$

Next consider applying the Cauchy-Schwarz inequality and a mean value expansion we obtain

$$||\hat{\Psi}^{L,J\prime}(\delta^*(\hat{\theta}_\lambda, s^n, P^R) - \delta^*(\theta_{\lambda 0}, s^n, P^R))||/J \le ||\hat{\Psi}^{L,J}\frac{\partial \delta^*(\tilde{\theta}_\lambda, s^n, P^R)}{\partial \theta_\lambda}/J|| \cdot ||\hat{\theta}_\lambda - \theta_{\lambda 0}|| \le CLJ^{-\tau} \tag{70}$$

where $\tilde{\theta}_\lambda$ is an intermediate value between $\hat{\theta}_\lambda$ and $\theta_{\lambda 0}$.

Combining (65), (66), (67), (68), (69), (70), and by $\hat{\mathcal{T}}$ is nonsingular w.p.a.1, we obtain

$$\begin{aligned} ||\hat{\vartheta} - \vartheta_0|| \;\le\; & ||\hat{\mathcal{T}}^{-1}(\hat{\Psi}^{L,J})'\tilde{\varsigma}/J|| + ||\hat{\mathcal{T}}^{-1}(\hat{\Psi}^{L,J})'(\hat{\hat{g}}_L - \hat{g}_L)/J|| \\ & + ||\hat{\mathcal{T}}^{-1}(\hat{\Psi}^{L,J})'(\hat{g}_L - g_L)/J|| + ||\hat{\mathcal{T}}^{-1}(\hat{\Psi}^{L,J})'(g_L - g_0)/J|| + o_p(1) \\ = \; & O_p(1)\{\sqrt{L/J} + L^{1/2}J^{-2\iota} + L\zeta_0(k)\triangle_{J,1}\sqrt{k/J} + L\triangle_{J,2} + L^{-\varrho}\}. \end{aligned}$$

This implies $||\hat{\vartheta} - \vartheta_0|| = O_p(\sqrt{L/J} + L\zeta_0(k)\triangle_{J,1}\sqrt{k/J} + L\triangle_{J,2} + L^{-\varrho})$ and for (63) to be $o_p(1)$, the convergence rate should satisfy

$$L \cdot O_p(\sqrt{L/J} + L\zeta_0(k)\triangle_{J,1}\sqrt{k/J} + L\triangle_{J,2} + L^{-\varrho}) \to 0 \tag{71}$$

for consistency. Combining (59) and (71) (other order conditions are dominated by these two conditions), we obtain the rate condition for the consistency:

$$\begin{aligned}(L\zeta_1(L) + L^{3/2}\zeta_0(k)\sqrt{k/J} + L^{3/2})\triangle_J &\rightarrow 0 \\ \sqrt{L^3/J} + L^2\zeta_0(k)\triangle_{J,1}\sqrt{k/J} + L^2\triangle_{J,2} + L^{1-\varrho} &\rightarrow 0\end{aligned}$$

and we conclude

$$||\hat{\vartheta} - \vartheta_0|| = O_p(\triangle_{J,\vartheta}) \equiv O_p(\sqrt{L/J} + L\triangle_J + L^{-\varrho})$$

since $\zeta_0(k)\sqrt{k/J} = o(1)$. From (63), we also find that $\hat{\mathcal{T}}$ becomes nonsingular w.p.a.1 under $\triangle_T \equiv L\triangle_{J,\vartheta} \rightarrow 0$.

Applying the triangle inequality, by (55), the Markov inequality, Assumption L1 (iv), and $\sum_{j=1}^J(\varphi^L(z_j,\mathbf{v}_j) - \hat{\bar{\varphi}}^L(z_j))(\varphi^L(z_j,\mathbf{v}_j) - \hat{\bar{\varphi}}^L(z_j))'/J$ is nonsingular w.p.a.1 (by Assumption L1 (ii) and (55)), we find

$$\begin{aligned}
&\sum\nolimits_{j=1}^J\left(\hat{f}(z_i,\mathbf{v}_j) - f_0(z_j,\mathbf{v}_j)\right)^2/J \qquad (72)\\
\leq\;& 3\sum\nolimits_{j=1}^J\left(\hat{f}(z_j,\mathbf{v}_j) - f_{Lj}^*\right)^2/J + 3\sum\nolimits_{j=1}^J\left(f_{Lj}^* - f_{Lj}\right)^2/J + 3\sum\nolimits_{j=1}^J\left(f_{Lj} - f_0(z_j,\mathbf{v}_j)\right)^2/J\\
\leq\;& O_p(1)||\hat{a}_L - a_L||^2\\
&+C_1\sum\nolimits_{j=1}^J||a_L||^2|\hat{\bar{\varphi}}^L(z_j) - \bar{\varphi}^L(z_j)||^2/J + C_2\sup_{\mathcal{W}}||a_L'\tilde{\varphi}^L(z,\mathbf{v}) - f_0(z,\mathbf{v})||^2\\
\leq\;& O_p(\triangle_{J,\vartheta}^2) + LO_p(L\zeta_0(k)^2\triangle_{J,1}^2k/J + L\triangle_{J,2}^2) + O_p(L^{-2\varrho}) = O_p(\triangle_{J,\vartheta}^2)
\end{aligned}$$

where we let $f_L^*(z_j,\mathbf{v}_j) = a_L'(\varphi^L(z_j,\mathbf{v}_j) - \hat{\bar{\varphi}}^L(z_j))$, $f_{Lj}^* = f_L^*(z_j,\mathbf{v}_j)$, and $f_{Lj} = f_L(z_j,\mathbf{v}_j)$. This also implies that $||\hat{\hat{g}} - g_0||^2/J = O_p(\triangle_{J,\vartheta}^2)$ and

$$\max_{1\leq j\leq J}|\hat{\hat{\hat{g}}}_j - g_{0j}| = O_p(\triangle_g) \equiv O_p(\zeta_0(L)\triangle_{J,\vartheta}) \qquad (73)$$

by a similar proof to Theorem 1 of Newey (1997).

C.2 Proof of Theorem C1

Under Condition S and Assumptions 3-6, 9-10, 13-15 and Assumptions C1, all the conditions for the consistency are satisfied. We take the pseudo-metrics as the uniform norm $||\cdot||_s = ||\cdot||_\infty$ and $||\cdot||_{\mathcal{F}} = ||\cdot||_\infty$. We can therefore conclude that $(\hat{\theta},\hat{\theta}_\lambda)$ and $\hat{f}$ are consistent from the consistency theorem. Under Assumptions C1, all the assumptions in Assumption L1 are satisfied (we take $\varrho_1 = s_\Pi/\dim(z)$, $\varrho_2 = s_\varphi/\dim(z)$, and $\varrho_\delta = s_\delta/\dim(z,\mathbf{v})$). For the consistency, we require the following rate conditions be satisfied: (i) $\zeta_0(k)^2k/J \rightarrow 0$ (such that $\mathbf{P}'\mathbf{P}/J$ is nonsingular w.p.a.1), (ii) $\zeta_0(L)^2L/J \rightarrow 0$ (such that $\dot{\mathcal{T}}$ is nonsingular w.p.a.1) and (iii)

$$\begin{aligned}L\triangle_J^\varphi &= (L\zeta_1(L) + L^{3/2}\zeta_0(k)\sqrt{k/J} + L^{3/2})\triangle_J \rightarrow 0\\ L\triangle_{J,\vartheta} &= \sqrt{L^3/J} + L^2\zeta_0(k)\triangle_{J,1}\sqrt{k/J} + L^2\triangle_{J,2} + L^{1-\varrho} \rightarrow 0.\end{aligned}$$

The other rate conditions are dominated by these three. For the polynomial approximations, we have $\zeta_\iota(L) \leq CL^{1+2\iota}$ and $\zeta_0(k) \leq Ck$ and for the spline approximations, we have $\zeta_\iota(L) \leq CL^{0.5+\iota}$ and $\zeta_0(k) \leq Ck^{0.5}$. Therefore for the polynomial approximations, the rate condition (iii) becomes $\sqrt{L^3/J} + \triangle_J(L^4 + L^2k^{3/2}/\sqrt{J}) + L^{1-\varrho} \to 0$ and for the spline approximations, it becomes $\sqrt{L^3/J} + \triangle_J(L^{5/2} + L^2k/\sqrt{J}) + L^{1-\varrho} \to 0$. We can take $\varrho = s_f/\dim(z, \mathbf{v})$ because f_0 is assumed to be in the Hölder class and we can apply the approximation theorems (e.g., see Timan (1963), Schumaker (1981), Newey (1997), and Chen (2007)). Therefore, the conclusion of Theorem C1 follows from Lemma L1 applying the dominated convergence theorem by $\hat{\hat{g}}$ and g_0 are bounded.

D Asymptotic Normality (Proof of Theorem AN1 and AN2)

D.1 Rate conditions

Along the proof, we obtain a list of rate conditions from bounding terms. We collect them here. We take $\varrho = s_f/\dim(z, \mathbf{v})$, $\varrho_1 = s_\Pi/\dim(z)$, and $\varrho_2 = s_\varphi/\dim(z)$. Define

$$\begin{aligned}
\triangle_J^\varphi &= (\zeta_1(L) + L^{1/2}\zeta_0(k)\sqrt{k/J} + L^{1/2})\triangle_J,\ \triangle_{J,\vartheta} = \sqrt{L/J} + L\triangle_J + L^{-\varrho} \\
\triangle_{\mathcal{T}} &= L\triangle_{J,\vartheta},\ \triangle_{\mathcal{T}_1} = \zeta_0(k)\sqrt{k/J} \\
\triangle_H &= \zeta_0(L)k^{1/2}/\sqrt{J} + k^{1/2}L^{1/2}\triangle_J^\varphi + L^{-\varrho}\zeta_0(L)k^{1/2},\ \triangle_{d\Psi} = L^{1/2}\triangle_{J,\vartheta}^2 \\
\triangle_{d\bar{\varphi}} &= \zeta_0(L)L\triangle_{J,2},\ \triangle_g = \zeta_0(L)\triangle_{J,\vartheta},\ \triangle_\Sigma = \boldsymbol{\Delta}_{\mathcal{T}} + \zeta_0(L)^2L/J \\
\triangle_{\hat{H}} &= (\zeta_1(L)\triangle_{J,\vartheta} + \zeta_0(k)\triangle_{J,1})L^{1/2}\zeta_0(k),\ \triangle_\omega = J^{-1}(\zeta_0(L)^2L + \zeta_0(k)^2k + \zeta_0(k)^2kL^4).
\end{aligned}$$

Then for the $\sqrt{J}$-consistency and the consistency of the variance estimator we require $\sqrt{J}k^{1/2}L^{-\varrho} \to 0, \sqrt{J}k^{-\varrho_1} \to 0, \sqrt{J}k^{-\varrho_2} \to 0$ and

$$\begin{gathered}
\sqrt{J}\triangle_{d\Psi} \to 0,\ k^{1/2}(\triangle_{\mathcal{T}_1} + \triangle_H + \triangle_{\mathcal{T}}) \to 0,\ \triangle_\omega \to 0,\ L^{1/2}\triangle_{J,\vartheta} \to 0 \\
k^{1/2}(\triangle_{\mathcal{T}_1} + \triangle_H) + L^{1/2}\triangle_{\mathcal{T}} + \triangle_{d\bar{\varphi}} \to 0,\ \triangle_g \to 0,\ L^{1/2}\triangle_{J,\vartheta} \to 0,\ \triangle_\Sigma \to 0,\ \triangle_{\hat{H}} \to 0.
\end{gathered}$$

For the polynomial approximations, the rate conditions become (dropping the dominated terms)

$$\begin{aligned}
&(1) \quad \sqrt{J}\triangle_{d\Psi} = \sqrt{J}L^{1/2}\triangle_{J,\vartheta}^2 = L^{3/2}/\sqrt{J} + \sqrt{J}L^{5/2}\triangle_J^2 + \sqrt{J}L^{1/2-2\varrho} \\
&(2) \quad k^{1/2}(\triangle_{\mathcal{T}_1} + \triangle_H) + L^{1/2}\triangle_{\mathcal{T}} + \triangle_{d\bar{\varphi}} \\
&= \quad (k^2 + Lk + L^2)/\sqrt{J} + (kL^{1/2} + L^2)\triangle_J^\varphi + L^{-\varrho}(\zeta_0(L)k + L^{3/2}) + \zeta_0(L)L\triangle_J \\
&= \quad \frac{k^2}{\sqrt{J}} + (kL^{7/2} + L^5 + \frac{Lk^{5/2} + L^{5/2}k^{3/2}}{\sqrt{J}})\triangle_J + L^{-\varrho}(Lk + L^{3/2}) \to 0 \\
&(3) \quad \triangle_{\hat{H}} = (L^{7/2}k\triangle_{J,\vartheta} + L^{1/2}k^2\triangle_{J,1}) \to 0.
\end{aligned}$$

Assuming $L^{-\varrho}, k^{-\varrho_1}$, and $k^{-\varrho_2}$ are small enough, these are all satisfied when $\frac{L^5k^{1/2} + L^{9/2}k^{3/2}}{\sqrt{J}} \to 0$.

For the spline approximations, the rate conditions (dropping the dominated terms) become

(1) $\sqrt{J}\triangle_{d\Psi} = L^{3/2}/\sqrt{J} + \sqrt{J}L^{5/2}\triangle_J^2 + \sqrt{J}L^{1/2-2\varrho} \to 0$
(2) $k^{1/2}(\triangle_{\mathcal{T}_1} + \triangle_H) + L^{1/2}\triangle_{\mathcal{T}} + \triangle_{d\bar{\varphi}} = (kL^2 + L^{7/2} + \frac{Lk^2 + L^{5/2}k}{\sqrt{J}})\triangle_J + L^{-\varrho}(L^{1/2}k + L^{3/2}) \to 0$
(3) $\triangle_{\hat{H}} = (L^2k^{1/2}\triangle_{J,\vartheta} + L^{1/2}k\triangle_{J,1}) \to 0.$

Assume $L^{-\varrho}, k^{-\varrho_1}$, and $k^{-\varrho_2}$ small enough. Then these are all satisfied if $\frac{L^{7/2}k^{1/2}+L^3k+L^2k^{3/2}}{\sqrt{J}} \to 0$.

D.2 Asymptotic variance terms

Let $\varphi_j^k = \varphi^k(Z_j)$ and $\Psi_{0j}^L = (\Psi_{\theta_{\lambda 0},j}', -\Psi_{\theta 0,j}^{L\prime})'$ where $\Psi_{\theta_{\lambda 0},j} = E[\frac{\partial \delta_j^*(\theta_{\lambda 0}, s^0, P^0)}{\partial \theta_\lambda}|z_j, \mathbf{v}_j]$ and $\Psi_{\theta 0,j}^L = (1, x_j, -p_j, x_j f_0(z_j,\mathbf{v}_j), (\bar{y}-p_j)f_0(z_j,\mathbf{v}_j), \frac{\partial g_{0j}}{\partial f_0}\tilde{\varphi}^L(z_j,\mathbf{v}_j)')'$. Also let $\varsigma_{0j} = \varsigma_j(\theta_{\lambda 0})$. Then define the followings:

$$
\begin{aligned}
\Sigma^J &= \sum\nolimits_{j=1}^J E[\Psi_{0j}^L \Psi_{0j}^{L\prime} \mathrm{var}(\varsigma_{0j}|Z_j,\mathbf{V}_j)]/J, \mathcal{T}^J = \sum_{j=1}^J E[\Psi_{0j}^L \Psi_{0j}^{L\prime}]/J, \mathcal{T}_1^J = \sum_{j=1}^J E[\varphi_j^k \varphi_j^{k\prime}]/J, \qquad (74)\\
\Sigma_1^J &= \sum\nolimits_{j=1}^J E[V_j^2 \varphi_j^k \varphi_j^{k\prime}]/J, \Sigma_{2,l}^J = \sum\nolimits_{j=1}^J E[(\varphi_l(Z_j,\mathbf{V}_j) - \bar{\varphi}_l(Z_j))^2 \varphi_j^k \varphi_j^{k\prime}]/J,\\
H_{11}^J &= \sum\nolimits_{j=1}^J E[\frac{\partial g_{0j}}{\partial f_0}\frac{\partial f_{0j}}{\partial V_j}\Psi_{0j}^L \varphi_j^{k\prime}]/J, \bar{H}_{11}^J = \sum_{j=1}^J \frac{\partial g_{0j}}{\partial f_0}\frac{\partial f_{0j}}{\partial v_j}\hat{\Psi}_j^L \varphi_j^{k\prime}/J\\
H_{12}^J &= \sum\nolimits_{j=1}^J E[\frac{\partial g_{0j}}{\partial f_0}E[\frac{\partial f_{0j}}{\partial V_j}|Z_j]\Psi_{0j}^L \varphi_j^{k\prime}]/J, \bar{H}_{12}^J = \sum_{j=1}^J \frac{\partial g_{0j}}{\partial f_0}E[\frac{\partial f_{0j}}{\partial V_j}|Z_j]\hat{\Psi}_j^L \varphi_j^{k\prime}/J\\
H_{2,l}^J &= \sum\nolimits_{j=1}^J E[a_l \frac{\partial g_{0j}}{\partial f_0}\Psi_{0j}^L \varphi_j^{k\prime}]/J, \bar{H}_{2,l}^J = \sum_{j=1}^J a_l \frac{\partial g_{0j}}{\partial f_0}\hat{\Psi}_j^L \varphi_j^{k\prime}/J, H_1^J = H_{11}^J - H_{12}^J, \bar{H}_1^J = \bar{H}_{11}^J - \bar{H}_{12}^J\\
\bar{\Omega}^J &= A(\mathcal{T}^J)^{-1}[\Sigma^J + H_1^J(\mathcal{T}_1^J)^{-1}\Sigma_1^J(\mathcal{T}_1^J)^{-1}H_1^{J\prime} + \sum_{l=1}^L H_{2,l}^J(\mathcal{T}_1^J)^{-1}\Sigma_{2,l}^J(\mathcal{T}_1^J)^{-1}H_{2,l}^{J\prime}](\mathcal{T}^J)^{-1}A'\\
\bar{\Omega}_2^J &= \frac{1}{nJ}A(\mathcal{T}^J)^{-1}(\Psi_0^{L,J\prime}H_{\delta 0}^{-1}V_2 H_{\delta 0}^{-1\prime}\Psi_0^{L,J})(\mathcal{T}^J)^{-1}A'\\
\bar{\Omega}_3^J &= \frac{1}{RJ}A(\mathcal{T}^J)^{-1}(\Psi_0^{L,J\prime}H_{\delta 0}^{-1}V_3 H_{\delta 0}^{-1\prime}\Psi_0^{L,J})(\mathcal{T}^J)^{-1}A'.
\end{aligned}
$$

Here we note $A^J = (\sum_{j=1}^J E[r_{0j}r_{0j}']/J)^{-1}\sum_{j=1}^J E[r_{0j}\Psi_{0j}^{L\prime}]/J$ and $A^J = A$ where $A = \lim_{J\to\infty} A^J$, so we do not distinguish A and A^J to save notation. We also often write $C \to 0$ to denote $||C|| \to 0$ for a sequence of matrix C. Below we show that $\bar{\Omega}^J + \bar{\Omega}_2^J + \bar{\Omega}_3^J \to \Omega + \Omega_2 + \Omega_3$ as $J, k, L \to \infty$ and in Section D.4 we show $\hat{\Omega} - \bar{\Omega}^J \to_p 0$, $\hat{\Omega}_2 - \bar{\Omega}_2^J \to_p 0$, $\hat{\Omega}_3 - \bar{\Omega}_3^J \to_p 0$ and therefore $\hat{\Omega} + \hat{\Omega}_2 + \hat{\Omega}_3 \to_p \Omega + \Omega_2 + \Omega_3$. We let $\mathcal{T}_1^J = I$ without loss of generality for ease of notation. Then $\bar{\Omega}^J = A(\mathcal{T}^J)^{-1}\left[\Sigma^J + H_1^J\Sigma_1^J H_1^{J\prime} + \sum_{l=1}^L H_{2,l}^J \Sigma_{2,l}^J H_{2,l}^{J\prime}\right](\mathcal{T}^J)^{-1}A'$. Let Γ^J be a symmetric square root of $(\bar{\Omega}^J + \bar{\Omega}_2^J + \bar{\Omega}_3^J)^{-1}$. Because $\mathcal{T}^J$ is nonsingular for all J large enough and $\mathrm{var}(\varsigma_{0j}|Z_j,\mathbf{V}_j)$ is bounded away from zero for all j, $C\Sigma^J - I$ is positive semidefinite for some positive constant C for all J large enough. It follows that

$$
\begin{aligned}
||\Gamma^J A(\mathcal{T}^J)^{-1}|| &= \{\mathrm{tr}(\Gamma^J A(\mathcal{T}^J)^{-1}(\mathcal{T}^J)^{-1}A'\Gamma^{J\prime})\}^{1/2} \le C\{\mathrm{tr}(\Gamma^J A(\mathcal{T}^J)^{-1}\Sigma^J(\mathcal{T}^J)^{-1}A'\Gamma^{J\prime})\}^{1/2}\\
&\le \{\mathrm{tr}(C\Gamma^J(\bar{\Omega}^J + \bar{\Omega}_2^J + \bar{\Omega}_3^J)\Gamma^{J\prime})\}^{1/2} \le C
\end{aligned}
$$

and therefore $||\Gamma^J A(\mathcal{T}^J)^{-1}||$ is bounded. Now we show that $\bar{\Omega}^J \to \Omega$ as $J, k, L \to \infty$. Note $A = (\sum_{j=1}^J E[r_{0j}r_{0j}']/J)^{-1}\sum_{j=1}^J E[r_{0j}\Psi_{0j}^{L\prime}]/J$ and $\omega_j^{*J} = (\sum_{j=1}^J E[r_{0j}r_{0j}']/J)^{-1}r_{0j}$. Let $\omega_{Lj}^* =$

$A(\mathcal{T}^J)^{-1}\Psi_{0j}^L$. Then note

$$\sum_{j=1}^{J} E[||\omega_j^{*J} - \omega_{Lj}^*||^2]/J \to 0 \tag{75}$$

because (i) we can view $\tilde{r}_{0j} \equiv (\sum_{j=1}^J E[r_{0j}\Psi_{0j}^{L\prime}]/J)(\mathcal{T}^J)^{-1}\Psi_{0j}^L$ is a projection of r_{0j} on Ψ_{0j}^L and (ii) r_{0j} is smooth and the second moment of r_{0j} is bounded (Assumption N1 (i)). It follows that

$$\begin{aligned}
&\sum_{j=1}^{J}\{E[\omega_{Lj}^*\mathrm{var}(\varsigma_j|Z_j,\mathbf{V}_j)\omega_{Lj}^{*\prime}] - E[\omega_j^{*J}\mathrm{var}(\varsigma_{0j}|Z_j,\mathbf{V}_j)\omega_j^{*J\prime}]\}/J \\
= &\sum_{j=1}^{J}\{A(\mathcal{T}^J)^{-1}E[\Psi_{0j}^L\mathrm{var}(\varsigma_{0j}|Z_j,\mathbf{V}_j)\Psi_{0j}^{L\prime}](\mathcal{T}^J)^{-1}A' - E[\omega_j^{*J}\mathrm{var}(\varsigma_{0j}|Z_j,\mathbf{V}_j)\omega_j^{*J\prime}]\}/J \to 0.
\end{aligned}$$

This concludes that $A(\mathcal{T}^J)^{-1}\Sigma^J(\mathcal{T}^J)^{-1}A' - \sum_{j=1}^J E[\omega_j^{*J}\mathrm{var}(\varsigma_{0j}|Z_j,\mathbf{V}_j)\omega_j^{*J\prime}]/J \to 0$ as $J,k,L\to\infty$ where the limit of the latter is the first term in Ω.

Next let

$$b_{Lj'} = \frac{1}{J}\sum_{j=1}^{J} E\left[\omega_{Lj}^*\frac{\partial g_{0j}}{\partial f_0}\left(\frac{\partial f_{0j}}{\partial V_j} - E[\frac{\partial f_{0j}}{\partial V_j}|Z_j]\right)\varphi_j^{k\prime}\right]\varphi_{j'}^k$$

and $b_{j'} = \frac{1}{J}\sum_{j=1}^J E\left[\omega_j^*\frac{\partial g_{0j}}{\partial f_0}\left(\frac{\partial f_{0j}}{\partial V_j} - E[\frac{\partial f_{0j}}{\partial V_j}|Z_j]\right)\varphi_j^{k\prime}\right]\varphi_{j'}^k$. Note that because $(\mathcal{T}_1^J)^{-1} = I$, b_{Lj} and b_j are least squares mean projections respectively of $\omega_{Lj}^*\frac{\partial g_{0j}}{\partial f_0}\left(\frac{\partial f_{0j}}{\partial V_j} - E[\frac{\partial f_{0j}}{\partial V_j}|Z_j]\right)$ on φ_j^k and $\omega_j^{*J}\frac{\partial g_{0j}}{\partial f_0}\left(\frac{\partial f_{0j}}{\partial V_j} - E[\frac{\partial f_{0j}}{\partial V_j}|Z_j]\right)$ on φ_j^k. Then $\frac{1}{J}\sum_{j=1}^J E[||b_{Lj}-b_j||^2] \le \frac{1}{J}\sum_{j=1}^J CE[||\omega_{Lj}^* - \omega_j^{*J}||^2] \to 0$ because the mean square error of a least squares projection cannot be larger than the MSE of the variable being projected. Also note that $\frac{1}{J}\sum_{j=1}^J E[||\rho_v(Z_j) - b_j||^2] \to 0$ as $J,k\to\infty$ because b_j is a least squares projection of $\omega_j^{*J}\frac{\partial g_{0j}}{\partial f_0}\left(\frac{\partial f_{0j}}{\partial \mathbf{V}_j} - E\left[\frac{\partial f_{0j}}{\partial \mathbf{V}_j}|Z_j\right]\right)$ on φ_j^k and it converges to the conditional mean as $k\to\infty$. Finally note that

$$\begin{aligned}
&E[b_{Lj}\mathrm{var}(V_j|Z_j)b_{Lj}'] \\
= &A(\mathcal{T}^J)^{-1}\sum_{j=1}^{J}\frac{1}{J}E\left[\Psi_{0j}^L\frac{\partial g_{0j}}{\partial f_0}\left(\frac{\partial f_{0j}}{\partial V_j} - E\left[\frac{\partial f_{0j}}{\partial V_j}|Z_j\right]\right)\varphi_j^{k\prime}\right]E[\mathrm{var}(V_j|Z_j)\varphi_j^k\varphi_j^{k\prime}] \\
&\times\sum_{j=1}^{J}\frac{1}{J}E\left[\varphi_j^k\frac{\partial g_{0j}}{\partial f_0}\left(\frac{\partial f_{0j}}{\partial V_j} - E\left[\frac{\partial f_{0j}}{\partial V_j}|Z_j\right]\right)\Psi_{0j}^{L\prime}\right](\mathcal{T}^J)^{-1}A'
\end{aligned}$$

and therefore $\sum_{j=1}^J E[b_{Lj}\mathrm{var}(V_j|Z_j)b_{Lj}']/J - A(\mathcal{T}^J)^{-1}H_1^J\Sigma_1^J H_1^{J\prime}(\mathcal{T}^J)^{-1}A' \to 0$. This also conclude that

$$A(\mathcal{T}^J)^{-1}H_1^J\Sigma_1^J H_1^{J\prime}(\mathcal{T}^J)^{-1}A' - \sum_{j=1}^{J}E[\rho_v(Z_j)\mathrm{var}(p_j|Z_j)\rho_v(Z_j)']/J \to 0$$

as $J,k,L\to\infty$ where the limit of the latter is the second term in Ω. Similarly we can show that for all l as $J, k, L \to \infty$

$$A(\mathcal{T}^J)^{-1}H_{2,l}^J\Sigma_{2,l}^J H_{2,l}^{J\prime}(\mathcal{T}^J)^{-1}A' - \sum_{j=1}^J E[\rho_{\bar{\varphi}_l}(Z_j)\mathrm{var}(\varphi_l(Z_j,\mathbf{V}_j)|Z_j)\rho_{\bar{\varphi}_l}(Z_j)']/J \to 0$$

where the limit of the latter is the third term in Ω. Therefore we conclude $\bar{\Omega}^J \to \Omega$ as $J, k, L \to \infty$.

Next we show $\bar{\Omega}_2^J \to \Omega_2$ and $\bar{\Omega}_3^J \to \Omega_3$ as $J, k, L \to \infty$. Remember that $V_2 = nE[\varepsilon^n\varepsilon^{n\prime}]$ and $V_3 = RE[\varepsilon^R(\theta_{\lambda 0})\varepsilon^R(\theta_{\lambda 0})']$. Note that $\omega_j^{*J} = (\Xi^J)^{-1}r_{0j}$ and $\omega_{Lj}^* = A(\mathcal{T}^J)^{-1}\Psi_{0j}^L$. Let $\omega^{*JJ\prime} = (\omega_1^*, \ldots, \omega_J^*)$ and $\omega_L^{*J\prime} = (\omega_{L1}^*, \ldots, \omega_{LJ}^*)$. Then from (75)

$$\begin{aligned}
& \frac{1}{J}E[\omega_L^{*J}H_{\delta 0}^{-1}V_2H_{\delta 0}^{-1\prime}\omega_L^{*J\prime}] - \frac{1}{J}E[\omega^{*JJ}H_{\delta 0}^{-1}V_2H_{\delta 0}^{-1\prime}\omega^{*JJ\prime}] \qquad (76)\\
= \; & \frac{1}{J}A(\mathcal{T}^J)^{-1}E[\Psi_0^{L,J}H_{\delta 0}^{-1}V_2H_{\delta 0}^{-1\prime}\Psi_0^{L,J\prime}](\mathcal{T}^J)^{-1}A' - \frac{1}{J}E[\omega^{*JJ}H_{\delta 0}^{-1}V_2H_{\delta 0}^{-1\prime}\omega^{*JJ\prime}] \to 0
\end{aligned}$$

and we find by definition of ω^{*JJ} and Assumption N3,

$$\frac{1}{J}E[\omega^{*JJ}H_{\delta 0}^{-1}V_2H_{\delta 0}^{-1\prime}\omega^{*JJ\prime}] - (\Xi^J)^{-1}\Phi_2(\Xi^J)^{-1\prime} \to 0. \qquad (77)$$

This concludes $\bar{\Omega}_2^J \to \Omega_2$ as $J, k, L \to \infty$. By similar argument we conclude $\bar{\Omega}_3^J \to \Omega_3$ as $J, k, L \to \infty$. We therefore conclude $\bar{\Omega}^J + \bar{\Omega}_2^J + \bar{\Omega}_3^J \to \Omega + \Omega_2 + \Omega_3$ as $J, k, L \to \infty$. This also implies that $\Gamma^J \to (\Omega + \Omega_2 + \Omega_3)^{-1/2}$ and Γ^J is bounded for all J large enough.

D.3 Influence functions and asymptotic normality

Next we derive the asymptotic normality of $\sqrt{J}((\hat{\theta}_\lambda', \hat{\theta}') - (\theta_{\lambda 0}', \theta_0'))'$. After we establish the asymptotic normality, we will show the convergence of the each term in (26) and (32) to the corresponding terms in (74). We show some of them first, which will be useful to derive the asymptotic normality. From the proofs in the convergence rate section, we obtain $||\hat{\mathcal{T}} - \mathcal{T}^J|| = O_p(\triangle_{\mathcal{T}}) = o_p(1)$ (see (63)-(71)) and obtain $||\hat{\mathcal{T}}_1 - \mathcal{T}_1^J|| = O_p(\triangle_{\mathcal{T}_1}) = o_p(1)$. We also have $||\Gamma^J A(\hat{\mathcal{T}}^{-1} - (\mathcal{T}^J)^{-1})|| = o_p(1)$ and $||\Gamma^J A\hat{\mathcal{T}}^{-1/2}||^2 = O_p(1)$ (see proof in Lemma A1 of Newey, Powell, and Vella (1999)). We next show $||\bar{H}_{11}^J - H_{11}^J|| = o_p(1)$. Let $H_{11L}^J = \sum_{j=1}^J E[\frac{\partial g_{0j}}{\partial f_0}\sum_{l=1}^L a_l \frac{\partial\varphi_l(Z_j,\mathbf{V}_j)}{\partial V_j}\Psi_{0j}^L\varphi_j^{k\prime}]/J$ and $\bar{H}_{11L}^J = \sum_{j=1}^J \frac{\partial g_{0j}}{\partial f_0}\sum_{l=1}^L a_l \frac{\partial\varphi_l(Z_j,\mathbf{V}_j)}{\partial V_j}\Psi_{0j}^L\varphi_j^{k\prime}/J$. Similarly define H_{12L}^J and $\bar{H}_{12L}^J$ and let $H_{1L}^J = H_{11L}^J - H_{12L}^J$. By Assumption N1 (ii), Assumption L1 (iii), and the Cauchy-Schwarz inequality,

$$\begin{aligned}
& ||H_1^J - H_{1L}^J||^2 \\
\leq \; & C\frac{1}{J}\sum_{j=1}^J\{E[||\frac{\partial g_{0j}}{\partial f_0}\{(\frac{\partial f_{0j}}{\partial V_j} - E[\frac{\partial f_{0j}}{\partial V_j}|Z_j]) - \sum_l a_l(\frac{\partial\varphi_l(Z_j,\mathbf{V}_j)}{\partial V_j} - E[\frac{\partial\varphi_l(Z_j,\mathbf{V}_j)}{\partial V_j}|Z_j])\}\Psi_{0j}^L\varphi_j^{k\prime}||^2]\} \\
\leq \; & CL^{-2\varrho}E[C_j||\Psi_{0j}^L||^2\sum_{i=1}^k\varphi_{ij}^2] = O(L^{-2\varrho}\zeta_0(L)^2k).
\end{aligned}$$

Next consider that by Assumption L1 (iii) and the Cauchy-Schwarz inequality,

$$E[\sqrt{J}||\bar{H}_{11L}^J - H_{11L}^J||] \leq C\left(\frac{1}{J}\sum_{j=1}^J E[(\frac{\partial g_{0j}}{\partial f_0}\sum_{l=1}^L a_l \frac{\partial \varphi_l(Z_j, \mathbf{V}_j)}{\partial V_j})^2 ||\Psi_{0j}^L||^2 \sum_{i=1}^k \varphi_{ij}^2]\right)^{1/2} \leq C\zeta_0(L)k^{1/2}$$

and that by a similar argument with (58) and (62) (applying a triangle inequality), the Cauchy-Schwarz inequality, and the Markov inequality,

$$\begin{aligned}\left\|\bar{H}_{11}^J - \bar{H}_{11L}^J\right\| &\leq CJ^{-1}\sum\nolimits_{j=1}^J |\frac{\partial g_{0j}}{\partial f_0}\sum_{l=1}^L a_l \frac{\partial \varphi_l(Z_j, \mathbf{V}_j)}{\partial V_j}| \cdot ||\hat{\Psi}_j^L - \Psi_{0j}^L|| \cdot ||\varphi_j^k|| \\ &\leq C\left(\sum\nolimits_{j=1}^J C_j ||\hat{\Psi}_j^L - \Psi_{0j}^L||^2/J\right)^{1/2} \cdot \left(\sum\nolimits_{j=1}^J ||\varphi_j^k||^2/J\right)^{1/2} \leq O_p(k^{1/2}L^{1/2}\triangle_J^{\varphi}).\end{aligned}$$

Therefore, we have $||\bar{H}_{11}^J - H_{11}^J|| = O_p(\zeta_0(L)k^{1/2}/\sqrt{J} + k^{1/2}L^{1/2}\triangle_J^{\varphi} + L^{-\alpha}\zeta_0(L)\sqrt{k}) = O_p(\triangle_H) = o_p(1)$. Similarly we can show that $||\bar{H}_{12}^J - H_{12}^J|| = o_p(1)$ and $||\bar{H}_{2,l}^J - H_{2,l}^J|| = o_p(1)$ for all l. Therefore we have $\bar{H}_1^J = H_1 + o_p(1)$ and $\bar{H}_{2,l}^J = H_{2,l} + o_p(1)$ for all l.

Now we derive the asymptotic expansion to obtain the influence functions. Recall definitions of $\hat{\hat{g}}_j$, $\hat{g}_{Lj}$, and g_{Lj} and further define

$$\begin{aligned}\hat{g}_{Lj} &= c_0 + x_j'\beta_{10} - \alpha_0 p_j + (1 + x_j'\gamma_0 + (\bar{y} - p_j)\gamma_{p0})\tilde{f}_L(z_j, \hat{\mathbf{v}}_j), \\ g_{Lj} &= c_0 + x_j'\beta_{10} - \alpha_0 p_j + (1 + x_j'\gamma_0 + (\bar{y} - p_j)\gamma_{p0})f_L(z_j, \mathbf{v}_j),\end{aligned}$$

where $\tilde{f}_L(z_j, \hat{\mathbf{v}}_j) = a_L'(\varphi^L(z_j, \hat{\mathbf{v}}_j) - E[\varphi^L(Z_j, \hat{\mathbf{V}}_j)|z_j])$ and again let $\hat{\hat{g}}, \hat{g}_L, \hat{g}_L$, and g_L stack the J observations of $\hat{\hat{g}}, \hat{g}_L, \hat{g}_L$, and g_L, respectively.

From the first order condition, we obtain the expansion[26] similarly to (66).

$$\begin{aligned}o_p(1) &= \hat{\Psi}^{L,J\prime}(\delta^*(\hat{\theta}_\lambda, s^n, P^R) - \hat{\hat{g}})/\sqrt{J} &(78)\\ &= \hat{\Psi}^{L,J\prime}\{\tilde{\varsigma} + (\delta^*(\hat{\theta}_\lambda, s^n, P^R) - \delta^*(\theta_{\lambda 0}, s^n, P^R)) - (\hat{\hat{g}} - \hat{g}_L)\}/\sqrt{J} \\ &\quad -\hat{\Psi}^{L,J\prime}\{(\hat{g}_L - \hat{g}_L) - (\hat{g}_L - g_L) - (g_L - g_0)\}/\sqrt{J} &(79)\\ &= \hat{\Psi}^{L,J\prime}\{\tilde{\varsigma} - \hat{\Psi}^{L,J}(\hat{\vartheta} - \vartheta_0) - (\tilde{\Psi}^{L,J} - \hat{\Psi}^{L,J})(\hat{\vartheta} - \vartheta_0)\}/\sqrt{J} \\ &\quad -\hat{\Psi}^{L,J\prime}\{(\hat{g}_L - \hat{g}_L) - (\hat{g}_L - g_L) - (g_L - g_0)\}/\sqrt{J}.\end{aligned}$$

First consider that similar to (67), by $\hat{\Psi}^{L,J}(\hat{\Psi}^{L,J\prime}\hat{\Psi}^{L,J})^{-1}\hat{\Psi}^{L,J\prime}$ idempotent, the triangle inequality, the Markov inequality, Cauchy-Schwarz inequality,

$$\begin{aligned}&||\hat{\mathcal{T}}^{-1}\hat{\Psi}^{L,J\prime}(\tilde{\Psi}^{L,J} - \hat{\Psi}^{L,J})(\hat{\vartheta} - \vartheta_0)/\sqrt{J}|| &(80)\\ &\leq O_p(1)(\quad \sqrt{J}L^{1/2}\triangle_{J,\vartheta}^2) = O_p(\sqrt{J}\triangle_{d\Psi}) = o_p(1).\end{aligned}$$

Next similar to (68) by $\hat{\Psi}^{L,J}(\hat{\Psi}^{L,J\prime}\hat{\Psi}^{L,J})^{-1}\hat{\Psi}^{L,J\prime}$ idempotent and Assumption L1 (iii),

$$||\hat{\mathcal{T}}^{-1}\hat{\Psi}^{L,J\prime}(g_L - g_0)/\sqrt{J}|| = O_p(\sqrt{J}L^{-\varrho}) \qquad (81)$$

[26]Take the minimization error (tolerance) in estimation arbitrary small to justify this asymptotic expansion.

From (78), (80), and (81), we have

$$\sqrt{J}\Gamma^J((\hat{\theta}_\lambda', \hat{\theta}')' - (\theta_{\lambda 0}', \theta_0')') = \sqrt{J}\Gamma^J A(\hat{\vartheta} - \vartheta_0) = \Gamma^J A\hat{\mathcal{T}}^{-1}\hat{\Psi}^{L,J\prime}(\tilde{\varsigma} - (\hat{\hat{g}}_L - \hat{g}_L) - (\hat{g}_L - g_L))/\sqrt{J} + o_p(1). \tag{82}$$

D.3.1 Influence function for the first stage

Now we derive the stochastic expansion of $\Gamma^J A\hat{\mathcal{T}}^{-1}\hat{\Psi}^{L,J\prime}(\hat{g}_L - g_L)/\sqrt{J}$. Note that by a second order mean-value expansion of each $\tilde{f}_{Lj}$ around v_j,

$$\begin{aligned}
&\Gamma^J A\hat{\mathcal{T}}^{-1}\sum\nolimits_{j=1}^{J}\hat{\Psi}_j^L(\hat{g}_{Lj} - g_{Lj})/\sqrt{J} = \Gamma^J A\hat{\mathcal{T}}^{-1}\sum\nolimits_{j=1}^{J}\frac{\partial g_{Lj}}{\partial f_L}\hat{\Psi}_j^L(\tilde{f}_{Lj} - f_{Lj})/\sqrt{J} \\
= \;&\Gamma^J A\hat{\mathcal{T}}^{-1}\sum\nolimits_{j=1}^{J}\frac{\partial g_{Lj}}{\partial f_L}\hat{\Psi}_j^L(\frac{df_{Lj}}{dv_j} - E[\frac{df_{Lj}}{dV_j}|Z_j])(\hat{\Pi}_j - \Pi_j)/\sqrt{J} + \hat{\kappa} \\
= \;&\Gamma^J A\hat{\mathcal{T}}^{-1}\bar{H}_1^J\hat{\mathcal{T}}_1^{-1}\sum\nolimits_{j=1}^{J}\varphi_j^k v_j/\sqrt{J} + \Gamma^J A\hat{\mathcal{T}}^{-1}\bar{H}_1^J\hat{\mathcal{T}}_1^{-1}\sum\nolimits_{j=1}^{J}\varphi_j^k(\Pi_j - \varphi_j^{k\prime}\lambda_k^1)/\sqrt{J} \qquad (83) \\
&+\Gamma^J A\hat{\mathcal{T}}^{-1}\sum\nolimits_{j=1}^{J}\frac{\partial g_{Lj}}{\partial f_L}\hat{\Psi}_j^L(\frac{df_{Lj}}{dv_j} - E[\frac{df_{Lj}}{dV_j}|Z_j])(\varphi_j^{k\prime}\lambda_k^1 - \Pi_j)/\sqrt{J} + \hat{\kappa}.
\end{aligned}$$

and the remainder term $||\hat{\kappa}|| \le C\sqrt{J}||\Gamma^J A\hat{\mathcal{T}}^{-1/2}||\zeta_0(L)\sum_{j=1}^{J}C_j||\hat{\Pi}_j - \Pi_j||^2/J = O_p(\sqrt{J}\zeta_0(L)\triangle_{J,1}^2) = o_p(1)$. Then by the essentially same proofs ((A.18) to (A.23)) in Lemma A2 of Newey, Powell, and Vella (1999), we can show the second term and the third term in (83) are $o_p(1)$ under $\sqrt{J}k^{-\varrho_1} \to 0$ (so that $\sqrt{J}\left|\Pi_0(z) - \lambda_k^{1\prime}\varphi^k(z)\right|_0 \to 0$ by Assumption L1 (iv)) and $k^{1/2}(\triangle_{\mathcal{T}_1} + \triangle_H) + L^{1/2}\triangle_{\mathcal{T}} \to 0$ (so that we can replace $\hat{\mathcal{T}}_1$ with $\mathcal{T}_1^J$, $\bar{H}_1^J$ with H_1^J, and $\hat{\mathcal{T}}$ with $\mathcal{T}^J$ respectively). Therefore we obtain

$$\Gamma^J A\hat{\mathcal{T}}^{-1}\hat{\Psi}^{L,J\prime}(\hat{g}_L - g_L)/\sqrt{J} = \Gamma^J A(\mathcal{T}^J)^{-1}H_1^J\sum\nolimits_{j=1}^{J}\varphi_j^k v_j/\sqrt{J} + o_p(1). \tag{84}$$

This derives the influence function that comes from estimating V_j in the first step.

D.3.2 Influence function for the second stage

Next we derive the stochastic expansion of $\Gamma^J A\hat{\mathcal{T}}^{-1}\hat{\Psi}^{L,J\prime}(\hat{\hat{g}}_L - \hat{g}_L)/\sqrt{J}$:

$$\begin{aligned}
&\Gamma^J A\hat{\mathcal{T}}^{-1}\sum_{j=1}^{J}\hat{\Psi}_j^L(\hat{\hat{g}}_{Lj} - \hat{g}_{Lj})/\sqrt{J} = \Gamma^J A\hat{\mathcal{T}}^{-1}\sum_{j=1}^{J}\frac{\partial\hat{g}_{Lj}}{\partial f_L}\hat{\Psi}_j^L a_L'(\hat{\bar{\varphi}}^L(z_j) - E[\varphi^L(Z_j, \hat{\mathbf{V}}_j)|z_j])/\sqrt{J} \\
= \;&\Gamma^J A\hat{\mathcal{T}}^{-1}\{\sum\nolimits_l \bar{H}_{2,l}^J\hat{\mathcal{T}}_1^{-1}\sum\nolimits_{j=1}^{J}\varphi_j^k\tilde{\varphi}_l(z_j, \mathbf{v}_j)/\sqrt{J} + \sum\nolimits_l \bar{H}_{2,l}^J\hat{\mathcal{T}}_1^{-1}\sum\nolimits_{j=1}^{J}\varphi_j^k(\bar{\varphi}_l(z_j) - \varphi_j^{k\prime}\lambda_{l,k}^2)/\sqrt{J}\} \\
&+\Gamma^J A\hat{\mathcal{T}}^{-1}\sum\nolimits_{j=1}^{J}\frac{\partial\hat{g}_{Lj}}{\partial f_L}\hat{\Psi}_j^L\sum\nolimits_l a_l(\varphi_j^{k\prime}\lambda_{l,k}^2 - \bar{\varphi}_l(z_j))/\sqrt{J} + \Gamma^J A\hat{\mathcal{T}}^{-1}\sum\nolimits_{j=1}^{J}\frac{\partial\hat{g}_{Lj}}{\partial f_L}\hat{\Psi}_j^L\rho_j/\sqrt{J} \qquad (85)
\end{aligned}$$

where $\rho_j \equiv \sum_l a_l\{\varphi_j^{k\prime}\hat{\mathcal{T}}_1^{-1}\sum_{j'=1}^{J}\varphi_{j'}^k(\varphi_l(z_{j'}, \hat{\mathbf{v}}_{j'}) - \varphi_l(z_{j'}, \mathbf{v}_{j'}))/J - (E[\varphi_l(z_j, \hat{\mathbf{v}}_j)|z_j] - \bar{\varphi}_l(z_j))\}$. We focus on the last term in (85). Note that $\varphi_j^{k\prime}\hat{\mathcal{T}}_1^{-1}\sum_{j=1}^{J}\varphi_j^k(\varphi_l(z_j, \hat{\mathbf{v}}_j) - \varphi_l(z_j, \mathbf{v}_j))/J$ is a least squares projection of $\varphi_l(z_j, \hat{\mathbf{v}}_j) - \varphi_l(z_j, \mathbf{v}_j)$ on φ_j^k and it converges to the conditional mean $E[\varphi_l(Z_j, \hat{\mathbf{V}}_j)|z_j] - \bar{\varphi}_l(z_j)$. Therefore $\rho_j = \sum_{l=1}^{L}a_l\rho_{jl}$ and ρ_{jl} is the projection residual from the least squares projection of $\varphi_l(z_j,\hat{\mathbf{v}}_j)-\varphi_l(z_j,\mathbf{v}_j)$ on φ_j^k for each l. It follows that $E[\rho_j|Z_1,\ldots,Z_J]=0$ and therefore

$$\sum_{j=1}^{J}E[||\rho_j||^2|Z_1,\ldots,Z_J]/J\leq\sum_{j=1}^{J}E[L\sum_{l}||\rho_{jl}||^2|Z_1,\ldots,Z_J]/J\leq L^2O_p(\triangle_{J,2}^2)$$

where the first inequality holds by the Cauchy-Schwarz inequality and the second inequality holds by the similar proof to (46). It follows that by Assumption L1 (iii) and the Cauchy-Schwarz inequality,[27]

$$E[\|\sum_{j=1}^{J}\frac{\partial\hat{g}_{Lj}}{\partial f_L}\hat{\Psi}_j^L\rho_j/\sqrt{J}\||Z_1,\ldots,Z_J]\leq(\frac{1}{J}\sum_{j=1}^{J}E[C_j||\hat{\Psi}_j^L||^2||\rho_j||^2|Z_1,\ldots,Z_J])^{1/2}\leq C\zeta_0(L)L\triangle_{J,2}.$$

This implies that $\sum_{j=1}^{J}\frac{\partial\hat{g}_L}{\partial f_L}\hat{\Psi}_j^L\rho_j/\sqrt{J}=O_p(\zeta_0(L)L\triangle_{J,2})=O_p(\triangle_{d\bar{\varphi}})=o_p(1)$.

Then again by the essentially same proofs ((A.18) to (A.23)) in Lemma A2 of Newey, Powell, and Vella (1999), we can show the second term and the third term in (85) are $o_p(1)$ under $\sqrt{J}k^{-\varrho_1}\to 0$ (so that $\sqrt{J}|\bar{\varphi}_l(z)-\lambda_{l,k}^{2\prime}\varphi^k(z)|_0\to 0$ for all l by Assumption L1 (iv)), $\sqrt{J}k^{1/2}L^{-\varrho}\to 0$ (so that $\sqrt{J}k^{1/2}|f_0(z,\mathbf{v})-a_L'\tilde{\varphi}^L(z,\mathbf{v})|_0\to 0$ by Assumption L1 (iv)), and $k^{1/2}(\triangle_{\mathcal{T}_1}+\triangle_H)+L^{1/2}\triangle_{\mathcal{T}}+\triangle_{d\varphi}\to 0$ (so that we can replace $\hat{\mathcal{T}}_1$ with $\mathcal{T}_1^J$, $\bar{H}_{2,l}^J$ with $H_{2,l}^J$, and $\hat{\mathcal{T}}$ with $\mathcal{T}^J$ respectively). Therefore we obtain

$$\Gamma^J A\hat{\mathcal{T}}^{-1}\hat{\Psi}^{L,J\prime}(\hat{g}_L-\hat{g}_L)/\sqrt{J}=\Gamma^J A(\mathcal{T}^J)^{-1}\sum\nolimits_l H_{2,l}^J\sum\nolimits_{j=1}^{J}\varphi_j^k\tilde{\varphi}_l(z_j,\mathbf{v}_j)/\sqrt{J}+o_p(1). \tag{86}$$

This derives the influence function that comes from estimating $E[\varphi_{lj}|z_j]$'s in the middle step.

D.3.3 Influence function due to the sampling and simulation errors

Next we analyze the influence function terms due to the sampling and the simulation errors, i.e. we derive the stochastic expansion of $\Gamma^J A\hat{\mathcal{T}}^{-1}\hat{\Psi}^{L,J\prime}(\tilde{\varsigma}-\varsigma)/\sqrt{J}$. Note that $\omega_j^{*J}=(\Xi^J)^{-1}r_{0j}$, $\omega_{Lj}^{*J}=A(\mathcal{T}^J)^{-1}\Psi_{0j}^L$, and (75) and note that replacing $\hat{\Psi}_j^L$ with Ψ_{0j}^L does not influence the stochastic expansion by (58) and (62) and $|f_L-f_0|_\iota=O(L^{-\varrho})$ by Assumption N1 (ii). We therefore have

$$\Gamma^J A\hat{\mathcal{T}}^{-1}\hat{\Psi}^{L,J\prime}(\tilde{\varsigma}-\varsigma)/\sqrt{J}=\Gamma^J(\Xi^J)^{-1}\mathbf{r}_0(\mathbf{z},\mathbf{v})'(\tilde{\varsigma}-\varsigma)/\sqrt{J}+o_p(1). \tag{87}$$

Further note that by the intermediate value expansion

$$\tilde{\varsigma}-\varsigma=\delta^*(\theta_{\lambda 0},s^n,P^R)-\delta^*(\theta_{\lambda 0},s^0,P^0)=\tilde{H}_\delta^{-1}\{\varepsilon^n-\varepsilon^R(\theta_{\lambda 0})\} \tag{88}$$

where $\tilde{H}_\delta=H_\delta(\tilde{\delta}^*,\theta_{\lambda 0},\tilde{P})$ for some intermediate $(\tilde{\delta}^*,\tilde{P})$. Combining (87) and (88) we can write

$$\begin{aligned}\Gamma^J A\hat{\mathcal{T}}^{-1}\hat{\Psi}^{L,J\prime}(\tilde{\varsigma}-\varsigma)/\sqrt{J} &= \Gamma^J(\Xi^J)^{-1}\mathbf{r}_0(\mathbf{z},\mathbf{v})'(\tilde{\varsigma}-\varsigma)/\sqrt{J}+o_p(1) \\ &= \Gamma^J(\Xi^J)^{-1}\upsilon_J(\tilde{\delta}^*,\theta_{\lambda 0},\tilde{P})+o_p(1) \\ &= \Gamma^J(\Xi^J)^{-1}\upsilon_J(\delta^*(\theta_{\lambda 0},s^0,P^0),\theta_{\lambda 0},P^0)+o_p(1)\end{aligned} \tag{89}$$

[27]Note that in our definition of $\hat{g}_{Lj}$, g_{Lj}, and g_{0j}, $\frac{\partial\hat{g}_{Lj}}{\partial f_L}=\frac{\partial g_{Lj}}{\partial f_L}=\frac{\partial g_{0i}}{\partial f_0}=(1+\gamma_0'x_j+(\bar{y}-p_j)\gamma_{p0})$.

where the third equality holds by Assumption N4 (Stochastic Equicontinuity).

Therefore by (82), (84), (86), and (89) we obtain the stochastic expansion,

$$\begin{aligned}\sqrt{J}\Gamma^J((\hat{\theta}_\lambda', \hat{\theta}')' - (\theta_{\lambda 0}', \theta_0')') &= \Gamma^J A(\mathcal{T}^J)^{-1}(\Psi_0^{L,J\prime}\varsigma - H_1^J \textstyle\sum_{j=1}^J \varphi_j^k \upsilon_j/\sqrt{J} - \sum_l H_{2,l}^J \sum_{j=1}^J \varphi_j^k \tilde{\varphi}_{lj}/\sqrt{J}) \\ &\quad + \Gamma^J(\Xi^J)^{-1}\upsilon_J(\delta^{*0}, \theta_{\lambda 0}, P^0) + o_p(1).\end{aligned}$$

To apply the Lindeberg-Feller theorem for the first three terms, we check the Lindeberg condition. For any vector q with $||q|| = 1$, let $W_{jJ} = q'\Gamma^J A(\mathcal{T}^J)^{-1}(\Psi_{0j}^L \varsigma_j - H_1^J \varphi_j^k \upsilon_j - \sum_l H_{2,l}^J \varphi_j^k \tilde{\varphi}_{lj})/\sqrt{J}$. Note that W_{jJ} is a triangular array r.v. and by construction, $E[W_{jJ}] = 0$ and $\mathrm{var}(W_{jJ}) = O(1/J)$. Also note that $||\Gamma^J A(\mathcal{T}^J)^{-1}|| \le C$, $||\Gamma^J A(\mathcal{T}^J)^{-1} H_l^J|| \le C||\Gamma^J A(\mathcal{T}^J)^{-1}|| \le C$ by $CI - H_l^J H_l^{J\prime}$ being positive semidefinite for $l = 1, (2,1), \ldots, (2,L)$. Also note that $(\sum_{l=1}^L \tilde{\varphi}_{lj})^4 \le L^2(\sum_{l=1}^L \tilde{\varphi}_{lj}^2)^2 \le L^3 \sum_{l=1}^L \tilde{\varphi}_{lj}^4$. It follows that for any $\epsilon > 0$,

$$\begin{aligned} & JE[1(|W_{jJ}| > \epsilon)W_{jJ}^2] = J\epsilon^2 E[1(|W_{jJ}| > \epsilon)(W_{jJ}/\epsilon)^2] \le J\epsilon^{-2}E[|W_{jJ}|^4] \\ \le \; & CJ\epsilon^{-2}\{E[||\Psi_{0j}^L||^4 E[\varsigma_j^4|Z_j, \mathbf{V}_j]] + E[||\varphi_j^k||^4 E[\mathbf{V}_j^4|Z_j]] + L^3 \sum_l E[||\varphi_j^k||^4 E[\tilde{\varphi}_{lj}^4|Z_j]]\}/J^2 \\ \le \; & CJ^{-1}(\zeta_0(L)^2 L + \xi(k)^2 k + \xi(k)^2 k L^4) = o(1). \end{aligned}$$

For the second term $\Gamma^J(\Xi^J)^{-1}\nu_J(\delta^{*0}, \theta_{\lambda 0}, P^0)$ we can apply the Lyapunov Central Limit Theorem for triangular arrays by Assumption N3 such that

$$\Gamma^J(\Xi^J)^{-1}\nu_J(\delta^{*0}, \theta_{\lambda 0}, P^0) \to_d \Gamma^J(\Xi^J)^{-1}(\Phi_2 + \Phi_3)(\Xi^J)^{-1\prime}\Gamma^{J\prime}.$$

Therefore, $\sqrt{J}\Gamma^J((\hat{\theta}_\lambda', \hat{\theta}')' - (\theta_{\lambda 0}', \theta_0')') \to_d N(0, I)$ by the Lindeberg-Feller and the Lyapunov Central Limit Theorem. We have shown that $\bar{\Omega}^J + \bar{\Omega}_2^J + \bar{\Omega}_3^J \to \Omega + \Omega_2 + \Omega_3$ and Γ^J is bounded for all J large enough. We therefore also conclude $\sqrt{J}((\hat{\theta}_\lambda', \hat{\theta}')' - (\theta_{\lambda 0}', \theta_0')') \to_d N(0, \Omega + \Omega_2 + \Omega_3)$. This proves the asymptotic normality results in Theorem AN1 and AN2.

D.4 Consistency of the estimate of the asymptotic variance

Now we show the convergence of the each term in (26) and (32) to the corresponding terms in (74). Let $\hat{\delta}_j^* = \delta_j^*(\hat{\theta}_\lambda, s^n, P^R)$, $\hat{\varsigma}_j = \hat{\delta}_j^* - \hat{g}(z_j, \hat{\mathbf{v}}_j)$, and $\varsigma_j = \delta_j^{*0} - g_0(z_j, \mathbf{v}_j)$. Note that

$$\begin{aligned}\hat{\varsigma}_j^2 - \varsigma_j^2 &= 2\varsigma_j\{(\hat{\delta}_j^* - \delta_j^{*0}) - (\hat{\hat{g}}_j - g_{0j})\} + \{(\hat{\delta}_j^* - \delta_j^{*0}) - (\hat{\hat{g}}_j - g_{0j})\}^2 \\ &\le 2\varsigma_j\{(\hat{\delta}_j^* - \delta_j^{*0}) - (\hat{\hat{g}}_j - g_{0j})\} + 2(\hat{\delta}_j^* - \delta_j^{*0})^2 + 2(\hat{\hat{g}}_j - g_{0j})^2\end{aligned}$$

and that $\max_{j \le J}|\hat{\hat{g}}_j - g_{0j}| = O_p(\triangle_g) = o_p(1)$ by (73). Let $(H_\delta^{-1})_j$ and $(\overline{H_\delta^{-1}})_j$ denotes the j-th row of H_δ^{-1} and $\overline{H_\delta^{-1}}$, respectively where $\overline{H_\delta^{-1}}$ is defined in (29).

Note $(\overline{H_\delta^{-1}})_j = \underline{s}_j^{-1}\mathbf{e}_j + \frac{\mathbf{i}}{\underline{s}_0} = J(J\underline{s}_j)^{-1}\mathbf{e}_j + \frac{J\mathbf{i}}{J\underline{s}_0}$ where $\mathbf{e}_j$ is the j-th row of the $J \times J$ identify matrix and note

$$\Pr[\sum_{j=1}^J (\varepsilon_j^n)^2 > \epsilon] \le J \max_{1 \le j \le J} \Pr[(s_j^n - s_j^0)^2 > \epsilon] \le J\exp(-\epsilon n) \tag{90}$$

where the last inequality is obtained by Bernstein's inequality since s^n is a sum of n independent random variables each bounded by one. By similar argument we obtain

$$\Pr[\sum_{j=1}^{J}(\varepsilon_j^R(\theta_{\lambda 0}))^2 > \epsilon] \leq J\exp(-\epsilon R). \tag{91}$$

It follows that under Condition S for all $j \leq J$,

$$\begin{aligned} |(H_\delta(\cdot,\tilde{s},\tilde{P})^{-1})_j(\varepsilon^n - \varepsilon^R(\theta_{\lambda 0}))| &\leq \max_{1\leq j'\leq J}(\overline{H_\delta^{-1}})_{jj'}\sum_{j=1}^{J}|\varepsilon_j^n - \varepsilon_j^R(\theta_{\lambda 0})| \\ &\leq O(J)\times\sqrt{J}(\sum_{j=1}^{J}(\varepsilon_j^n)^2 + \sum_{j=1}^{J}(\varepsilon_j^R(\theta_{\lambda 0}))^2)^{1/2} \end{aligned}$$

where the second inequality is obtained applying the Markov inequality. Therefore $|(H_\delta(\cdot,\tilde{s},\tilde{P})^{-1})_j(\varepsilon^n - \varepsilon^R(\theta_{\lambda 0}))| = o_p(1)$ as long as $\frac{\log(J^4)}{n}\to 0$ and $\frac{\log(J^4)}{R}\to 0$ by (90) and (91).

Then applying the intermediate value expansions we obtain for all j

$$\begin{aligned} |\hat{\delta}_j^* - \delta_j^{*0}| &\leq |\hat{\delta}_j^* - \delta_j^*(\theta_{\lambda 0}, s^n, P^R)| + |\delta_j^*(\theta_{\lambda 0}, s^n, P^R) - \delta_j^{*0}| \\ &\leq |\frac{\partial\delta_j^*(\tilde{\theta}_\lambda, s^n, P^R)}{\partial\theta_\lambda}|\cdot||\hat{\theta}_\lambda - \theta_{\lambda 0}|| + |(H_\delta(\cdot,\tilde{s},\tilde{P})^{-1})_j(\varepsilon^n - \varepsilon^R(\theta_{\lambda 0}))| \\ &\leq O_p(||\hat{\theta}_\lambda - \theta_{\lambda 0}||) + o_p(1) \\ &= O_p(\triangle_{J,\vartheta}) + o_p(1). \end{aligned} \tag{92}$$

Let $\hat{D} = \Gamma^J A\hat{\mathcal{T}}^{-1}\hat{\Psi}^{L,J\prime}\mathrm{diag}\{1+|\varsigma_1|,\ldots,1+|\varsigma_J|\}\hat{\Psi}^{L,J}\hat{\mathcal{T}}^{-1}A'\Gamma^{J\prime}$ and $\tilde{D} = \Gamma^J A\tilde{\mathcal{T}}^{-1}\tilde{\Psi}^{L,J\prime}\mathrm{diag}\{1+|\varsigma_1|,\ldots,1+|\varsigma_J|\}\tilde{\Psi}^{L,J}\tilde{\mathcal{T}}^{-1}A'\Gamma^{J\prime}$ where $\tilde{\Psi}^{L,J}$ and $\tilde{\mathcal{T}}$ is obtained by replacing $\hat{f}_j$ with $\hat{f}_L(z_j,\hat{\mathbf{v}}_j)$. Then by (60), (63), $||\hat{\vartheta} - \vartheta_0|| = O_p(\triangle_{J,\vartheta})$, and the triangle inequality, we have $E[||\hat{D} - \tilde{D}||] = o(1)$ under $L^{1/2}\triangle_{J,\vartheta}\to 0$ and then by the Markov inequality, $||\hat{D} - \tilde{D}|| = o_p(1)$. Note $\tilde{\Psi}^{L,J}$ and $\tilde{\mathcal{T}}$ only depend on $(z_1,v_1),\ldots,(z_J,v_J)$ and thus $E[\tilde{D}|(Z_1,V_1),\ldots,(Z_J,V_J)] \leq C\Gamma^J A\tilde{\mathcal{T}}^{-1}A'\Gamma^{J\prime} = O_p(1)$. Therefore, $||\hat{D}|| = O_p(1)$ as well. Next let $\tilde{\Sigma} = \sum_{j=1}^{J}\hat{\Psi}_j^L\hat{\Psi}_j^{L\prime}\varsigma_j^2/J$ and $\hat{e}_j = -2\varsigma_j\{(\hat{\delta}_j^* - \delta_j^{*0}) - (\hat{\hat{g}}_j - g_{0j})\} + 2(\hat{\hat{g}}_j - g_{0j})^2$. Then,

$$\begin{aligned} ||\Gamma^J A\hat{\mathcal{T}}^{-1}(\hat{\Sigma} - \tilde{\Sigma})\hat{\mathcal{T}}^{-1}A'\Gamma^{J\prime}|| &\leq ||\Gamma^J A\hat{\mathcal{T}}^{-1}\hat{\Psi}^{L,J\prime}\mathrm{diag}\{\hat{e}_1,\ldots,\hat{e}_J\}\hat{\Psi}^{L,J}\hat{\mathcal{T}}^{-1}A'\Gamma^{J\prime}|| \\ &\quad + o(\max_{j\leq J}|\hat{\hat{g}}_j - g_{0j}| + \max_{j\leq J}|\hat{\delta}_j^* - \delta_j^{*0}|) \\ &\leq C\mathrm{tr}(\hat{D})\{\max_{j\leq J}|\hat{\hat{g}}_j - g_{0j}| + \max_{j\leq J}|\hat{\delta}_j^* - \delta_j^{*0}|\}(1+o(1)) \\ &= O_p(\triangle_{J,\vartheta}) + o_p(1) = o_p(1) \end{aligned} \tag{93}$$

where the last equality holds because of (30) and (92).

Then, by the essentially same proof in Lemma A2 of Newey, Powell, and Vella (1999), we obtain

$$\begin{aligned} ||\tilde{\Sigma} - \Sigma^J|| &= O_p(\Delta_{\mathcal{T}} + \zeta_0(L)^2 L/J) = O_p(\triangle_\Sigma) = o_p(1), \quad (94) \\ ||\Gamma^J A \hat{\mathcal{T}}^{-1}(\tilde{\Sigma} - \Sigma^J)\hat{\mathcal{T}}^{-1} A' \Gamma^{J\prime}|| &= o_p(1) \\ ||\Gamma^J A(\hat{\mathcal{T}}^{-1}\Sigma^J \hat{\mathcal{T}}^{-1} - \mathcal{T}^{-1}\Sigma^J \mathcal{T}^{-1})A'\Gamma^{J\prime}|| &= o_p(1). \end{aligned}$$

Then, by (93), (94), and the triangle inequality, we conclude $||\Gamma^J A\hat{\mathcal{T}}^{-1}\hat{\Sigma}\hat{\mathcal{T}}^{-1}A'\Gamma^{J\prime} - \Gamma^J A\mathcal{T}^{-1}\Sigma^J\mathcal{T}^{-1}A'\Gamma^{J\prime}|| = o_p(1)$. It remains to show that for $l = 1, (1,2), \ldots, (2, L)$,

$$\Gamma^J A(\hat{\mathcal{T}}^{-1}\hat{H}_l \hat{\mathcal{T}}_1^{-1}\hat{\Sigma}_l \hat{\mathcal{T}}_1^{-1}\hat{H}_l' \hat{\mathcal{T}}^{-1} - (\mathcal{T}^J)^{-1} H_l^J \Sigma_l^J H_l^{J\prime}(\mathcal{T}^J)^{-1})A'\Gamma^{J\prime} = o_p(1). \quad (95)$$

As we have shown $||\hat{\Sigma} - \Sigma^J|| = o_p(1)$, similarly we can show $||\hat{\Sigma}_l - \Sigma_l^J|| = o_p(1)$, $l = 1, (1,2), \ldots, (2, L)$. We focus on showing $||\hat{H}_l^J - \bar{H}_l^J|| = o_p(1)$ for $l = 1, (1,2), \ldots, (2, L)$. First note that

$$\begin{aligned} ||\hat{H}_{11}^J - \bar{H}_{11}^J|| &= ||\sum\nolimits_{j=1}^J \frac{\partial \hat{\hat{\hat{g}}}_j}{\partial \hat{f}}(\sum\nolimits_{l=1}^L \hat{a}_l \frac{\partial \varphi_l(z_j, \hat{\mathbf{v}}_j)}{\partial v_j} - a_l \frac{\partial \varphi_l(z_j, \mathbf{v}_j)}{\partial v_j})\hat{\Psi}_j^L \varphi_j^{k\prime}/J|| \quad (96) \\ &\quad + ||\sum\nolimits_{j=1}^J (\frac{\partial \hat{\hat{\hat{g}}}_j}{\partial \hat{f}_L} - \frac{\partial g_{0j}}{\partial f_0})\sum\nolimits_{l=1}^L a_l \frac{\partial \varphi_l(z_j, \mathbf{v}_j)}{\partial v_j}\hat{\Psi}_j^L \varphi_j^{k\prime}/J||. \end{aligned}$$

Note $\sum_{j=1}^J ||\frac{\partial \hat{\hat{\hat{g}}}_j}{\partial \hat{f}} - \frac{\partial g_{0j}}{\partial f_0}||^2/J \le ||\hat{\theta} - \theta_0||^2 \sum_{j=1}^J C_j^2/J = O_p(\triangle_{J,\vartheta}^2)$. By the Cauchy-Schwarz inequality, (57), (60), and Assumption L1 (iii), we have

$$\sum_{j=1}^J ||C_j \hat{\Psi}_j^L \varphi_j^{k\prime}||^2/J \le C \sum_{j=1}^J ||\hat{\Psi}_j^L||^2 ||\varphi_j^k||^2/J = O_p(L\xi_0(k)^2) \quad (97)$$

for any bounded $C_j > 0$. Therefore we bound the second term in (96) as $O_p(L^{1/2}\xi_0(k)\triangle_{J,\vartheta})$.

Also note that by the triangle inequality, the Cauchy-Schwarz inequality, and by Assumption C1 (vi) and (47), applying a mean value expansion to $\frac{\partial \varphi_l(z_j, \mathbf{v}_j)}{\partial v_j}$ w.r.t v_j,

$$\begin{aligned} &\sum\nolimits_{j=1}^J ||\sum\nolimits_{l=1}^L (\hat{a}_l \frac{\partial \varphi_l(z_j, \hat{\mathbf{v}}_j)}{\partial v_j} - a_l \frac{\partial \varphi_l(z_j, \mathbf{v}_j)}{\partial v_j})||^2/J \quad (98) \\ \le\ & 2\sum\nolimits_{j=1}^J ||\sum\nolimits_{l=1}^L (\hat{a}_l - a_l)\frac{\partial \varphi_l(z_j, \mathbf{v}_j)}{\partial v_j}||^2/J + 2\sum\nolimits_{j=1}^J ||\sum\nolimits_{l=1}^L \hat{a}_l (\frac{\partial \varphi_l(z_j, \hat{\mathbf{v}}_j)}{\partial v_j} - \frac{\partial \varphi_l(z_j, \mathbf{v}_j)}{\partial v_j})||^2/J \\ \le\ & C||\hat{a} - a_L||^2 \sum\nolimits_{j=1}^J ||\frac{\partial \tilde{\varphi}^L(z_j, \mathbf{v}_j)}{\partial v_j}||^2/J + C_1 \sum\nolimits_{j=1}^J ||\sum\nolimits_{l=1}^L \hat{a}_l \frac{\partial^2 \varphi_l(z_j, \tilde{\mathbf{v}}_j)}{\partial v_j^2}(\hat{\Pi}_j - \Pi_j)||^2/J \\ \le\ & C||\hat{a} - a_L||^2 \sum\nolimits_{j=1}^J ||\frac{\partial \tilde{\varphi}^L(z_j, \mathbf{v}_j)}{\partial v_j}||^2/J + C_1 \max_{1 \le j \le J} ||\hat{\Pi}_j - \Pi_j||^2 \cdot \sum\nolimits_{j=1}^J ||\sum\nolimits_{l=1}^L \hat{a}_l \frac{\partial^2 \varphi_l(z_j, \tilde{\mathbf{v}}_j)}{\partial v_j^2}||^2/J \\ =\ & O_p(\zeta_1^2(L)\triangle_{J,\vartheta}^2 + \xi_0^2(k)\triangle_{J,1}^2) \end{aligned}$$

where $\tilde{\mathbf{v}}_j$ lies between $\hat{\mathbf{v}}_j$ and $\mathbf{v}_j$, which may depend on l. Therefore we bound the first term in (96) as $O_p((\zeta_1(L)\triangle_{J,\vartheta} + \zeta_0(k)\triangle_{J,1})L^{1/2}\zeta_0(k))$ by the Cauchy-Schwarz inequality combining (97) and (98). Then we conclude by the triangle inequality, $||\hat{H}_{11}^J - \bar{H}_{11}^J|| \le O_p((\zeta_1(L)\triangle_{J,\vartheta} + \zeta_0(k)\triangle_{J,1})L^{1/2}\zeta_0(k)) = O_p(\triangle_{\hat{H}}) = o_p(1)$. Similarly we can show that $||\hat{H}_{12}^J - \bar{H}_{12}^J|| = o_p(1)$ and $||\hat{H}_{2,l}^J - \bar{H}_{2,l}^J|| = o_p(1)$

$l = 1, (2,1), \ldots, (2,L)$.

Then we have $||\bar{H}_l^J - H_l^J|| = o_p(1)$ for $l = 1, (2,1), \ldots, (2,L)$. Therefore, $||\hat{H}_l^J - H_l^J|| = o_p(1)$ for $l = 1, (2,1), \ldots, (2,L)$. Then by the similar proof like (93) and (94), the conclusion (95) follows.

Next we show $\hat{\Omega}_2 = \frac{1}{nJ} A\hat{\mathcal{T}}^{-1}((\hat{\Psi}^{L,J})'\hat{H}_\delta^{-1}\hat{V}_2\hat{H}_\delta^{-1\prime}\hat{\Psi}^{L,J})\hat{\mathcal{T}}^{-1}A'$ is consistent and consistency of $\hat{\Omega}_3$ is similarly obtained. First note that replacing $\hat{\mathcal{T}}$ with $\mathcal{T}$ and $\hat{\Psi}^{L,J}$ with $\Psi_0^{L,J}$ does not affect the consistency, so we have only to show $\frac{1}{nJ}(\Xi^J)^{-1}\mathbf{r}_0(\mathbf{z},\mathbf{v})'\hat{H}_\delta^{-1}\hat{V}_2\hat{H}_\delta^{-1\prime}\mathbf{r}_0(\mathbf{z},\mathbf{v})(\Xi^J)^{-1\prime} \to \Omega_2$ where we replace $A(\mathcal{T}^J)^{-1}\Psi_0^{L,J\prime}$ with $(\Xi^J)^{-1}\mathbf{r}_0(\mathbf{z},\mathbf{v})'$ applying the same argument in (76). Next note obviously $\lim_{n\to\infty}\hat{V}_2 = V_2 = nE_*[\varepsilon^n\varepsilon^{n\prime}]$, so we have

$$
\begin{aligned}
& \frac{1}{nJ}(\Xi^J)^{-1}\mathbf{r}_0(\mathbf{z},\mathbf{v})'\hat{H}_\delta^{-1}\hat{V}_2\hat{H}_\delta^{-1\prime}\mathbf{r}_0(\mathbf{z},\mathbf{v})(\Xi^J)^{-1\prime} \\
= \; & \frac{1}{nJ}(\Xi^J)^{-1}\mathbf{r}_0(\mathbf{z},\mathbf{v})'\hat{H}_\delta^{-1}(n\varepsilon^n\varepsilon^{n\prime})\hat{H}_\delta^{-1\prime}\mathbf{r}_0(\mathbf{z},\mathbf{v})(\Xi^J)^{-1\prime} + o_p(1) \\
= \; & (\Xi^J)^{-1}\left\{\frac{1}{nJ}\mathbf{r}_0(\mathbf{z},\mathbf{v})'H_{\delta 0}^{-1}(n\varepsilon^n\varepsilon^{n\prime})H_{\delta 0}^{-1\prime}\mathbf{r}_0(\mathbf{z},\mathbf{v})\right\}(\Xi^J)^{-1\prime} + o_p(1) \\
= \; & (\Xi^J)^{-1}\Phi_2(\Xi^J)^{-1\prime} + o_p(1) = \Omega_2 + o_p(1)
\end{aligned}
$$

where the second equality holds by the stochastic equicontinuity (Assumption N4) and the third equality holds by Assumption N3 and (30).

Then from (95) finally note that because Γ^J is bounded for all J large enough $||(\hat{\Omega}+\hat{\Omega}_2+\hat{\Omega}_3) - (\Omega+\Omega_2+\Omega_3)|| \le C||\Gamma^J(\hat{\Omega}+\hat{\Omega}_2+\hat{\Omega}_3)\Gamma^{J\prime} - \Gamma^J(\Omega+\Omega_2+\Omega_3)\Gamma^{J\prime}|| = o_p(1)$.

References

Ackerberg, D., X. Chen, and J. Hahn (2009): "A Practical Asymptotic Variance Estimator for Two-Step Semiparametric Estimators," Working Paper.

Ai, C., and X. Chen (2003): "Efficient Estimation of Models With Conditional Moment Restrictions," *Econometrica*, 71(6), 1795–1843.

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. (2007): "Large Sample Sieve Estimation of Semi-Nonparametric Models," *Handbook of Econometrics, Volume 6B*.

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.

Chintagunta, P., J. Dube, and K. Goh (2005): "Beyond the Endogeneity Bias: The Effect of Unmeasured Brand Characteristics on Household Level Brand Choice Models," *Management Science*, 51(5), 832–849.

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.

Goolsbee, A., and A. Petrin (2004): "The Consumer Gains from Direct Broadcast Satellites and the Competition with Cable TV," *Econometrica*, 72(2), 351–382.

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

Hall, P., and J. Horowitz (2005): "Nonparametric Methods for Inference in the Presence of Instrumental Variables," *Annals of Statistics*, 33, 2904–2929.

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.

Murphy, K., and R. Topel (1985): "Estimation and Inference in Two-Step Econometric Models," *Journal of Business and Economic Statistics*, 3, 370–379.

Newey, W. (1984): "A Method of Moments Interpretation of Sequential Estimators," *Economics Letters*, 14, 201–206.

——— (1997): "Convergence Rates and Asymptotic Normality for Series Estimators," *Journal of Econometrics*, 79, 147–168.

Newey, W., and J. Powell (2003): "Instrumental Variable Estimation of Nonparametric Models," *Econometrica*, 71, 1565–1578.

Newey, W., J. Powell, and F. Vella (1999): "Nonparametric Estimation of Triangular Simultaneous Equaitons Models," *Econometrica*, 67, 565–603.

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.

Palais, R. (1959): "Natural Operations on Differential Forms," *Transactions of the American Mathematical Society*, 92, 125–141.

Schumaker, L. (1981): *Spline Fuctions:Basic Theory.*

Timan, A. (1963): *Theory of Approximation of Functions of a Real Variable.*