NBER WORKING PAPER SERIES

A NEW CONTROL FUNCTION APPROACH FOR NON-PARAMETRIC REGRESSIONS WITH ENDOGENOUS VARIABLES

Kyoo il Kim
Amil Petrin

Working Paper 16679
http://www.nber.org/papers/w16679

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
January 2011

This paper previously circulated as "Revisiting Instrumental Variables and the Classic Control Function Approach, with Implications for Parametric and Non-Parametric Regressions." The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

NBER working papers are circulated for discussion and comment purposes. They have not been peer-reviewed or been subject to the review by the NBER Board of Directors that accompanies official NBER publications.

A New Control Function Approach for Non-Parametric Regressions with Endogenous Variables
Kyoo il Kim and Amil Petrin
NBER Working Paper No. 16679
January 2011, Revised April 2013
JEL No. C14

ABSTRACT

When the endogenous variable enters the structural equation non-parametrically the linear Instrumental Variable (IV) estimator is no longer consistent. Non-parametric IV (NPIV) can be used but it requires one to impose restrictions during estimation to make the problem well-posed. The non-parametric control function estimator of Newey, Powell, and Vella (1999) (NPV-CF) is an alternative approach that uses the residuals from the conditional mean decomposition of the endogenous variable as controls in the structural equation. While computationally simple identification relies upon independence between the instruments and the expected value of the structural error conditional on the controls, which is hard to motivate in many economic settings including estimation of returns to education, production functions, and demand or supply elasticities. We develop an estimator for non-linear and non-parametric regressions that maintains the simplicity of the NPV-CF estimator but allows the conditional expectation of the structural error to depend on both the control variables and the instruments. Our approach combines the conditional moment restrictions (CMRs) from NPIV with the controls from NPV-CF setting. We show that the CMRs place shape restrictions on the conditional expectation of the error given instruments and controls that are sufficient for identification. When sieves are used to approximate both the structural function and the control function our estimator reduces to a series of Least Squares regressions. Our monte carlos are based on the economic settings suggested above and illustrate that our new estimator performs well when the NPV-CF estimator is biased. Our empirical example replicates NPV-CF and we reject the maintained assumption of the independence of the instruments and the expected value of the structural error conditional on the controls in their setting.

Kyoo il Kim
Department of Economics
University of Minnesota
4-129 Hanson Hall
1925 4th Street South
Minneapolis, MN 55455
kyookim@umn.edu

Amil Petrin
Department of Economics
University of Minnesota
4-101 Hanson Hall
Minneapolis, MN 55455
and NBER
petrin@umn.edu

1 Introduction

The problem of endogenous regressors in simultaneous equations models has a long history in econometrics and empirical studies. The econometric problem is further complicated when the endogenous variables enter non-parametrically in the structural equation because linear Instrumental Variable approaches are no longer valid.

There are currently two approaches to identification for this case. The non-parametric instrumental variables (NPIV) approach uses conditional moment restrictions (CMRs) to identify the structural equation (see Newey and Powell (2003), Ai and Chen (2003), Hall and Horowitz (2005), Blundell, Chen, and Kristensen (2007) and Horowitz (2011a)). The structural equation is the solution to the integral equation implied by the CMRs. The main challenge arises from the non-continuity of these estimators with respect to the joint distribution of the data - the ill-posed inverse problem - and there is a large literature on restrictions that make the problem well-posed (see e.g. Newey and Powell (2003), Florens (2003), and Darolles, Florens, and Renault (2006)).[1]

A second approach is the non-parametric control function estimator of Newey, Powell, and Vella (1999) (NPV-CF). The approach starts by using conditional mean decomposition to create control variables that are orthogonal to all of the instruments. For identification they consider the function that is the expected value of the structural error conditional on the instruments and the control variables (and thus conditional also on the endogenous variables because of the way the control variables are constructed). They then assume that this conditional expectation is independent of the instruments conditional on the control variables. This suffices for identification because the control function - the expected value of the structural error conditional on the instruments and controls - is then only a function of the control variables and because the control variables are measurably separated from the regressors as long as there are valid instruments. This implies that including (possibly functions of) the control variables conditions out the variations in the endogenous variables that are correlated with the structural error.

A weakness of this identification assumption is that it does not hold in some economic settings where endogeneity is a first-order concern. These include estimation of returns to education, production functions, and demand or supply with non-separable reduced forms for equilibrium prices. We provide some examples in Section 3 to illustrate why this assumption does not hold.

Our main contribution is to develop an estimator for non-linear and non-parametric regressions

[1] For example, assuming the structural function belongs to a class of compact functions and using a sieve approach is one way to make the problem well-posed. This compactness assumption, however, can be relaxed and instead various regularization methods have been used to stabilize the inversion of the integral equation (see Tikhonov, Goncharsky, Stepanov, and Yagola (1995), Engl, Hanke, and Neubauer (1996), Kress (1999), Chernozhukov and Hansen (2005), Hall and Horowitz (2005), Blundell, Chen, and Kristensen (2007), Chen and Pouzo (2009), Chen and Pouzo (2011), and Darolles, Fan, Florens, and Renault (2011)). Hall and Horowitz (2005) and Chen and Reiss (2011) studied convergence properties of the nonparametric IV estimators (also see Johannes, Bellegem, and Vanhems (2011) and Horowitz (2011b)). Horowitz (2007) developed conditions under which the estimator in Hall and Horowitz (2005) is asymptotically normal.

that is consistent in the case when the conditional expectation of the structural error depends on both the control variables and the instruments. We also show our estimator is consistent in settings where both NPIV and NPV-CF estimators are not, although this is not the main focus of the paper. For identification we combine the CMRs from NPIV and the conditional mean controls from NPV. Together they imply that the expectation of the structural error conditional on instruments and control variables can both depend upon instruments and be distinguished from any function of the endogenous and exogenous regressors, implying identification of the structural function in the outcome equation. Section 4 discusses identification and provides a set of assumptions that falls outside the realm of both NPIV and NPV-CF but not our new estimator.

Our estimator based on this identification result is a multi-step sieve estimator and we develop asymptotic properties of this estimator in Sections 5-7. The results include $\sqrt{n}$-asymptotic normality of linear functionals of the structural function and consistent estimators for their standard errors.

Our approach shares the strength of the NPV-CF estimator in that it is easy to implement. In the case of sieve estimation it typically reduces to a series of Least Squares regressions, which makes point estimation simple and approximation of standard errors possible by bootstrap methods.

Our estimator uses the same CMR conditions as the NPIV estimators and we therefore face potentially the same ill-posed inverse problem. Our estimator provides another set of restrictions that our assumptions together with the use of sieve approximations for both the structural function and the control function make the problem well-posed. Our proofs for identification, consistency, and asymptotic normality are all based on extending the proofs from Newey, Powell, and Vella (1999) - which are quite different from the methods of proof used in the NPIV literature - and we therefore focus our discussion in the paper almost exclusively on the control function aspects of our estimator.

Our monte carlos in Section 8 are motivated by the economic examples we provide in Section 3, where the NPV-CF identification assumption is violated. They illustrate the ease of implementing our estimator. They also show that our new estimator performs well while the NPV-CF can be biased in non-linear settings.

In Section 9 we return to the empirical application from NPV and show that we reject the NPV-CF independence assumption when the CMRs are maintained instead. However, in terms of the economic significance of the violation in this application the point estimates of the coefficients in the structural function do not change appreciably.

In Section 10 we conclude. The Appendix explores the implications of our estimator for control functions in linear models with additively separable errors.[2] Other technical details are also presented in the Appendix.

[2]For the classic control function approach linear in parameters see, for example, Telser (1964), Hausman (1978), or Heckman (1978).

2 The Model

We consider a triangular non-parametric simultaneous equations model with additivity:

$$x_i = \Pi_0(z_i) + v_i, \quad E[v_i|z_i] = 0 \tag{1}$$

$$y_i = f_0(x_i, z_{1i}) + \varepsilon_i. \tag{2}$$

With $z_i = (z_{1i}, z_{2i})$ denoting the instrumental variables, (1) is a conditional mean decomposition of x_i with $\Pi_0(z_i)$ denoting $E[x_i|z_i]$ (so $E[v_i|z_i] = 0$ is not restrictive). The unknown function $f_0(x_i, z_{1i})$ in the second equation is the parameter of interest with z_{1i} a subset of z_i. The endogeneity arises because $E[\varepsilon_i|x_i] \neq 0$, i.e. the regressors x_i are endogenous.

Control function estimators express y_i as a function of (z_i, v_i)

$$y_i = f_0(x_i, z_{1i}) + h_0(z_i, v_i) + \eta_i \tag{3}$$

where $h_0(z_i, v_i) = E[\varepsilon_i|z_i, v_i]$ is the control function and $\eta_i = \varepsilon_i - h_0(z_i, v_i)$ (so $E[\eta_i|z_i, v_i] = 0$). The control function conditions out the part of the error correlated with the endogenous regressors. However, without further restrictions on $h_0(z_i, v_i)$ the function $f_0(x_i, z_{1i})$ is not identified because one cannot separate the effect of (x_i, z_{1i}) on f_0 from their impact on h_0.

Newey, Powell, and Vella (1999) (NPV-CF) achieve identification by assuming that the expected value of ε_i conditional on v_i is independent of z_i,

$$E[\varepsilon_i|z_i, v_i] = E[\varepsilon_i|v_i], \tag{4}$$

so that $h_0(z_i, v_i)$ does not depend on z_i given the control v_i. From (3) it is clear that this shape restriction on $h_0(z, v)$ can potentially rule out any additive functional relationship with (x, z_1). NPV achieve identification of $f_0(x_i, z_{1i})$ by combining this restriction with (i) differentiability of $f_0(x, z_1)$, $h_0(v) = E[\varepsilon|v]$, and $\Pi_0(z)$, (ii) zero mass on the boundary of the support of (z, v), and (iii) full rank of $\frac{\partial \Pi_0(z)}{\partial z_2'}$ with probability one. Indeed identification holds as long as (x, z_1) and a control variate v are measurably separated and imposing NPV's sufficient conditions is one way to achieve this separability (see Florens, Heckman, Meghir, and Vytlacil (2008) and their use of Matzkin (2003)'s control).

In Section 3 we show it can be hard to motivate the independence assumption in (4) from primitives in models of demand and supply.[3] In contrast conditional moment restrictions are usually motivated by economic primitives that lead to exclusion restrictions. Our main contribution is to show that in a control function approach framework, we can identify and consistently estimate

[3] The NPV-CF estimator is also not robust to conditional heteroskedasticity. For example, suppose $\varepsilon_i = \sigma(z_i)\tilde{\varepsilon}_i$ where $E[\tilde{\varepsilon}_i|z_i] = 0$. Then $E[\varepsilon_i|z_i, v_i] \neq E[\varepsilon_i|v_i]$ because $E[\varepsilon_i|z_i, v_i] = \sigma(z_i)E[\tilde{\varepsilon}_i|z_i, v_i] = \sigma(z_i)E[\tilde{\varepsilon}_i|v_i]$, so $E[\varepsilon_i|z_i, v_i]$ cannot be written as a function of v_i only.

$f_0(x, z_1)$ by restricting $h_0(z, v)$ to satisfy the conditional moment restrictions (CMR)

$$(\textbf{CMR}) \qquad E[\varepsilon_i|z_i] = 0.$$

CMR implies that the function $h_0(z_i, v_i)$ must satisfy $E[h_0(z_i, v_i)|z_i] = 0$ because by the law of iterated expectations

$$0 = E[\varepsilon_i|z_i] = E[E[\varepsilon_i|z_i, v_i]|z_i] = E[h_0(z_i, v_i)|z_i] = 0. \tag{5}$$

Our main contribution is to show that the shape restriction on $h_0(z_i, v_i)$ implied by (5) implies that $h_0(z_i, v_i)$ can both depend upon z_i and be distinguished from any function of z_i only, which leads to identification of $f_0(x_i, z_{1i})$ when combined with a suitable rank condition. Specifically, in Sections 4-7 we show that **CMR** and a completeness condition in the control function setting is sufficient for identification and estimation of $f_0(x_i, z_{1i})$ or any linear functional of it. Before turning to the formal development of the estimator we illustrate with a simple example.

2.1 Heuristic Example

While our approach is consistent for much more general functions $h_0(z_i, v_i)$, for this example we suppose $h_0(z_i, v_i)$ is given by

$$h_0(z_i, v_i) = \varphi(z_i) + a_1 v_i + a_2 v_i^2 + a_3' z_i v_i$$

where $\varphi(z_i)$ denotes any arbitrary function of z_i. We show how we can identify $f_0(x_i, z_{1i})$ from an additive regression of y_i on (x_i, z_{1i}) and $h_0(z_i, v_i)$ when $h_0(z_i, v_i)$ satisfies the **CMR** condition. The **CMR** condition implies that $h_0(z_i, v_i)$ satisfies

$$\begin{aligned} 0 &= E[h_0(z_i, v_i)|z_i] = E[\varphi(z_i)|z_i] + a_1 E[v_i|z_i] + a_2 E[v_i^2|z_i] + a_3' E[z_i v_i|z_i] \\ &= \varphi(z_i) + a_2 E[v_i^2|z_i] \end{aligned}$$

because by construction of v_i we have $E[v_i|z_i] = 0$. It follows that

$$\begin{aligned} &h_0(z_i, v_i) = h_0(z_i, v_i) - E[h_0(z_i, v_i)|z_i] \\ &= (\varphi(z_i) - \varphi(z_i)) + a_1 v_i + a_2(v_i^2 - E[v_i^2|z_i]) + a_3' z_i v_i = a_1 v_i + a_2 \tilde{v}_{2i} + a_3' z_i v_i \end{aligned}$$

where $\tilde{v}_{2i} = v_i^2 - E[v_i^2|z_i]$. Identification in this example is then equivalent to the non-existence of a linear functional relationship between any functions of (x_i, z_{1i}) and linear functions of $v_i, \tilde{v}_{2i}$, and $z_i v_i$.

2.2 Computation and Testing

A strength of our non-parametric sieve estimator is that it will typically be a series of least squares regressions that proceeds in three steps. In the first step we obtain the control $\hat{v}_i =$

$x_i - \hat{E}[x_i|z_i]$ from the first stage (possibly non-parametric) regression of x_i on z_i (equation (1)). In the second step we construct an approximation of $h(z_i, \hat{v}_i)$ using (e.g.) polynomials. We directly impose the restriction $E[h(z_i, v_i)|z_i] = 0$ at this point by demeaning (conditional on z_i) each term in the approximation to $h(z_i, \hat{v}_i)$. For example, one non-parametric sieve approximation is given by

$$h(z_i, \hat{v}_i) \approx \sum\nolimits_{l_1=1}^{L_1} a_{l_1,0}(\hat{v}_i^{l_1} - E[\hat{v}_i^{l_1}|z_i]) + \sum\nolimits_{l=2}^{L} \sum_{l_1 \geq 1, l_2 \geq 1 \text{ s.t. } l_1+l_2=l} a_{l_1,l_2}\varphi_{l_2}(z_i)(\hat{v}_i^{l_1} - E[\hat{v}_i^{l_1}|z_i])$$

where $\varphi_{l_2}(z_i)$ denotes functions of z_i and $E[\hat{v}_i^{l_1}|z_i]$ are estimated using (possibly non-parametric) regression.[4] We then estimate the parameters of $f(x_i, z_{1i})$ and $h(z_i, \hat{v}_i)$ simultaneously in the final regression step. In the example above estimation would involve a first regression where an estimate of v_i is recovered followed by a second regression of v_i^2 on z_i to estimate $\tilde{v}_{2i}$ followed by the final nonparametric regression of y_i on (x_i, z_{1i}) with three additive regressors $v_i, \tilde{v}_{2i}$, and $z_i v_i$.

As in other control function approaches (e.g. Smith and Blundell (1986), Rivers and Vuong (1988), Wooldridge (2005), and Wooldridge and Papke (2008)), we can test for endogeneity of the regressors x_i. In our setting this is equivalent to testing whether the control function $h_0(z_i, v_i)$ is a zero function or not. If x_i is exogenous it must be that $h_0(z_i, v_i) = 0$ because $E[\varepsilon_i|z_i, v_i] = E[\varepsilon_i|z_i, x_i] = 0$. In terms of the simple example above, testing whether $a_1 = 0, a_2 = 0,$ and $a_3 = 0$ is equivalent to testing the null hypothesis of exogenous x_i.

3 Does $E[\varepsilon_i|z_i, v_i] = E[\varepsilon_i|v_i]$ Hold in Models of Demand or Supply?

In this section we examine the control function assumption in demand and supply settings. On the demand side we consider a buy/not-buy binary choice setting with logit demands and a single-product monopolist.[5] We let the latent utilities of consumers be $u_{i0} = \epsilon_{i0}$ and $u_{i1} = \beta_0 + \beta_1 z_1 - \alpha p + \xi + \epsilon_{i1}$ with $(\epsilon_{i0}, \epsilon_{i1})$ i.i.d. extreme value and (z_1, p, ξ) denoting observed characteristics, price, and the unobserved characteristic (to the econometrician). The market share for good 1 is given by

$$s = \frac{\exp(\beta_0 + \beta_1 z_1 - \alpha p + \xi)}{1 + \exp(\beta_0 + \beta_1 z_1 - \alpha p + \xi)}, \tag{6}$$

which can be linearized as

$$\ln s - \ln(1 - s) = \beta_0 + \beta_1 z_1 - \alpha p + \xi.$$

If we let $mc(z_2, \omega)$ denote marginal costs given as a function of a cost shifter z_2 and a cost shock ω,

[4] For consistency $L_1, L \to \infty$ and $L_1/n, L/n \to 0$ as $n \to \infty$.

[5] The idea extends immediately to more general multi-firm and multi-product settings.

then the monopolist chooses price p to maximize the expected profit such that

$$p = \arg\max_p \, (p - mc(z_2, \omega)) \frac{\exp(\beta_0 + \beta_1 z_1 - \alpha p + \xi)}{1 + \exp(\beta_0 + \beta_1 z_1 - \alpha p + \xi)},$$

so $E[\xi|p] \neq 0$ in the linearized demand equation. Prices are evidently not separable in ξ and $z = (z_1, z_2)$. The control is given as $v = p - E[p|z]$. Since price is not separable in z and ξ conditioning on v will not generally make ξ independent of z implying $E[\xi|z, v] \neq E[\xi|v]$.

For the returns to education/production function setting - both of which are about input choices conditional on productivity - we use the setup from Imbens and Newey (2009) and Florens, Heckman, Meghir, and Vytlacil (2008). y denotes the outcome (wages/output), x is the agent's choice variable (schooling/input), and

$$y = f(x) + \varepsilon.$$

$c(x, z, \eta)$ is the cost function where z denotes a cost shifter. The agent sees a noisy signal η of ε, with η possibly a vector. The agent optimally chooses x by maximizing the expected profit given the information (z, η) so the observed x is the solution to

$$x = \arg\max_{\tilde{x}} \{E[f(\tilde{x}) + \varepsilon|z, \eta] - c(\tilde{x}, z, \eta)\}, \tag{7}$$

which leads to the endogeneity problem.

Assuming differentiability the optimal x solves

$$\partial f(x)/\partial x - \partial c(x, z, \eta)/\partial x = 0. \tag{8}$$

By the implicit function theorem we have $x = k(z, \eta)$ for some function $k(\cdot)$ and we also know that

$$\frac{\partial x}{\partial \eta} = \frac{\partial^2 c(x, z, \eta)/\partial x \partial \eta}{\partial^2 f(x)/\partial x^2 - \partial^2 c(x, z, \eta)/\partial x^2}.$$

Since the derivative of x with respect to η depends on z, z and η are not additively separable in $x = k(z, \eta)$. With $v = x - E[x|z]$ this implies that ε conditional on v is not generally independent of z (so $E[\varepsilon|z, v] \neq E[\varepsilon|v]$).

We illustrate further by considering the special case when

$$y = \varphi_0 + \varphi_1 x + \frac{1}{2}\varphi_2 x^2 + \varepsilon$$

for parameters $(\varphi_0, \varphi_1, \varphi_2)$ and

$$c(x, z, \eta) = c_0(z, \eta_0) + c_1(z, \eta_1)x + \frac{1}{2}c_2(z, \eta_2)x^2,$$

for cost shocks $\eta = (\eta_0, \eta_1, \eta_2)$ known to the agent and possibly correlated with ε (and z independent

of ε and η). The optimal x is

$$x = \frac{\varphi_1 - c_1(z, \eta_1)}{c_2(z, \eta_2) - \varphi_2} \tag{9}$$

which is not additively separable in η_1 or η_2, which means $E[\varepsilon|z, v]$ is not generally equal to $E[\varepsilon|v]$.[6]

4 Identification

We ask whether $f_0(x_i, z_{1i})$ and $h_0(z_i, v_i)$ are identified by equation (3) with restrictions (5). Our approach to identification closely follows Newey, Powell, and Vella (1999) and Newey and Powell (2003). We consider pairs of functions $f(x_i, z_{1i})$ and $h(z_i, v_i)$ that satisfy

$$E[y_i|z_i, v_i] = f(x_i, z_{1i}) + h(z_i, v_i) \tag{10}$$

and the **CMR** condition. Because conditional expectations are unique with probability one, if there exists a pair $\bar{f}(x_i, z_{1i})$ and $\bar{h}(z_i, v_i)$ that satisfies (10), it must be that

$$\Pr(f_0(x_i, z_{1i}) + h_0(z_i, v_i) = \bar{f}(x_i, z_{1i}) + \bar{h}(z_i, v_i)) = 1. \tag{11}$$

Identification of f_0 and h_0 means we must have $f_0 = \bar{f}$ and $h_0 = \bar{h}$ with probability one whenever (11) holds. Working with differences, we let $\delta(x_i, z_{1i}) = f_0(x_i, z_{1i}) - \bar{f}(x_i, z_{1i})$ and $\kappa(z_i, v_i) = h_0(z_i, v_i) - \bar{h}(z_i, v_i)$. Identification of f_0 and h_0 is then equivalent to

$$\Pr(\delta(x_i, z_{1i}) + \kappa(z_i, v_i) = 0) = 1 \text{ implying } \Pr(\delta(x_i, z_{1i}) = 0, \kappa(z_i, v_i) = 0) = 1.$$

Theorem 1. *Assume (1-2) and (5). If for all $\delta(x_i, z_{1i})$ with finite expectation $E[\delta(x_i, z_{1i})|z_i] = 0$ implies $\delta(x_i, z_{1i}) = 0$ a.s. then $f_0(x_i, z_{1i})$ and $h_0(z_i, v_i)$ are identified.*

Proof. Suppose it is not identified. Then there must exist functions $\bar{f}(x_i, z_{1i})$ and $\bar{h}(z_i, v_i)$ such that $\delta(x_i, z_{1i}) \neq 0$ and $\kappa(z_i, v_i) \neq 0$ but $\Pr(\delta(x_i, z_{1i}) + \kappa(z_i, v_i) = 0) = 1$. It follows that $E[\kappa(z_i, v_i)|z_i] = 0$ by construction so $0 = E[\delta(x_i, z_{1i}) + \kappa(z_i, v_i)|z_i] = E[\delta(x_i, z_{1i})|z_i]$. $E[\delta(x_i, z_{1i})|z_i] = 0$ then implies $\delta(x_i, z_{1i}) = 0$ a.s., so $\delta(x_i, z_{1i}) = 0$ and $\kappa(z_i, v_i) = 0$ with probability one. This is a contradiction. The result then implies that $h_0(z_i, v_i)$ is also identified because the conditional expectation $E[y_i|z_i, v_i]$ is nonparametrically identified and $h_0(z_i, v_i) = E[y_i|z_i, v_i] - f_0(x_i, z_{1i})$. □

A sufficient condition for identification is that the conditional distribution of x_i given z_i satisfies the completeness condition (see Newey and Powell (2003) or Hall and Horowitz (2005)), which

[6] Sufficient conditions for $E[\varepsilon|z, v] = E[\varepsilon|v]$ are that $c_1(z, \eta_1) = c_{1z}(z) + \eta_1$ and $c_2(z, \eta_2)$ is constant, in which case $v = -\frac{\eta_1}{c_2 - \varphi_2}$. The economic implication is that the linear cost coefficient must be separable in z and η_1 and the quadratic cost coefficient cannot depend on z nor cost shocks.

Also even when we use the Matzkin (2003)'s control $v^* = F_{x|z}$, the conditional CDF of x given z as in Florens, Heckman, Meghir, and Vytlacil (2008), the condition $E[\varepsilon|z, v^*] = E[\varepsilon|v^*]$ does not hold in general unless we restrict $c_1(z, \eta_1)$ and $c_2(z, \eta_2)$.

assumes that $E[\delta(x_i, z_{1i})|z_i] = 0$ implies $\delta(x_i, z_{1i}) = 0$ for any $\delta(x_i, z_{1i})$ with finite expectation.[7] To illustrate the implication for a parametric setting we let $f_0(x_i, z_{1i}) = \beta_0' x_i + \beta_{10}' z_{1i}$. Then an alternative function is $\bar{f}(x_i, z_{1i}) = \bar{\beta}' x_i + \bar{\beta}_1' z_{1i} \neq \beta_0' x_i + \beta_{10}' z_{1i}$, so $E[\delta(x_i, z_{1i})|z_i] = (\beta_0 - \bar{\beta})' E[x_i|z_i] + (\beta_{10} - \bar{\beta}_1)' z_{1i}$. If z_i satisfies the standard rank condition - e.g. it includes excluded instruments from z_{1i} that are correlated with x_i - then $E[\delta(x_i, z_{1i})|z_i] = 0$ implies $\delta(x_i, z_{1i}) = 0$, so $\beta_0 = \bar{\beta}$ and $\beta_{10} = \bar{\beta}_1$.

4.1 Generalization of NPIV and NPV-CF

While it is not the focus of this paper, there are sets of assumptions under which neither NPV-CF nor NPIV yields identification but our approach does yield identification. Let $z_i = (z_{1i}, z_{2i})$ and consider a setting where

$$E[\varepsilon_i|z_{1i}] = 0 \quad \text{and} \quad E[\varepsilon_i|z_i, v_i] = E[\varepsilon_i|z_{1i}, v_i]. \tag{12}$$

In this case $E[\varepsilon_i|z_{2i}] \neq 0$ so NPIV is not consistent. Also, since $E[\varepsilon_i|z_i, v_i] = E[\varepsilon_i|z_{1i}, v_i]$, even though ε_i is independent of z_{2i} conditional on v_i it is not independent of z_{1i}, meaning $E[\varepsilon_i|z_i, v_i] \neq E[\varepsilon_i|v_i]$. Thus NPV-CF is also not consistent.[8]

In order to see that our approach to identification works let $h(z_{1i}, v_i) = E[\varepsilon_i|z_{1i}, v_i]$. Then under (12) the new expression for (10) combined with the CMRs yields

$$\begin{aligned} E[y_i|z_i, v_i] &= f(x_i, z_{1i}) + h(z_{1i}, v_i) \\ E[h(z_{1i}, v_i)|z_{1i}] &= 0. \end{aligned} \tag{13}$$

For identification it must be that if both (f_0, h_0) and $(\bar{f}, \bar{h})$ satisfy (13),

$$\Pr(f_0(x_i, z_{1i}) = \bar{f}(x_i, z_{1i}), h_0(z_{1i}, v_i) = \bar{h}(z_{1i}, v_i)) = 1.$$

We show (f_0, h_0) are identified under (1-2) and (12) next, and Section 5.1 develops the estimator for this case.

Theorem 2. *Assume (1-2) and (12). Suppose* $\frac{\partial \Pi_0(z_i)}{\partial z_{2i}'}$ *has the full rank (i.e.* $rank(\frac{\partial \Pi_0(z_i)}{\partial z_{2i}'}) = \dim(x_i)$*) with probability one. Then* $f_0(x_i, z_{1i})$ *and* $h(z_{1i}, v_i)$ *are identified.*

Proof. We prove by contradiction. Suppose it is not identified. Then there must exist functions $\bar{f}(x_i, z_{1i})$ and $\bar{h}(z_{1i}, v_i)$ such that $\delta(x_i, z_{1i}) = f_0(x_i, z_{1i}) - \bar{f}(x_i, z_{1i}) \neq 0$ and $\kappa(z_{1i}, v_i) = h_0(z_{1i}, v_i) - \bar{h}(z_{1i}, v_i) \neq 0$ but $\delta(x_i, z_{1i}) + \kappa(z_{1i}, v_i) = 0$ with probability one (wp1). By taking derivatives w.r.t. z_{2i} to $\delta(x_i, z_{1i}) + \kappa(z_{1i}, v_i) = 0$, we obtain $\frac{\partial \delta(x_i, z_{1i})}{\partial z_{2i}} = 0$ wp1 because $\kappa(z_{1i}, v_i)$ is not a function

[7] The completeness condition is the nonparametric analog of the rank condition for identification in the linear setting.

[8] This last assumption allows, for example, the error ε_i to be conditionally heteroskedastic in z_{1i}.

of z_{2i}. Then because $0 = \frac{\partial \delta(x_i,z_{1i})}{\partial z_{2i}'} = \frac{\partial \delta(x_i,z_{1i})}{\partial x_i'}\frac{\partial x_i}{\partial z_{2i}'} = \frac{\partial \delta(x_i,z_{1i})}{\partial x_i'}\frac{\partial \Pi_0(z_i)}{\partial z_{2i}'}$ and $\frac{\partial \Pi_0(z_i)}{\partial z_{2i}'}$ has the full rank, we also find $\frac{\partial \delta(x_i,z_{1i})}{\partial x_i} = 0$ wp1. Therefore $\delta(x_i, z_{1i})$ must be a function of z_{1i} only wp1. We then find $0 = E[\delta(x_i, z_{1i})|z_{1i}] + E[\kappa(z_{1i}, v_i)|z_{1i}] = \delta(x_i, z_{1i})$ because $E[\kappa(z_{1i}, v_i)|z_{1i}] = 0$ by construction of $h_0(z_{1i}, v_i)$ and $\bar{h}(z_{1i}, v_i)$. We then conclude $\delta(x_i, z_{1i}) = 0$ and $\kappa(z_{1i}, v_i) = 0$ wp1, which is a contradiction. □

5 Estimation

Our estimator is obtained in three steps. We focus on sieve estimation because it is convenient to impose the restriction (5). We use capital letters to denote random variables and lower case letters to denote their realizations. We assume the tuple $\{(Y_i, X_i, Z_i)\}$ for $i = 1, \ldots, n$ are i.i.d. We let X_i be $d_x \times 1$, Z_{1i} be $d_1 \times 1$, Z_{2i} be $d_2 \times 1$, $d_z = d_1 + d_2$ and $d = d_z + d_x$, with $d_x = 1$ for ease of exposition. Let $\{p_j(Z), j = 1, 2, \ldots\}$ denote a sequence of approximating basis functions (e.g. orthonormal polynomials or splines). Let $p^{k_n} = (p_1(Z), \ldots, p_{k_n}(Z))'$, $P = (p^{k_n}(Z_1), \ldots, p^{k_n}(Z_n))'$, and $(P'P)^-$ denote the Moore-Penrose generalized inverse, where k_n tends to infinity but $k_n/n \to 0$. Similarly we let $\{\phi_j(X, Z_1), j = 1, 2, \ldots\}$ denote a sequence of approximating basis functions, $\phi^{Kn} = (\phi_1(X, Z_1), \ldots, \phi_{K_n}(X, Z_1))'$, where K_n tends to infinity but $K_n/n \to 0$.[9]

In the first step to estimate the controls we estimate $\Pi_0(z)$ using

$$\hat{\Pi}(z) = p^{k_n}(z)'(P'P)^- \sum\nolimits_{i=1}^{n} p^{k_n}(z_i)x_i$$

and obtain the control variable as $\hat{v}_i = x_i - \hat{\Pi}(z_i)$.

In the second step we construct approximating basis functions using $\hat{v}$ and z, where we impose the CMR condition (5) by subtracting out the conditional means (conditional on Z). We start by assuming v is known and then show how the setup changes when $\hat{v}$ replaces v. We write basis functions when v is known as

$$\tilde{\varphi}_l(z, v) = \varphi_l(z, v) - \bar{\varphi}_l(z)$$

where $\bar{\varphi}_l(z) = E[\varphi_l(Z, V)|Z = z]$ and $\{\varphi_l(z, v), l = 1, 2, \ldots\}$ denotes a sequence of approximating basis functions generated using $(z, v) \in \text{supp}(Z, V) \equiv \mathcal{W}$, the support of (Z, V). We let $\mathcal{H}$ denote a space of functions that includes h_0, and we let $\|\cdot\|_{\mathcal{H}}$ be a pseudo-metric on $\mathcal{H}$. We define the sieve space $\mathcal{H}_n$ as the collection of functions

$$\mathcal{H}_n = \{h : h = \sum\nolimits_{l \le L_n} a_l \tilde{\varphi}_l(z, v), \|h\|_{\mathcal{H}} < \bar{C}_h, (z, v) \in \mathcal{W}\}$$

for some bounded positive constant $\bar{C}_h$, with $L_n \to \infty$ so that $\mathcal{H}_n \subseteq \mathcal{H}_{n+1} \subseteq \ldots \subseteq \mathcal{H}$ (and $L_n/n \to 0$).

[9] We state specific rate conditions in the next section for our convergence rate results and also for $\sqrt{n}$-consistency and asymptotic normality of linear functionals.

Because v is not known we use instead estimates of the approximating basis functions, which we denote as $\hat{\tilde{\varphi}}_l(z,\hat{v}) = \varphi_l(z,\hat{v}) - \hat{\tilde{\varphi}}_l(z)$, where $\hat{\tilde{\varphi}}_l(z) = \hat{E}[\varphi_l(Z,\hat{V})|Z=z]$. We then construct the approximation of $h(z,v)$ as [10]

$$\begin{aligned}\hat{h}_{L_n}(z,\hat{v}) &= \sum_{l=1}^{Ln} a_l\{\varphi_l(z,\hat{v}) - \hat{E}[\varphi_l(Z,\hat{V})|Z=z]\} \\ &= \sum_{l=1}^{Ln} a_l\{\varphi_l(z,\hat{v}) - p^{k_n}(z)'(P'P)^{-}\sum_{i=1}^{n} p^{k_n}(z_i)\varphi_l(z_i,\hat{v}_i)\},\end{aligned} \tag{14}$$

with coefficients, $(a_1,\ldots,a_{L_n})$ to be estimated in the last step. We approximate the sieve space $\mathcal{H}_n$ with $\hat{\mathcal{H}}_n$ using (14), so $\hat{\mathcal{H}}_n$ is given by

$$\hat{\mathcal{H}}_n = \{h : h = \sum_{l\le L_n} a_l\hat{\tilde{\varphi}}_l(z,\hat{v}), \|h\|_{\mathcal{H}} < \bar{C}_h, (z,\hat{v}) \in \mathcal{W}\}. \tag{15}$$

In the last step we define $\mathcal{F}$ as the space of functions that includes f_0, and we let $\|\cdot\|_{\mathcal{F}}$ be a pseudo-metric on $\mathcal{F}$. We define the sieve space $\mathcal{F}_n$ as the collection of functions

$$\mathcal{F}_n = \{f : f = \sum_{l\le K_n} \beta_l\phi_l(x,z_1), \|f\|_{\mathcal{F}} < \bar{C}_f, (x,z_1) \in \text{supp}(X,Z_1)\}$$

for some bounded positive constant $\bar{C}_f$, with $K_n \to \infty$ so that $\mathcal{F}_n \subseteq \mathcal{F}_{n+1} \subseteq \ldots \subseteq \mathcal{F}$ (and $K_n/n \to 0$). Then our multi-step sieve estimator is obtained by solving

$$(\hat{f},\hat{h}) = \text{arginf}_{(f,h)\in\mathcal{F}_n\times\hat{\mathcal{H}}_n} \sum_{i=1}^{n}\{y_i - (f(x_i,z_{1i}) + h(z_i,\hat{v}_i))\}^2/n \tag{16}$$

where $\hat{v}_i = x_i - \hat{\Pi}(z_i)$.

Equivalently we can write the last step estimation as

$$\min_{(\beta_1,\ldots,\beta_{K_n},a_1,\ldots,a_{L_n})} \sum_{i=1}^{n}\{y_i - (\sum_{k=1}^{K_n}\beta_k\phi_k(x_i,z_{1i}) + \sum_{l=1}^{Ln} a_l\hat{\tilde{\varphi}}_l(z_i,\hat{v}_i))\}^2/n. \tag{17}$$

With fixed k_n, L_n, and K_n our estimator is just a three-stage least squares estimator. Once we obtain the estimates $(\hat{f},\hat{h})$ we can also estimate linear functionals of (f_0,h_0) using plug-in methods (see Section 7).

5.1 Estimation of Model (12)-(13)

To estimate the model (12)-(13) we have only to replace the control function with $h(z_1,v)$ and

[10] We can use different sieves (e.g., power series, splines of different lengths) to approximate $E[\varphi_l(Z,V)|Z=z]$ and $\Pi(z)$ depending on their smoothness, but we assume one uses the same sieves for notational simplicity.

its approximation with

$$\begin{aligned}\hat{h}_{L_n}(z_1,\hat{v}) &= \sum_{l=1}^{Ln} a_l\hat{\tilde{\varphi}}_l(z_{1i},\hat{v}_i) \\ &= \sum_{l=1}^{Ln} a_l\{\varphi_l(z_1,\hat{v}) - \hat{E}[\varphi_l(Z_1,\hat{V})|Z_1=z_1]\} \\ &= \sum_{l=1}^{Ln} a_l\{\varphi_l(z_1,\hat{v}) - p_1^{k_n}(z_1)'(P_1'P_1)^- \sum_{i=1}^{n} p_1^{k_n}(z_{1i})\varphi_l(z_{1i},\hat{v}_i)\}\end{aligned}$$

where $p_1^{k_n}(z_1) = (p_1(z_1),\ldots,p_{k_n}(z_1))'$ and $P_1 = (p_1^{k_n}(z_{11}),\ldots,p_1^{k_n}(z_{1n}))'$, so we demean the approximating functions $\varphi_l(z_1,\hat{v})$ w.r.t. z_1 only. We then estimate the function f together with the control function as

$$(\hat{\beta},\hat{a}) = arg\min_{(\beta_1,\ldots,\beta_{K_n},a_1,\ldots,a_{L_n})} \sum_{i=1}^{n}\{y_i - (\sum_{k=1}^{K_n}\beta_k\phi_k(x_i,z_{1i}) + \sum_{l=1}^{Ln} a_l\hat{\tilde{\varphi}}_l(z_{1i},\hat{v}_i))\}^2/n$$

such that $\hat{f} = \sum_{k=1}^{K_n}\hat{\beta}_k\phi_k(x_i,z_{1i})$ and $\hat{h}(z_1,\hat{v}) = \sum_{l=1}^{Ln}\hat{a}_l\hat{\tilde{\varphi}}_l(z_{1i},\hat{v}_i)$.

In the following Sections 6-7 we develop asymptotic properties of the estimator for the model (1-2) and (5). Similar results can be obtained for the estimator of the model (12)-(13) with minor modifications.

6 Convergence Rates

We obtain the convergence rates building on Newey, Powell, and Vella (1999). We differ from their approach as we have another nonparametric estimation stage in the middle step of estimation that arises due to our identification approach different from NPV. This creates additional terms in the convergence rate results.

We introduce additional notation. We write $g_0(z_i,v_i) \equiv g_0(x_i,z_{1i}\cup z_i,v_i) = f_0(x_i,z_{1i}) + h_0(z_i,v_i)$ for ease of notation. For a random matrix D, let $\|D\| = (\mathrm{tr}(D'D))^{1/2}$, and let $\|D\|_\infty$ be the infimum of constants C such that $\Pr(\|D\| < C) = 1$. We derive the convergence rates of the nonparametric estimator $\hat{g} = \hat{f} + \hat{h}$ to g_0 and $\hat{f}$ to f_0 only for the purpose of obtaining the $\sqrt{n}$-consistency and the asymptotic normality of the linear functional estimators of g_0 or f_0. Below Assumptions C1 and C2 along with identification of (f_0,h_0) as shown in Section 4 ensure the rate results we derive.

Assumption 1 (C1)**.** *(i)* $\{(Y_i,X_i,Z_i)\}_{i=1}^n$ *are i.i.d.,* $V = X - E[X|Z]$*, and var*$(X|Z)$*, var*$(Y|Z,V)$*, and var*$(\varphi_l(Z,V)|Z)$ *for all* l *are bounded; (ii)* (Z,X) *are continuously distributed with densities that are bounded away from zero on their supports, which are compact and the distribution of* X *conditional on* Z *satisfies the completeness condition; (iii)* $\Pi_0(z)$ *is continuously differentiable of order* s_1 *and all the derivatives of order* s_1 *are bounded on the support of* Z*; (iv)* $\bar{\varphi}_l(Z)$ *is continuously differentiable of order* s_2 *and all the derivatives of order* s_2 *are bounded for all* l *on the support of* Z*; (v)* $h_0(Z,V)$ *is Lipschitz and is continuously differentiable of order* s *and all the derivatives of order* s *are bounded on the support of* (Z,V)*; (vi)* $\varphi_l(Z,V)$ *is Lipschitz and is twice continuously differentiable in* v *and its first and second derivatives are bounded for all* l*; (vii)* $f_0(X,Z_1)$ *is continuously differentiable of order* s *and all the derivatives of order* s *are bounded on the support of* (X,Z_1)*.*

Assumptions C1 (iii), (iv), (v), and (vii) ensure that the unknown functions $\Pi_0(Z)$, $\bar{\varphi}_l(Z)$, $h_0(Z,V)$, and $f_0(X,Z_1)$ belong to a Hölder class of functions, so they can be approximated up to the orders of $O(k_n^{-s_1/d_z})$, $O(k_n^{-s_2/d_z})$, $O(L_n^{-s/d})$, and $O(K_n^{-s/(d_x+d_1)})$ respectively when polynomials or splines approximation is used (see Timan (1963), Schumaker (1981), Newey (1997), and Chen (2007)). Assumption C1 (vi) is satisfied for polynomial and spline basis functions with appropriate orders. Assumptions C1 (i)-(ii) are about the structure of the data. Assumption C1 (ii) includes the completeness condition for identification and other conditions can be relaxed with additional complexity (e.g., a trimming device as in Newey, Powell, and Vella (1999)). Assumption C1 (v) and (vii) maintain that f_0 and h_0 have the same order of smoothness for ease of notation, but it is possible to allow them to differ.

Next we impose the rate conditions that restrict the growth of k_n, K_n, and L_n as n tends to infinity. We write $\mathbf{L}_n = K_n + L_n$.

Assumption 2 (C2)**.** *Let* $\triangle_{n,1} = k_n^{1/2}/\sqrt{n} + k_n^{-s_1/d_z}$, $\triangle_{n,2} = k_n^{1/2}/\sqrt{n} + k_n^{-s_2/d_z}$, *and* $\triangle_n = \max\{\triangle_{n,1}, \triangle_{n,2}\}$. *For polynomial approximations* $\mathbf{L}_n^{1/2}(L_n^3 + L_n^{1/2}k_n^{3/2}/\sqrt{n} + L_n^{1/2})\triangle_n \to 0$, $\mathbf{L}_n^3/n \to 0$ *and* $k_n^3/n \to 0$. *For spline approximations* $\mathbf{L}_n^{1/2}(L_n^{3/2} + L_n^{1/2}k_n/\sqrt{n} + L_n^{1/2})\triangle_n \to 0$, $\mathbf{L}_n^2/n \to 0$ *and* $k_n^2/n \to 0$.

For any differentiable function $c(w)$, let $|\mu| = \sum_{j=1}^{\dim(w)} \mu_j$ and define $\partial^\mu c(w) = \frac{\partial^{|\mu|} c(w)}{\partial w_1 \cdots \partial w_{\dim(w)}}$. Also for integer δ define $|c(w)|_\delta = \max_{|\mu| \le \delta} \sup_{w \in \text{supp}(w)} ||\partial^\mu c(w)||$ with $|c(w)|_0 = \sup_{w \in \text{supp}(w)} ||c(w)||$. Also let $\phi^K(x, z_1) \equiv (\phi_1(x, z_1), \ldots, \phi_K(x, z_1))'$.

Theorem 3. *Suppose Assumptions C1-C2 are satisfied. Then*

$$\begin{aligned} (a)\ \Big(\int [\hat{g}(z,v) - g_0(z,v)]^2 d\mu_0(z,v)\Big)^{1/2} &= O_p(\sqrt{\mathbf{L}_n/n} + L_n\triangle_n + \mathbf{L}_n^{-s/d}),\ \textit{and} \\ (b)\ |\hat{f}(x,z_1) - f_0(x,z_1)|_\delta &= O_p(\zeta_\delta(K_n)[\sqrt{\mathbf{L}_n/n} + L_n\triangle_n + \mathbf{L}_n^{-s/d}] + K_n^{-s/(d_x+d_1)}) \end{aligned}$$

where $\mu_0(z,v)$ *denotes the distribution function of* (z,v) *and* $|\phi^K(x,z_1)|_\delta \le \zeta_\delta(K)$.

In Theorem 3 the term $L_n\triangle_n$ arises because of the estimation error from the first and second steps of estimation. With no estimation error from these stages we would obtain standard convergence rates of series estimators.

7 Asymptotic Normality

Following Newey (1997) and Newey, Powell, and Vella (1999) we consider inference for the linear functionals of g, $\theta = \alpha(g)$.[11] The estimator $\hat{\theta} = \alpha(\hat{g})$ of $\theta_0 = \alpha(g_0)$ is a well-defined "plug-in" estimator, and because of the linearity of $\alpha(g)$ we have

$$\hat{\theta} = \mathcal{A}\hat{\beta}, \mathcal{A} = (\alpha(\phi_1), \ldots, \alpha(\phi_{K_n}), \alpha(\tilde{\varphi}_1), \ldots, \alpha(\tilde{\varphi}_{L_n}))$$

[11] This includes θ's being functionals of f only.

where we let $\hat{\beta} = (\hat{\beta}_1, \ldots, \hat{\beta}_{K_n}, \hat{a}_1, \ldots, \hat{a}_{L_n})'$ that solves (17). This setup includes (e.g.) partially linear models, where f contains some parametric components, and the weighted average derivative, where one estimates the average response of y with respect to the marginal change of x or z_1. For example if f is partially linear, then each row vector of $\mathcal{A}$ only consists of ones and zeros such that ones select particular parameters from the parametric components. More generally, if $\mathcal{A}$ depends on unknown population objects, we can estimate it using $\hat{\mathcal{A}} = \partial\alpha(\hat{\psi}_i^{\mathbf{L}\prime}\beta)/\partial\beta'|_{\beta=\hat{\beta}}$ where $\hat{\psi}_i^{\mathbf{L}} = (\phi_1(x_i, z_{1i}), \ldots, \phi_K(x_i, z_{1i}), \hat{\tilde{\varphi}}_1(z_i, \hat{v}_i), \ldots, \hat{\tilde{\varphi}}_L(z_i, \hat{v}_i))'$, so that $\hat{\theta} = \hat{\mathcal{A}}\hat{\beta}$ (see Newey (1997)).

We focus on conditions that provide for $\sqrt{n}$-asymptotics and allow for a straightforward consistent estimator for the standard errors of $\hat{\theta}$.[12] If there exists a Riesz representer $\nu^*(Z, V)$ such that

$$\alpha(g) = E[\nu^*(Z, V)g(Z, V)] \tag{18}$$

for any $g = (f, h) \in \mathcal{F} \times \mathcal{H}$ that can be approximated by power series or splines in the mean-squared norm, then we can obtain $\sqrt{n}$-consistency and asymptotic normality for $\hat{\theta}$, expressed as

$$\sqrt{n}(\hat{\theta} - \theta_0) \to_d N(0, \Omega),$$

for some asymptotic variance matrix Ω. In Assumption C1 we take both $\mathcal{F}$ and $\mathcal{H}$ as Hölder spaces of functions, which ensures the approximation of g in the mean-squared norm (see e.g., Newey (1997), Newey, Powell, and Vella (1999), and Chen (2007)). Defining $\rho_v(Z) = E[\nu^*(Z, V)(\frac{\partial h_0(Z,V)}{\partial V} - E[\frac{\partial h_0(Z,V)}{\partial V}|Z])|Z]$ and $\rho_{\bar{\varphi}_l}(Z) = E[a_l\nu^*(Z, V)|Z]$, the asymptotic variance of the estimator $\hat{\theta}$ is given by

$$\begin{aligned} \Omega &= E[\nu^*(Z, V)\mathrm{var}(Y|Z, V)\nu^*(Z, V)'] + E[\rho_v(Z)\mathrm{var}(X|Z)\rho_v(Z)'] \\ &\quad + \lim_{n\to\infty} \sum_{l=1}^{L_n} E[\rho_{\bar{\varphi}_l}(Z)\mathrm{var}(\varphi_l(Z, V)|Z)\rho_{\bar{\varphi}_l}(Z)']. \end{aligned} \tag{19}$$

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

Assumptions C1, R1, N1, and N2 below are sufficient to characterize the asymptotic normality of $\hat{\theta}$ and also a consistent estimator for the asymptotic variance of $\hat{\theta}$. Below let $\tilde{\varphi}^L(z, v) \equiv (\tilde{\varphi}_1(z, v), \ldots, \tilde{\varphi}_L(z, v))'$ and $\psi^{\mathbf{L}}(z_i, v_i) \equiv (\phi^K(x, z_1)', \tilde{\varphi}^L(z, v)')'$.

Assumption 3 (R1)**.** *There exist $\nu^*(Z, V)$ and $\beta_{\mathbf{L}}$ such that $E[||\nu^*(Z, V)||^2] < \infty$, $\alpha(g_0) = E[\nu^*(Z, V)g_0(Z, V)]$, $\alpha(\phi_k) = E[\nu^*(Z, V)\phi_k]$ for $k = 1, \ldots, K$, $\alpha(\tilde{\varphi}_l) = E[\nu^*(Z, V)\tilde{\varphi}_l]$ for $l = 1, \ldots, L$, and $E[||\nu^*(Z, V) - \psi^{\mathbf{L}}(Z, V)'\beta_{\mathbf{L}}||^2] \to 0$ as $\mathbf{L} \to \infty$.*

Assumption R1 restricts the class of linear functionals we consider that yield the $\sqrt{n}$-consisteny and also requires $\nu^*(Z, V)$ be well approximated by the approximating basis functions we use to

[12] Developing the asymptotic distributions of the functionals that do not yield the $\sqrt{n}$-consistency is also possible based on the convergence rates result we obtained and alternative assumptions on the functionals of interest (see Newey, Powell, and Vella (1999)).

approximate g_0. To present the theorem, we need additional notation and assumptions. Let $a_L = (a_1, \ldots, a_L)'$ with an abuse of notation.

Assumption 4 (N1). *(i) there exist δ, γ, and $\beta_{\mathbf{L}}$ such that $|g_0(z,v) - \beta_{\mathbf{L}}'\psi^{\mathbf{L}}(z,v)|_\delta \leq C\mathbf{L}^{-\gamma}$ (which also implies $|h_0(z,v) - a_L'\tilde{\varphi}^L(z,v)|_\delta \leq CL^{-\gamma}$); (ii) var$(Y_i|Z_i,V_i)$ is bounded away from zero, $E[\eta_i^4|Z_i,V_i]$ and $E[V_i^4|Z_i]$ are bounded and $E[\tilde{\varphi}_l(Z_i,V_i)^4|Z_i]$ is bounded for all l.*

Assumption N1 (i) is satisfied for f_0 and h_0 that belong to the Hölder class and then we can take (e.g.) $\gamma = s/d$. The bounded conditional fourth moments are imposed to apply for appropriate central limit theorems. Next we impose the rate conditions that restrict the growth of k_n and $\mathbf{L}_n = K_n + L_n$ as n tends to infinity.

Assumption 5 (N2). *Let $\triangle_{n,1} = k_n^{1/2}/\sqrt{n} + k_n^{-s_1/d_z}$, $\triangle_{n,2} = k_n^{1/2}/\sqrt{n} + k_n^{-s_2/d_z}$, and $\triangle_n = \max\{\triangle_{n,1}, \triangle_{n,2}\}$. $\sqrt{n}k_n^{-s_1/d_z} \to 0, \sqrt{n}k_n^{-s_2/d_z} \to 0, \sqrt{n}k_n^{1/2}L_n^{-s/d} \to 0, \sqrt{n}\mathbf{L}_n^{-s/d} \to 0$ and they are sufficiently small. For the polynomial approximations $\frac{\mathbf{L}_n^2 + \mathbf{L}_n L_n^3 k_n + \mathbf{L}_n^{1/2}(L_n^4 k_n^{3/2} + k_n^{5/2})}{\sqrt{n}} \to 0$ and for the spline approximations $\frac{\mathbf{L}_n^{3/2} + \mathbf{L}_n L_n^{3/2} k_n^{1/2} + \mathbf{L}_n^{1/2}(L_n^{5/2} k_n + k_n^{3/2}) + L_n^{3/2} k_n^{3/2}}{\sqrt{n}} \to 0$.*

Theorem 4. *Suppose Assumptions C1, R1, and N1-N2 are satisfied. Then*

$$\sqrt{n}(\hat{\theta} - \theta_0) \to_d N(0, \Omega).$$

Based on this asymptotic distribution, one can construct the confidence intervals of θ_0 and calculate standard errors in a straightforward manner. Let $\hat{g}(z_i, \hat{v}_i) = \hat{f}(x_i, z_{1i}) + \hat{h}(z_i, \hat{v}_i)$ and $\hat{g}_i = \hat{g}(z_i, \hat{v}_i)$. Define $\hat{\psi}_i^{\mathbf{L}} = (\phi_1(x_i, z_{1i}), \ldots, \phi_K(x_i, z_{1i}), \hat{\tilde{\varphi}}^L(z_i, \hat{v}_i)')'$ where $\hat{\tilde{\varphi}}^L(z_i, v_i) = (\hat{\tilde{\varphi}}_1(z_i, v_i), \ldots, \hat{\tilde{\varphi}}_L(z_i, v_i))'$. Let

$$\begin{aligned}
\hat{\mathcal{I}} &= \textstyle\sum_{i=1}^n \hat{\psi}_i^{\mathbf{L}} \hat{\psi}_i^{\mathbf{L}\prime}/n, \hat{\Sigma} = \sum_{i=1}^n (y_i - \hat{g}(z_i, \hat{v}_i))^2 \hat{\psi}_i^{\mathbf{L}} \hat{\psi}_i^{\mathbf{L}\prime}/n \qquad (20)\\
\hat{\mathcal{I}}_1 &= P'P/n, \hat{\Sigma}_1 = \sum_{i=1}^n \hat{v}_i^2 p^k(z_i) p^k(z_i)'/n, \hat{\Sigma}_{2,l} = \sum_{i=1}^n \{\varphi_l(z_i, \hat{v}_i) - \hat{\tilde{\varphi}}_l(z_i)\}^2 p^k(z_i) p^k(z_i)'/n\\
\hat{H}_{11} &= \sum_{i=1}^n \sum_{l=1}^L \hat{a}_l \frac{\partial \varphi_l(z_i, \hat{v}_i)}{\partial v_i} \hat{\psi}_i^{\mathbf{L}} p^k(z_i)'/n,\\
\hat{H}_{12} &= \sum_{i=1}^n p^k(z_i)'((P'P)^- \sum_{j=1}^n p^k(z_j) \frac{\partial \sum_{l=1}^L \hat{a}_l \varphi_l(z_j, \hat{v}_j)}{\partial v_j}) \hat{\psi}_i^{\mathbf{L}} p^k(z_i)'/n,\\
\hat{H}_{2,l} &= \textstyle\sum_{i=1}^n \hat{a}_l \hat{\psi}_i^{\mathbf{L}} p^k(z_i)'/n, \hat{H}_1 = \hat{H}_{11} - \hat{H}_{12}.
\end{aligned}$$

Then, we can estimate Ω consistently by

$$\hat{\Omega} = \mathcal{A}\hat{\mathcal{I}}^{-1} \left[\hat{\Sigma} + \hat{H}_1 \hat{\mathcal{I}}_1^{-1} \hat{\Sigma}_1 \hat{\mathcal{I}}_1^{-1} \hat{H}_1' + \textstyle\sum_{l=1}^{L_n} \hat{H}_{2,l} \hat{\mathcal{I}}_1^{-1} \hat{\Sigma}_{2,l} \hat{\mathcal{I}}_1^{-1} \hat{H}_{2,l}' \right] \hat{\mathcal{I}}^{-1} \mathcal{A}'. \qquad (21)$$

Theorem 5. *Suppose Assumptions C1, R1, and N1-N2 are satisfied. Then $\hat{\Omega} \to_p \Omega$.*

This $\hat{\Omega}$ is the heteroskedasticity robust variance estimator that accounts for the first and second steps of estimation. The first variance term $\mathcal{A}\hat{\mathcal{T}}^{-1}\hat{\Sigma}\hat{\mathcal{T}}^{-1}\mathcal{A}'$ corresponds to the variance estimator without error from the first and second steps of estimation. The second variance term accounts for the estimation of v (and corresponds to the second term in (19)). The third variance term accounts for the estimation of $\bar{\varphi}_l(\cdot)$'s (and corresponds to the third term in (19)). If we view our model as a parametric one with fixed k_n, K_n, and L_n, the same variance estimator $\hat{\Omega}$ can be used as the estimator of the variance for the parametric model (e.g, Newey (1984) and Murphy and Topel (1985)).

7.1 Discussion

We discuss Assumption R1 for the partially linear model and the weighted average derivative. Consider a partially linear model of the form

$$f_0(x, z_1) = x_1'\beta_{10} + f_{20}(x_{-1}, z_1)$$

where x can be multi-dimensional and x_1 is a subvector of x such that $x = (x_1, x_{-1})$. Then we have

$$\beta_{10} = \alpha(g_0) = E[\nu^*(Z, V)g_0(Z, V)]$$

where $\nu^*(z, v) = (E[q(Z, V)q(Z, V)'])^{-1}q(z, v)$ and $q(z, v)$ is the residual from the mean-square projection of x_1 on the space of functions that are additive in functions of (x_{-1}, z_1) and any $h(z, v)$ such that $E[h(Z, V)|Z] = 0$.[13] Thus we can approximate $q(z, v)$ by the mean-square projection residual of x_1 on $\psi^{\mathbf{L}}_{-1}(z_i, v_i) \equiv (\phi_1(x_{-1i}, z_{1i}), \ldots, \phi_K(x_{-1i}, z_{1i}), \tilde{\varphi}^L(z_i, v_i)')'$, and then use these estimates to approximate $\nu^*(z, v)$.

Next consider a weighted average derivative of the form

$$\alpha(g_0) = \int_{\bar{\mathcal{W}}} \varpi(x, z_1, \kappa(z, v))\frac{\partial g_0(z, v)}{\partial x}d(z, v) = \int \varpi(x, z_1, \kappa(z, v))\frac{\partial f_0(x, z_1)}{\partial x}d(z, v)$$

where the weight function $\varpi(x, z_1, \kappa(z, v))$ puts zero weights outside $\bar{\mathcal{W}} \subset \mathcal{W}$ and $\kappa(z, v)$ is some function such that $E[\kappa(Z, V)|Z] = 0$. This is a linear functional of g_0. Integration by parts shows that

$$\alpha(g_0) = -\int_{\bar{\mathcal{W}}} \text{proj}(\mu_0(z, v)^{-1}\frac{\partial \varpi(x, z_1, \kappa(z, v))}{\partial x}|\mathcal{S})g_0(z, v)d\mu_0(z, v) = E[\nu^*(Z, V)g(Z, V)]$$

where $\text{proj}(\cdot|\mathcal{S})$ denotes the mean-square projection on the space of functions that are additive in functions of (x, z_1) and any $h(z, v)$ such that $E[h(Z, V)|Z] = 0$ (so the Riesz representer $\nu^*(z, v)$ is well-defined), and $\nu^*(z, v) = -\text{proj}(\mu_0(z, v)^{-1}\frac{\partial \varpi(x, z_1, \kappa(z, v))}{\partial x}|\mathcal{S})$ with $\mu_0(z, v)$ denoting the distribution of (z, v). We can then approximate $\nu^*(z, v)$ using a mean-square projection of

[13] Note that existence of the Riesz representer in this setting requires $E[q(Z, V)q(Z, V)']$ to be nonsingular.

$\mu_0(z,v)^{-1}\frac{\partial\varpi(x,z_1,\kappa(z,v))}{\partial x}$ on $\psi^{\mathbf{L}}(z,v)$.

8 Simulation Study

We conduct two monte carlos simulations to evaluate the performance of the NPV-CF estimator and our CMR-CF estimator. The first set of monte carlos is based on the economic examples provided in Section 3 where the structural function $f(x)$ is parametric and the second set uses a nonlinear setup from Newey and Powell (2003) where the structural function is estimated nonparametrically.

8.1 Monte Carlos Based on Parametric Models

In the first set we consider six models. The outcome equations are parametric so $f(x)$ is known up to a finite set of parameters. The selection equations are treated as unknown to the practitioner and we use nonparametric regressions for them in the simulation.

The six designs are given as:

$$\begin{aligned}
[1]\ y_i &= \alpha+\beta x_i+\gamma x_i^2+\varepsilon_i\ ;\ x_i=z_i+(3\varepsilon_i+\varsigma_i)\cdot\log(z_i)\\
[2]\ y_i &= \alpha+\beta x_i+\gamma x_i^2+\varepsilon_i\ ;\ x_i=z_i+(3\varepsilon_i+\varsigma_i)/\exp(z_i)\\
[3]\ y_i &= \alpha+\beta x_i+\gamma\log x_i+\varepsilon_i\ ;\ x_i=z_i+(3\varepsilon_i+\varsigma_i)/\exp(z_i)\\
[4]\ y_i &= \alpha+\beta x_i+\gamma\log x_i+\varepsilon_i\ ;\ x_i=z_i+(3\varepsilon_i+\varsigma_i+\varepsilon_i\cdot\varsigma_i)/\exp(z_i)\\
[5]\ y_i &= \alpha+\beta x_i+\varepsilon_i\ ;\ x_i=z_i+(3\varepsilon_i+\varsigma_i)/\exp(z_i)\\
[6]\ y_i &= \alpha+\beta x_i+\gamma x_i^2+\varepsilon_i\ ;\ x_i=z_i+(3\varepsilon_i+\varsigma_i).
\end{aligned}$$

These designs can be obtained from the underlying decision problem of (7) by varying the structural function $f(x)$ and the cost function $c(x,z,\eta)$.[14]

We generate simulation data based on the following distributions: $\varepsilon_i\sim U_\varepsilon$, $\varsigma_i\sim U_\varsigma$, $z_i=2+2U_z$, where each U_ε, U_ς, and U_z independently follows the uniform distribution supported on $[-1/2,1/2]$ so all three random variables ε_i, ς_i, and z_i are independent of one another. In all designs x_i is correlated with ε_i and the CMR condition, $E[\varepsilon_i|z_i]=0$ holds. The NPV-CF restriction (4) is violated in designs [1]-[5] and holds in design [6].[15] We set the true parameter values at $(\alpha_0,\beta_0,\gamma_0)=(1,1,-1)$ and the data is generated with the sample size of $n=1,000$.

All of the estimators are based on a first stage estimation residual $\hat{v}_i=x_i-(\hat{\pi}_0+\hat{\pi}_1z_i+\hat{\pi}_2z_i^2)$ although estimates are robust to adding higher order terms.[16] The structural function $f(x_i)$ is given

[14] For example we obtain design [1] by letting $c_2(z,\eta_2)$ be constant and $c_1(z,\eta_1)$ include the leading term z and the interaction term $\eta_1\log(z)$, where $\eta_1=3\varepsilon+\varsigma$ is a noisy signal of ε. The selection equation (9) is then $x=\frac{\varphi_1-c_1(z,\eta_1)}{c_2(z,\eta_2)-\varphi_2}=z+(3\varepsilon+\varsigma)\cdot\log(z)$. The other designs can be derived in a similar way.

[15] For example, in design [2] we have $v_i=x_i-E[x_i|z_i]=(3\varepsilon_i+\varsigma_i)/\exp(z_i)$. Then we have $\varepsilon_i=(\exp(z_i)v_i-\varsigma_i)/3$ and therefore $E[\varepsilon_i|z_i,v_i]=(\exp(z_i)v_i-E[\varsigma_i|z_i,v_i])/3$, and this cannot be written as a function of v_i only.

[16] Root mean-squared errors were similar across all estimators whether we used two or more higher order terms. Thus if we followed Newey, Powell, and Vella (1999) and used cross validation (CV) to discriminate between alternative

by the designs [1]-[6]. The classic control function (CCF) approach just includes the control in an additive manner. Our analysis begins with this CCF estimator which then posits

$$y_i = f(x_i) + \rho\hat{v}_i + \eta_i$$

and estimates the model using least squares. The NPV-CF estimator is obtained by estimating

$$y_i = f(x_i) + h(\hat{v}_i) + \eta_i,$$

where we approximate $h(\hat{v}_i)$ as $h(\hat{v}_i) = \sum_{l=1}^{5} a_l \hat{v}_i^l$. Since the NPV-CF does not separately identify the constant term we normalize $h(0) = 0$ so that the constant term α is also identified. Our results are robust to adding higher orders of polynomials to fit $h(\hat{v}_i)$.

We obtain the CMR-CF estimator by using the first stage estimation residual $\hat{v}_i$ to construct approximating functions $\tilde{v}_{1i} = \hat{v}_i$, $\tilde{v}_{2i} = \hat{v}_i^2 - \hat{E}[\hat{v}_i^2|z_i]$, $\tilde{v}_{3i} = \hat{v}_i^3 - \hat{E}[\hat{v}_i^3|z_i]$ where $\hat{E}[\cdot|z_i]$ is estimated using least squares with regressors $(1, z_i, z_i^2)$. Interactions with polynomials of z_i like $z_i\hat{v}_i$ and $z_i^2\hat{v}_i$ are defined similarly. In the last step we estimate the parameters as

$$(\hat{\alpha}, \hat{\beta}, \hat{\gamma}, \hat{a}) = \text{argmin} \sum_{i=1}^{n} \{y_i - (f(x_i; \alpha, \beta, \gamma) + h(z_i, \hat{v}_i))\}^2/n$$

where $h(z_i, \hat{v}_i) = \sum_{l=1}^{L} a_l \tilde{v}_{li}$ depends on the simulation designs. The choice of the basis in the finite sample is not a consistency issue but it is an efficiency issue and we vary this choice across specifications. In design [1] we use $\tilde{v}_{1i}$ and $z_i\tilde{v}_i$ as the controls. In designs [2], [5], and [6] we use the controls $\tilde{v}_{1i}, \tilde{v}_{2i}$, and $z_i\tilde{v}_i$. In design [3] we use the controls $\tilde{v}_{1i}, \tilde{v}_{2i}, z_i\tilde{v}_i$, and $z_i^2\tilde{v}_i$, and in design [4] we use $\tilde{v}_{1i}, \tilde{v}_{2i}, \tilde{v}_{3i}, \tilde{v}_{4i}, z_i\tilde{v}_i$.

We report the biases and the RMSE's based on 200 repetitions of the estimations. The simulation results in Tables I-VI show that CCF and NPV-CF are biased in all designs except [5] and [6] for which the theory says they should be consistent. The CMR-CF is robust regardless of the designs. In design [5] all three approaches produce correct estimates because the outcome equation is linear, which is consistent with our discussion in Appendix A. In design [6] all three approaches are consistent because the restriction (4) holds. We conclude that our CMR-CF approach is consistent in these designs regardless of whether the model is linear or nonlinear or whether the restriction (4) holds while the CCF and NPV-CF approaches are not robust when the restriction (4) does not hold.

specifications we would be indifferent between this simplest specification and the ones with the higher order terms.

Table I: Design [1], $\alpha_0 = 1, \beta_0 = 1, \gamma_0 = -1$

Nonlinear & The condition (4) does not hold

		mean	bias	RMSE
CCF	α	0.7076	-0.2924	0.2952
	β	1.3078	0.3078	0.3094
	γ	-1.0679	-0.0679	0.0682
NPV-CF	α	0.6655	-0.3345	0.3395
	β	1.3677	0.3677	0.3738
	γ	-1.0917	-0.0917	0.0938
CMR-CF	α	0.9978	-0.0022	0.0548
	β	1.0021	0.0021	0.0503
	γ	-1.0005	-0.0005	0.0109

Table II: Design [2], $\alpha_0 = 1, \beta_0 = 1, \gamma_0 = -1$

Nonlinear & The condition (4) does not hold

		mean	bias	RMSE
CCF	α	1.5331	0.5331	0.5452
	β	0.4056	-0.5944	0.6055
	γ	-0.8496	0.1504	0.1529
NPV-CF	α	1.3535	0.3535	0.3767
	β	0.6283	-0.3717	0.3948
	γ	-0.9090	0.0910	0.0966
CMR-CF	α	0.9933	-0.0067	0.1478
	β	1.0079	0.0079	0.1611
	γ	-1.0021	-0.0021	0.0405

Table III: Design [3],$\alpha_0 = 1, \beta_0 = 1, \gamma_0 = -1$

Nonlinear & The condition (4) does not hold

		mean	bias	RMSE
CCF	α	0.5818	-0.4182	0.4235
	β	1.5048	0.5048	0.5108
	γ	-1.9246	-0.9246	0.9367
NPV-CF	α	0.7750	-0.2250	0.2405
	β	1.3042	0.3042	0.3200
	γ	-1.5861	-0.5861	0.6156
CMR-CF	α	0.9943	-0.0057	0.1103
	β	1.0076	0.0076	0.1255
	γ	-1.0144	-0.0144	0.2249

Table IV: Design [4], $\alpha_0 = 1, \beta_0 = 1, \gamma_0 = -1$

Nonlinear & The condition (4) does not hold

		mean	bias	RMSE
CCF	α	0.6109	-0.3891	0.3950
	β	1.4702	0.4702	0.4769
	γ	-1.8617	-0.8617	0.8751
NPV-CF	α	0.7794	-0.2206	0.2371
	β	1.3333	0.3333	0.3497
	γ	-1.6687	-0.6687	0.6988
CMR-CF	α	1.0003	0.0003	0.1117
	β	1.0005	0.0005	0.1267
	γ	-1.0016	-0.0016	0.2262

Table V: Design [5], $\alpha_0 = 1, \beta_0 = 1$

Linear & The condition (4) does not hold

		mean	bias	RMSE
CCF	α	0.9993	-0.0007	0.0343
	β	1.0004	0.0004	0.0172
NPV-CF	α	1.0010	0.0010	0.0417
	β	0.9997	-0.0003	0.0192
CMR-CF	α	0.9991	-0.0009	0.0343
	β	1.0005	0.0005	0.0171

Table VI: Design [6], $\alpha_0 = 1, \beta_0 = 1, \gamma_0 = -1$

Nonlinear & The condition (4) holds

		mean	bias	RMSE
CCF	α	0.9991	-0.0009	0.0354
	β	1.0010	0.0010	0.0200
	γ	-1.0002	-0.0002	0.0024
NPV-CF	α	0.9997	-0.0003	0.0350
	β	1.0004	0.0004	0.0210
	γ	-1.0001	-0.0001	0.0032
CMR-CF	α	0.9975	-0.0025	0.0891
	β	1.0068	0.0068	0.1204
	γ	-1.0021	-0.0021	0.0304

8.2 Monte Carlos Based on Non-Parametric Models

Next we conduct two small-scale simulation studies where we estimate the structural function $f(x)$ nonparametrically. Design [A] has a first stage selection equation that satisfies the NPV-CF restriction (4) and Design [B] does not.

For the first specification we follow the setup from Newey and Powell (2003) given as

$$\begin{aligned} y_i &= f(x_i) + \varepsilon_i = \ln(|x_i - 1| + 1)\text{sgn}(x_i - 1) + \varepsilon_i \\ [A]\ x_i &= z_i + \eta_i \end{aligned}$$

where the errors ε_i and η_i and instruments z_i are generated by

$$\begin{pmatrix} \varepsilon_i \\ \eta_i \\ z_i \end{pmatrix} \sim \text{i.i.d } N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho & 0 \\ \rho & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \right)$$

with $\rho = 0.5$. This design satisfies the restriction (4) with $v_i = x_i - E[x_i|z_i]$ because $v_i = \eta_i$.

In the second specification we use the same outcome equation but change the first stage equation to

$$[B]\ x_i = z_i + \eta_i / \exp(|z_i|)$$

and we use $\rho = 0.5$ and $\rho = 0.9$ varying the degree of endogeneity. The restriction (4) is violated because $v_i = x_i - E[x_i|z_i] = \eta_i / \exp(|z_i|)$.

Following Newey and Powell (2003) we use the Hermite series approximation of $f(x)$ as

$$f(x) \approx x\beta + \sum_{j=1}^{J} \gamma_j \exp(-x^2) x^{j-1}. \tag{22}$$

We estimate $f(x)$ using the nonparametric least squares (NPLS), the NPV-CF estimator and our CMR-CF estimator. We fix $J = 5$ for design [A] and $J = 7$ for design [B] and we use four different sample sizes (n=100, 400, 1000, and 2,000). In all of the designs we obtain the control using the first stage regression residual $\hat{v}_i = x_i - (\hat{\pi}_0 + \hat{\pi}_1 z_i + \hat{\pi}_2 z_i^2)$. We experimented with adding several higher order terms in the first stage and found very similar simulation results across all three estimators. We also experimented with different choices of approximating functions of $h(v)$ and $h(z, v)$ for design [B]. We consider $h(v) \approx \sum_{l=1}^{4} a_l v^l$ (NPV-CF1) or $\sum_{l=1}^{5} a_l v^l$ (NPV-CF2) for NPV-CF estimators and $h(z, v) = a_1 v + a_2 zv$ (CMR-CF1), $a_1 v + a_2 zv + a_3 z^2 v$ (CMR-CF2), or $a_1 v + a_2 zv + a_3 z^2 v + a_4 \tilde{v}_2$ (CMR-CF3) for CMR-CF estimators.

The results are summarized in Tables A and B. We report the root mean-squared-error (RMSE) averaged across the 500 replications and the realized values of x. In both designs RMSE decreases as the sample size increases for all estimators. The RMSEs for nonparametric least squares (NPLS) are larger than RMSEs for the estimators that correct for endogeneity. In the design [A] as expected both NPV-CF and CMR-CF estimators perform similarly although the CMR-CF estimator shows

slightly larger RMSEs because it adds an irrelevant correction term (zv) in the control function. In the design [B] the CMR-CF estimators dominate the NPV-CF estimators in terms of RMSE. CMR-CF2 is our preferred specification that shows the smallest RMSE among CMR-CF estimators. Comparison with CMR-CF1 suggests that including the term z^2v significantly reduces RMSE's. The RMSE's of NPLS and NPV-CF estimators tend to increase as the degree of endogeneity increases while RMSEs decrease in CMR-CF2 and CMR-CF3 as the degree of endogeneity increases. We conclude that our proposed CMR-CF estimator is robust to violations of the restriction (4) while the NPV-CF estimator is not.

Table A: Design [A], RMSE

	NPLS	NPV-CF	CMR-CF
Control Functions	None	$h(v) \approx a_1\hat{v}$	$h(z,v) \approx a_1\hat{v} + a_2 z\hat{v}$
n=100	0.4121	0.2685	0.2732
n=400	0.3844	0.1667	0.1692
n=1000	0.3788	0.1308	0.1317
n=2000	0.3695	0.1149	0.1165

Table B: Design [B], RMSE

	NPLS	NPV-CF1	NPV-CF2	CMR-CF1	CMR-CF2	CMR-CF3
CF's	None	$\sum_{l=1}^{4} a_l\hat{v}^l$	$\sum_{l=1}^{5} a_l\hat{v}^l$	$a_1\hat{v} + a_2 z\hat{v}$	$a_1\hat{v} + a_2 z\hat{v} + a_3 z^2\hat{v}$	$a_1\hat{v} + a_2 z\hat{v} + a_3 z^2\hat{v} + a_4\tilde{v}_2$
	$\rho = 0.5$					
n=100	0.3896	0.3233	0.3241	0.3049	0.3104	0.3231
n=400	0.2775	0.1679	0.1671	0.1540	0.1422	0.1456
n=1000	0.2511	0.1364	0.1358	0.1190	0.0999	0.1014
n=2000	0.2440	0.1150	0.1156	0.0968	0.0737	0.0745
	$\rho = 0.9$					
n=100	0.5277	0.3059	0.3018	0.2833	0.2734	0.2889
n=400	0.4462	0.2042	0.2003	0.1680	0.1296	0.1322
n=1000	0.4375	0.1885	0.1865	0.1483	0.0941	0.0950
n=2000	0.4311	0.1762	0.1773	0.1356	0.0690	0.0696

9 Empirical Example

We apply our estimator to the empirical example in Newey, Powell, and Vella (1999, hereafter NPV). They investigate the relationship between hourly wage rate and annual hours worked using the 1989 wave of the Michigan Panel Survey of Income Dynamics (PSID). The model is as follows:

$$y_i = z_{1i}'\beta + g_{20}(x_i) + \varepsilon_i \qquad x_i = z_i'\gamma + v_i$$

where y_i is the log of the hourly wage rate of individual i, z_{1i} is a vector of individual characteristics,

x_i is annual hours worked, z_i is a vector of exogenous variables that includes z_{1i} and g_{20} is a non-parametric, unknown function. ε_i and v_i are mean zero error terms such that $E[\varepsilon|z,v] \neq 0$, $E[v|z] = 0$, and $E[\varepsilon|v] \neq 0$. We mimic NPV's specification by defining z_{1i} as education dummies, union status, tenure status, full time work experience, regional variables and a racial dummy. The variables included in z_i but excluded from z_{1i} are marital status, health status, presence of young children, nonlabor income and a rural dummy. After following the data cleaning from NPV we have 2545 observations on males aged between 22 and 55 who had worked between 1000 and 3500 hours in the previous year.

To construct our estimator we rewrite the model as follows:

$$y_i = z_{1i}'\beta + g_{20}(x_i) + h(z_i,v_i) + \eta_i \qquad x_i = z_i'\gamma + v_i$$

with $E[\varepsilon_i|z_i,v_i] = h(z_i,v_i)$ and $E[\eta_i|z_i,v_i] = 0$. In the first step we mimic NPV's control $\hat{v}_i = x_i - \hat{E}[x_i|z_i]$ exactly by regressing hours worked on the z_i they use including the higher order terms for the regressors related to tenure and full-time experience. In the second step, for NPV we approximate $h(z_i,\hat{v}_i) = h(\hat{v}_i)$ as a third-order polynomial in $\hat{v}_i$ and for our estimator we approximate $h(z_i,\hat{v}_i)$ as a demeaned (w.r.t. $\hat{E}[\cdot|z_i]$) third-order polynomial in $\hat{v}_i$ and with interaction terms between z_{1i} and the demeaned third-order polynomial in $\hat{v}_i$ (45 interaction terms in total). We approximate $g_{20}(x_i)$ as a fourth-order polynomial in x_i.[17] We then regress y_i on our approximations to $f_0(x_i,z_{1i}) = z_{1i}'\beta + g_{20}(x_i)$ and $h(z_i,\hat{v}_i)$.

Table VII reports the results. Columns 1-3 mimic the specifications from NPV (estimates for all regressors except the control and hours worked are suppressed). Column 1 is OLS and shows that hours worked is not significant without the control, suggesting endogeneity problem. Column 2 is 2SLS and suggests that the linear model is misspecified when the higher order terms of hours worked are omitted. Column 3 is NPV's estimator using their preferred specification. Column 4 is our CMR-CF estimator. For our estimator there are 45 additional terms that we add to the control function relative to NPV and an F-test of their significance rejects the null with a p-value of less than 0.01 as the adjusted R-squared increases to 0.456 from 0.439. Thus it appears that the expected value of ε_i does depend on z_{1i} conditional on v_i. However, correcting for the additional terms does not appear to change the coefficients much as all of the NPV estimates fall within the confidence interval of their counterpart for the CMR-CF estimator.

10 Conclusion

We show that the CF estimator of Newey, Powell, and Vella (1999) can be modified to allow the conditional mean of the error to depend on both the instruments and controls. We do so by adding conditional moment restrictions which, when combined with a suitable rank condition, imply the control function is distinguishable from functions of the endogenous and exogenous regressors, yielding identification of the structural function. We also show our approach yields identification in settings where neither NPIV nor NPV-CF does. When sieves are used to approximate both the

[17] The choice of a fourth-order polynomial in x_i and a third-order polynomial in $\hat{v}_i$ is due to NPV, who state that this is the preferred specification according to the cross validation (CV) criterion.

structural function and the control function our estimator is simple to implement as it reduces to a series of Least Squares regressions. Our monte carlos are designed to mimic common economic settings where the NPV-CF independence assumption will not generally hold and we show our new estimator is consistent in these settings when the NPV-CF estimator is biased. Our empirical example revisits the example from NPV and we show under the CMRs we reject that the conditional mean of the error is fully independent of the instruments.

Table VII: Wage as a Function of Hours Worked (x)

	(1) OLS	(2) 2SLS	(3) NPV	(4) CMR-CF
	lnwage	lnwage	lnwage	lnwage
x	-1.51e-05	0.00102***	-0.00621**	-0.00603
	(1.97e-05)	(0.000112)	(0.00306)	(0.00389)
x^2			4.78e-06**	5.07e-06**
			(2.08e-06)	(2.44e-06)
x^3			-1.34e-09**	-1.54e-09**
			(6.33e-10)	(7.15e-10)
x^4			1.33e-13*	1.61e-13**
			(7.17e-14)	(8.00e-14)
v		-0.00106***	-0.00107***	0.00267
		(0.000114)	(0.000170)	(0.002554)
v^2			-1.32e-07	-1.43e-07
			(1.12e-07)	(1.83e-06)
v^3			1.81e-10*	-2.42e-09
			(1.10e-10)	(3.24e-09)
R-squared	0.416	0.435	0.444	0.471
Adj R-squared	0.412	0.432	0.439	0.456
F-test on Interactions	n/a	n/a	n/a	2.820
Prob > F	n/a	n/a	n/a	0.000

Standard errors in parentheses account for pre-stage estimations.

*** p<0.01, ** p<0.05, * p<0.1

x is the number of hours worked.

v is the residual from the first stage.

Appendix

A Linear setting with additive errors

We revisit the linear-in-parameters setting in light of our new estimator. In mean-deviated form the equation of interest as

$$y_i \quad = \quad x_i\beta_0 + \varepsilon_i, \tag{23}$$

with y_i the dependent variable and x_i a scalar explanatory variable that is potentially correlated with ε_i, so $E[x_i\varepsilon_i] \neq 0$ (and $E[y_i] = 0$ and $E[x_i] = 0$). We let z_i denote an instrument vector satisfying

$$E[z_i\,\varepsilon_i] = 0, \quad E[z_i\,x_i] \neq 0, \tag{24}$$

(and $E[z_i] = 0$). The classic control function (CCF) estimator for the linear case posits the same control $v_i = x_i - E[x_i|z_i]$ and follows directly from assuming

$$(\textbf{Classic CF Restriction})\ E[\varepsilon_i|z_i, x_i] = E[\varepsilon_i|z_i, v_i] = \rho v_i. \tag{25}$$

(25) implies the following estimating equation

$$y_i = x_i\beta + \rho v_i + \eta_i, \tag{26}$$

with $\eta_i = \varepsilon_i - E[\varepsilon_i|z_i, x_i]$. The CCF estimator differs from our CMR-CF estimator which considers an (unrestricted) general specification for the conditional expectation of the error

$$E[\varepsilon_i|z_i, v_i] \equiv h(z_i, v_i) = \tilde{\rho} v_i + \tilde{h}(z_i, v_i),$$

with the function characterizing $E[\varepsilon_i|z_i, v_i]$ having a leading term in v_i and a remaining term denoted by the function $\tilde{h}(z_i, v_i)$.

It is well-known by projection theory that the CCF estimator and the two-stage least squares (2SLS) estimator produce numerically identical estimates. Our CMR-CF estimator will differ in finite samples from CCF/2SLS estimator because of the additional regressors used to approximate $\tilde{h}(z_i, v_i)$. In this section we show that under the CMR condition the asymptotic correlation between x_i and $\tilde{h}(z_i, v_i)$ conditional on v_i vanishes so all three estimators for β_0 are asymptotically equivalent.

In the linear case "CMR-CF" is a misnomer because - like CCF and 2SLS - we only require $E[z_i\varepsilon_i] = 0$. We relegate the proof that CCF and 2SLS are numerically equivalent to a footnote below.[18]

[18] We provide a simple proof using projection that shows why they are equivalent. Let $Y = (y_1, \ldots, y_n)'$, $X = (x_1, \ldots, x_n)'$, $Z = (z_1, \ldots, z_n)'$, $\hat{X} = (\hat{x}_1, \ldots, \hat{x}_n)'$ and $\hat{V} = (\hat{v}_1, \ldots, \hat{v}_n)'$ below.

We then consider the generalization of equation (26), with

$$Y = X\beta_0 + \tilde{\rho}\hat{V} + \tilde{H}(Z,\hat{V}) + \hat{\eta} \tag{27}$$

where $Y = (y_1,\ldots,y_n)'$, $X = (x_1,\ldots,x_n)'$, $Z = (z_1,\ldots,z_n)'$, $\tilde{H}(Z,\hat{V}) = (\tilde{h}(z_1,\hat{v}_1),\ldots,\tilde{h}(z_n,\hat{v}_n))'$ and $\hat{V} = X - Z\hat{\pi}$, the vector of controls, which are the ordinary least squares residuals from the regression of X on Z, and $\hat{\eta}$ is the residual after the estimated vector of controls are included. We can rewrite (27) as

$$Y = (Z\hat{\pi} + \hat{V})\beta_0 + \tilde{\rho}\hat{V} + \tilde{H}(Z,\hat{V}) + \hat{\eta}. \tag{28}$$

Define $M_{\hat{V}} = I - \hat{V}(\hat{V}'\hat{V})^{-1}\hat{V}'$, the matrix that projects off of $\hat{V}$, and note that $M_{\hat{V}}\hat{V} = 0$ and $M_{\hat{V}}Z = Z$ (because $\hat{V}'Z = 0$). Then by partitioned regression theory, estimation of β_0 in (28) is numerically equivalent to the estimation of

$$\begin{aligned} Y &= M_{\hat{V}}(Z\hat{\pi} + \hat{V})\beta_0 + M_{\hat{V}}\tilde{H}(Z,\hat{V}) + M_{\hat{V}}\hat{\eta} \\ &= Z\hat{\pi}\beta + M_{\hat{V}}\tilde{H}(Z,\hat{V}) + M_{\hat{V}}\hat{\eta}. \end{aligned}$$

If $M_{\hat{V}}\tilde{H}(Z,\hat{V})$ is asymptotically uncorrelated with $Z\hat{\pi}$, i.e., if $Z'M_{\hat{V}}\tilde{H}(Z,\hat{V})/n$ converges to zero as the sample size goes to infinity, then the least squares estimator of β_0 in (27) is consistent whether or not we include $\tilde{H}(Z,\hat{V})$ in the regression as long as $\hat{V}$ is included as a regressor in (27). Note that because $Z'\hat{V} = 0$ we have

$$Z'M_{\hat{V}}\tilde{H}(Z,\hat{V})/n \;=\; Z'(I - \hat{V}(\hat{V}'\hat{V})^{-1}\hat{V}')\tilde{H}(Z,\hat{V})/n = Z'\tilde{H}(Z,\hat{V})/n = \sum\nolimits_{i=1}^{n} z_i\tilde{h}(z_i,\hat{v}_i)/n$$

and therefore consistency holds if

$$\sum\nolimits_{i=1}^{n} z_i\tilde{h}(z_i,\hat{v}_i)/n \to_p 0. \tag{29}$$

Theorem 6 couples (24) with weak regularity conditions which are sufficient for (29) to hold.

Theorem 6. *Assume (i) $E[\|z_i\| \cdot ||\tilde{h}(z_i,v_i)||] < \infty$, (ii) $\tilde{h}(z,v)$ is differentiable with respect to v, (iii) for $v_i(\pi) \equiv x_i - z_i'\pi$, assume $\sup_{\pi^*\in\Pi_0} E[\|z_i\|^2 \left\|\frac{\partial\tilde{h}(z_i,v_i(\pi^*))}{\partial v_i}\right\|] < \infty$ for Π_0 some neighborhood of π_0, (iv) assume $E[\|z_i\|^2 \left\|\frac{\partial\tilde{h}(z_i,v_i(\pi))}{\partial v_i}\right\|]$ is continuous at $\pi = \pi_0$, and (v) $\hat{\pi} \to_p \pi_0$. If (24) holds then (29) holds.*

Proof. Write $\sum_{i=1}^n z_i\tilde{h}(z_i,\hat{v}_i)/n = \sum_{i=1}^n z_i\tilde{h}(z_i,v_i)/n + \sum_{i=1}^n z_i(\tilde{h}(z_i,\hat{v}_i) - \tilde{h}(z_i,v_i))/n$. We have

Remark 1 (2SLS-CF Numerical Equivalence)**.** *If $\hat{\beta}_{2SLS} = (\hat{X}'\hat{X})^{-1}\hat{X}'Y$ and $(\hat{\beta}_{CF},\hat{\rho}_{CF})' = ((X,\hat{V})'(X,\hat{V}))^{-1}(X,\hat{V})'Y$ are well-defined and exist, then $\hat{\beta}_{2SLS} = \hat{\beta}_{CF}$.*

Proof. From projection theory the same numerical estimate obtains for the coefficient on x_i from either regressing Y on $(X,\hat{V})$ or regressing Yon the projection of X off of $\hat{V}$. Numerical equivalence follows because the projection of X off of $\hat{V}$ is equal to $\hat{X}$ because $(I - \hat{V}(\hat{V}'\hat{V})^{-1}\hat{V}')X = (I - \hat{V}(\hat{V}'\hat{V})^{-1}\hat{V}')(\hat{X} + \hat{V}) = \hat{X}$, as $\hat{V}'\hat{X} = 0$ by projection. □

$\sum_{i=1}^n z_i\tilde{h}(z_i,v_i)/n \to_p E[z_i\tilde{h}(z_i,v_i)]$ by the law of large numbers under (i). Obtain $||\sum_{i=1}^n z_i(\tilde{h}(z_i,\hat{v}_i)-\tilde{h}(z_i,v_i))/n|| \leq \|\hat{\pi}^*-\pi_0\| \sum_{i=1}^n \|z_i\|^2 \, ||\frac{\partial\tilde{h}(z_i,v_i(\hat{\pi}^*))}{\partial v_i}||/n$ by applying the mean-value expansion, where $\hat{\pi}^*$ lies between $\hat{\pi}$ and π_0 and $v_i(\pi) = x_i - z_i'\pi$. Then the term $\sum_{i=1}^n z_i(\tilde{h}(z_i,\hat{v}_i)-\tilde{h}(z_i,v_i))/n \to_p 0$ by the consistency of $\hat{\pi}$ and $\sum_{i=1}^n \|z_i\|^2 \left\|\frac{\partial\tilde{h}(z_i,v_i(\hat{\pi}^*))}{\partial v_i}\right\|/n \to_p E[\|z_i\|^2 \left\|\frac{\partial\tilde{h}(z_i,v_i(\pi_0))}{\partial v_i}\right\|] < \infty$ under (iii) and (iv). Therefore $\sum_{i=1}^n z_i\tilde{h}(z_i,\hat{v}_i)/n \to_p E[z_i\tilde{h}(z_i,v_i)]$. Finally note $E[z_i\tilde{h}(z_i,v_i)] = 0$ if (24) holds and v_i is such that $E[v_i|z_i] = 0$ because by the law of iterated expectations

$$\begin{aligned} 0 &= E[z_i\,\varepsilon_i] = E[z_i\,E[E[\varepsilon_i|z_i,v_i]|z_i]\,] = E[z_i(\tilde{\rho}E[v_i|z_i] + E[\tilde{h}(z_i,v_i)|z_i])] \\ &= E[z_i E[\tilde{h}(z_i,v_i)|z_i]] = E[z_i\,\tilde{h}(z_i,v_i)] \end{aligned}$$

and therefore the conclusion follows. □

The theorem also makes it clear that the classic control function approach does *not* generally yield consistency for the expected value of the error conditional on the control and exogenous variables unless $\tilde{H}(Z,\hat{V})$ is also included in the regression equation. Although this is not typically the object of interest of either the classic CF estimator or the 2SLS estimator, an exception is when one tests for endogeneity based on the estimate of ρ in (26).[19]

A simple example is illustrative of these points. Consider the case that z_i is a scalar and

$$E[\varepsilon_i|z_i,v_i] = \rho_1 v_i + \rho_2 v_i z_i, \tag{30}$$

but the researcher only includes x_i and $\hat{v}_i$ as regressors. Even though the researcher omits the *relevant* variable $\hat{v}_i z_i$, the ordinary least squares estimator $\hat{\beta}$ is consistent for β_0 because for the term corresponding to (29)

$$\sum_{i=1}^n z_i(\hat{v}_i z_i)/n \to_p E[v_i z_i^2] = 0,$$

which follows from $\hat{v}_i \to_p v_i$ (because $\hat{\pi} \to_p \pi_0$) and $E[v_i|z_i] = 0$ and by LLN under standard regularity conditions ($E[\|v_i\|\,\|z_i\|^2] < \infty$ and $E[\|z_i\|^3] < \infty$). However, $\hat{\rho}_1\hat{v}_i$ is not a consistent estimator of $E[\varepsilon_i|z_i,v_i]$. If one desired a consistent estimator of this conditional expectation, then $\hat{v}_i z_i$ would have to be included in the regression, and $\hat{\rho}_1\hat{v}_i + \hat{\rho}_2\hat{v}_i z_i$ would be consistent for $E[\varepsilon_i|z_i,v_i]$.

B Proof of convergence rates (Theorem 3)

We first introduce notation and prove Lemma L1 below that is useful to prove the convergence rate results and also the asymptotic normality of linear functional estimators.

[19] See e.g. Smith and Blundell (1986). In this case the misspecification of this conditional expectation may reduce the power of the test or call into question the test's consistency.

Define $h_L(z,v) = a_L'\tilde{\varphi}^L(z,v)$ and $\hat{h}_L(z,v) = a_L'\hat{\tilde{\varphi}}^L(z,v)$ where a_L[20] satisfies Assumption L1 (iv). Define $\psi_i^{\mathbf{L}}(z_i,v_i) = (\phi_1(x_i,z_{1i}),\ldots,\phi_K(x_i,z_{1i}),\tilde{\varphi}^L(z_i,v_i)')'$ where $\tilde{\varphi}^L(z_i,v_i) = (\tilde{\varphi}_1(z_i,v_i),\ldots,\tilde{\varphi}_L(z_i,v_i))'$ and $\hat{\psi}_i^{\mathbf{L}}(z_i,v_i) = (\phi_1(x_i,z_{1i}),\ldots,\phi_K(x_i,z_{1i}),\hat{\tilde{\varphi}}^L(z_i,v_i)')'$ with $\hat{\tilde{\varphi}}^L(z_i,v_i) = (\hat{\tilde{\varphi}}_1(z_i,v_i),\ldots,\hat{\tilde{\varphi}}_L(z_i,v_i))'$. We further let $\hat{\tilde{\psi}}_i^{\mathbf{L}} = \hat{\psi}^{\mathbf{L}}(z_i,\hat{v}_i)$, $\psi_i^{\mathbf{L}} = \psi^{\mathbf{L}}(z_i,v_i)$, and $\hat{\psi}_i^{\mathbf{L}} = \hat{\psi}^{\mathbf{L}}(z_i,v_i)$. We further let $\psi^{\mathbf{L},n} = (\psi_1^{\mathbf{L}},\ldots,\psi_n^{\mathbf{L}})'$, $\hat{\psi}^{\mathbf{L},n} = (\hat{\psi}_1^{\mathbf{L}},\ldots,\hat{\psi}_n^{\mathbf{L}})'$, and $\hat{\tilde{\psi}}^{\mathbf{L},n} = (\hat{\tilde{\psi}}_1^{\mathbf{L}},\ldots,\hat{\tilde{\psi}}_n^{\mathbf{L}})'$.

Let C (also C_1,C_2, and others) denote a generic positive constant and let $C(Z,V)$ or $C(X,Z_1)$ (also $C_1(\cdot)$, $C_2(\cdot)$, and others) denote a generic bounded positive function of (Z,V) or (X,Z_1). We often write $C_i = C(x_i,z_{1i})$. Recall $\mathcal{W} = \text{supp}(Z,V)$.

Assumption 6 (L1). *(i) (X,Z,V) is continuously distributed with bounded density; (ii) for each k, K, L, and $\mathbf{L} = K+L$ there are nonsingular matrices B_1, B_2, B_3, and B such that for $p_{B_1}^k(z) = B_1p^k(z)$, $\phi_{B_2}^K(x,z_1) = B_2\phi^K(x,z_1)$, $\tilde{\varphi}_{B_3}^L(z,v) = B_3\tilde{\varphi}^L(z,v)$, and $\psi_B^{\mathbf{L}}(z,v) = B\psi^{\mathbf{L}}(z,v)$, $E[p_{B_1}^k(Z_i)p_{B_1}^k(Z_i)']$, $E[\phi_{B_2}^K(X_i,Z_{1i})\phi_{B_2}^K(X_i,Z_{1i})']$, $E[\tilde{\varphi}_{B_3}^L(Z_i,V_i)\tilde{\varphi}_{B_3}^L(Z_i,V_i)']$, and $E[\psi_B^{\mathbf{L}}(Z_i,V_i)\psi_B^{\mathbf{L}}(Z_i,V_i)']$ have smallest eigenvalues that are bounded away from zero, uniformly in k, K, L, and $\mathbf{L}$; (iii) for each integer $\delta > 0$, there exist $\zeta_\delta(K)$, $\zeta_\delta(\mathbf{L})$, and $\xi_\delta(k)$ such that $|\phi^K(x,z_1)|_\delta \leq \zeta_\delta(K)$, $|\psi^{\mathbf{L}}(z,v)|_\delta \leq \zeta_\delta(\mathbf{L})$ (this also implies that $|\tilde{\varphi}^L(z,v)|_\delta \leq \zeta_\delta(L)$), and $|p^k(z)|_\delta \leq \xi_\delta(k)$; (iv) There exist γ_1, γ_2, γ_f, $\gamma > 0$, and β^K, a_L, $\beta_{\mathbf{L}}$, λ_k^1, and $\lambda_{l,k}^2$ such that $|\Pi_0(z) - \lambda_k^{1\prime}p^k(z)|_\delta \leq Ck^{-\gamma_1}$, $|\bar{\varphi}_{0l}(z) - \lambda_{l,k}^{2\prime}p^k(z)|_\delta \leq Ck^{-\gamma_2}$ for all l, $|f_0(x,z_1) - \beta^{K\prime}\phi^K(x,z_1)|_\delta \leq CK^{-\gamma_f}$, $|h_0(z,v) - a_L'\tilde{\varphi}^L(z,v)|_\delta \leq CL^{-\gamma}$, and $|g_0(z,v) - \beta_{\mathbf{L}}'\psi^{\mathbf{L}}(z,v)|_\delta \leq C\mathbf{L}^{-\gamma}$; (v) both $\mathcal{Z}$ and $\mathcal{X}$ are compact.*

Let $\triangle_{n,1} = k_n^{1/2}/\sqrt{n} + k_n^{-\gamma_1}$ and $\triangle_{n,2} = k_n^{1/2}/\sqrt{n} + k_n^{-\gamma_2}$ and $\triangle_n = \max\{\triangle_{n,1},\triangle_{n,2}\}$.

Lemma 1 (L1). *Suppose Assumptions L1 and Assumptions C1 hold. Further suppose $\mathbf{L}^{1/2}(\zeta_1(L) + L^{1/2}\xi_0(k)\sqrt{k/n} + L^{1/2})\triangle_n \to 0$, $\xi_0(k)^2k/n \to 0$, and $\zeta_0(\mathbf{L})^2\mathbf{L}/n \to 0$. Then,*

$$\begin{aligned}
\left(\sum_{i=1}^n \left(\hat{g}(z_i,v_i) - g_0(z_i,v_i)\right)^2/n\right)^{1/2} &= O_p(\sqrt{\mathbf{L}/n} + L\xi_0(k)\triangle_{n,1}\sqrt{k/n} + L\triangle_{n,2} + \mathbf{L}^{-\gamma}) \\
\max_{i\leq n}|\hat{g}(z_i,v_i) - g_0(z_i,v_i)| &= O_p(\zeta_0(\mathbf{L})[\sqrt{\mathbf{L}/n} + L\xi_0(k)\triangle_{n,1}\sqrt{k/n} + L\triangle_{n,2} + \mathbf{L}^{-\gamma}]).
\end{aligned}$$

B.1 Proof of Lemma L1

Without loss of generality, we will let $p^k(z) = p_{B_1}^k(z)$, $\phi^K(x,z_1) = \phi_{B_2}^K(x,z_1)$, $\tilde{\varphi}^L(z,v) = \tilde{\varphi}_{B_3}^L(z,v)$, and $\psi^{\mathbf{L}}(z,v) = \psi_B^{\mathbf{L}}(z,v)$. Let $\hat{\Pi}_i = \hat{\Pi}(z_i)$ and $\Pi_i = \Pi_0(z_i)$. Let $\hat{\bar{\varphi}}_{li} = \hat{\bar{\varphi}}_l(z_i)$ and $\bar{\varphi}_{li} = \bar{\varphi}_l(z_i)$. Let $\hat{\hat{\bar{\varphi}}}_{li} = \hat{\bar{\varphi}}_l(z_i,\hat{v}_i)$ and $\tilde{\varphi}_{li} = \tilde{\varphi}_l(z_i,v_i)$. Also let $\hat{\hat{\tilde{\varphi}}}_i^L = \hat{\tilde{\varphi}}^L(z_i,\hat{v}_i)$ and $\tilde{\varphi}_i^L = \tilde{\varphi}^L(z_i,v_i)$. Further define $\dot{\bar{\varphi}}_l(z) = p^k(z)'(P'P)^-\sum_{i=1}^n p^k(z_i)\varphi_l(z_i,v_i)$ where we have $\hat{\bar{\varphi}}_l(z) = p^k(z)'(P'P)^-\sum_{i=1}^n p^k(z_i)\varphi_l(z_i,\hat{v}_i)$. Let $\dot{\tilde{\varphi}}^L(z) = (\dot{\bar{\varphi}}_1(z),\ldots,\dot{\bar{\varphi}}_L(z))'$ and $\bar{\varphi}^L(z) = (\bar{\varphi}_1(z),\ldots,\bar{\varphi}_L(z))'$. We also let
$\varphi^L(z_i,\hat{v}_i) = (\varphi_1(z_i,\hat{v}_i),\ldots,\varphi_L(z_i,\hat{v}_i))'$ and $\varphi^L(z_i,v_i) = (\varphi_1(z_i,v_i),\ldots,\varphi_L(z_i,v_i))'$.

First note $(P'P)/n$ becomes nonsingular w.p.a.1 as $\xi_0(k)^2k/n \to 0$ by Assumption L1 (ii) and by the same proof in Theorem 1 of Newey (1997). Then by the same proof (A.3) of Lemma A1 in

[20] With abuse of notation we write $a_L = (a_1,\ldots,a_L)'$.

Newey, Powell, and Vella (1999), we obtain

$$\sum_{i=1}^{n} ||\hat{\Pi}_i - \Pi_i||^2/n = O_p(\triangle_{n,1}^2) \text{ and } \sum_{i=1}^{n} ||\dot{\bar{\varphi}}_{li} - \bar{\varphi}_{li}||^2/n = O_p(\triangle_{n,2}^2) \text{ for all } l. \tag{31}$$

Also by Theorem 1 of Newey (1997), it follows that

$$\begin{aligned} \max_{i\leq n} ||\hat{\Pi}_i - \Pi_i|| &= O_p(\xi_0(k)\triangle_{n,1}) & (32)\\ \max_{i\leq n} ||\dot{\bar{\varphi}}_{li} - \bar{\varphi}_{li}|| &= O_p(\xi_0(k)\triangle_{n,2}) \text{ for all } l. & (33) \end{aligned}$$

Define $\hat{\mathcal{T}} = (\hat{\bar{\psi}}^{\mathbf{L},n})'\hat{\bar{\psi}}^{\mathbf{L},n}/n$ and $\dot{\mathcal{T}} = (\psi^{\mathbf{L},n})'\psi^{\mathbf{L},n}/n$. Our goal is to show that $\hat{\mathcal{T}}$ is nonsingular w.p.a.1. We first note that $\dot{\mathcal{T}}$ is nonsingular w.p.a.1 by Assumption L1 (ii) as $\zeta_0(\mathbf{L})^2\mathbf{L}/n \to 0$ by the same proof in Lemma A1 of Newey, Powell, and Vella (1999).

For ease of notation along the proof, we will assume some rate conditions are satisfied. Then we collect those rate conditions in Section B.2 and derive conditions under which all of them are satisfied. Next note that

$$\begin{aligned} \left\|\hat{\bar{\varphi}}_i^L - \tilde{\varphi}_i^L\right\| &\leq \left\|\varphi^L(z_i,\hat{v}_i) - \varphi^L(z_i,v_i)\right\| + \left\|\hat{\bar{\varphi}}^L(z_i) - \bar{\varphi}^L(z_i)\right\| & (34)\\ &\leq \left\|\varphi^L(z_i,\hat{v}_i) - \varphi^L(z_i,v_i)\right\| + \left\|\hat{\bar{\varphi}}^L(z_i) - \dot{\bar{\varphi}}^L(z_i)\right\| + \left\|\dot{\bar{\varphi}}^L(z_i) - \bar{\varphi}^L(z_i)\right\|. \end{aligned}$$

We find $\left\|\varphi^L(z_i,\hat{v}_i) - \varphi^L(z_i,v_i)\right\| \leq C\zeta_1(L)||\hat{\Pi}_i - \Pi_i||$ applying a mean value expansion because $\varphi_l(z_i,v_i)$ is Lipschitz in v_i (so in Π_i) for all l (Assumption C1 (vi)). Combined with (31), it implies that

$$\sum_{i=1}^{n} \left\|\varphi^L(z_i,\hat{v}_i) - \varphi^L(z_i,v_i)\right\|^2/n = O_p(\zeta_1(L)^2\triangle_{n,1}^2). \tag{35}$$

Next let $\hat{\omega}_l = (\varphi_l(z_1,\hat{v}_1) - \varphi_l(z_1,v_1),\ldots,\varphi_l(z_n,\hat{v}_n) - \varphi_l(z_n,v_n))'$. Then we can write for any $l = 1,\ldots,L$,

$$\begin{aligned} \sum_{i=1}^{n} \left\|\hat{\bar{\varphi}}_l(z_i) - \dot{\bar{\varphi}}_l(z_i)\right\|^2/n &= \text{tr}\left\{\sum_{i=1}^{n} p^k(z_i)'(P'P)^- P'\hat{\omega}_l\hat{\omega}_l'P(P'P)^- p^k(z_i)\right\}/n & (36)\\ &= \text{tr}\left\{(P'P)^- P'\hat{\omega}_l\hat{\omega}_l'P(P'P)^- \sum_{i=1}^{n} p^k(z_i)p^k(z_i)'\right\}/n\\ &= \text{tr}\left\{(P'P)^- P'\hat{\omega}_l\hat{\omega}_l'P\right\}/n\\ &\leq C\max_{i\leq n}||\hat{\Pi}_i - \Pi_i||^2\text{tr}\left\{(P'P)^- P'P\right\}/n \leq C\xi_0(k)^2\triangle_{n,1}^2 k/n \end{aligned}$$

where the first inequality is obtained by (32) and applying a mean value expansion to $\varphi_l(z_i,v_i)$ which is Lipschitz in v_i (so in Π_i) for all l (Assumption C1 (vi)). From (31), (34), (35), and (36), we conclude

$$\begin{aligned} \sum_{i=1}^{n} ||\hat{\bar{\varphi}}^L(z_i) - \bar{\varphi}^L(z_i)||^2/n &= O_p(L\xi_0(k)^2\triangle_{n,1}^2 k/n) + O_p(L\triangle_{n,2}^2) = o_p(1), & (37)\\ \sum_{i=1}^{n} \left\|\hat{\bar{\varphi}}_i^L - \tilde{\varphi}_i^L\right\|^2/n &= O_p(\zeta_1(L)^2\triangle_{n,1}^2) + O_p(L\xi_0(k)^2\triangle_{n,1}^2 k/n) + O_p(L\triangle_{n,2}^2) = o_p(1). \end{aligned}$$

The latter also implies that by the triangle inequality and the Markov inequality,

$$\sum_{i=1}^{n} ||\hat{\tilde{\varphi}}_i^L||^2/n \leq 2\sum_{i=1}^{n} ||\hat{\tilde{\varphi}}_i^L - \tilde{\varphi}_i^L||^2/n + 2\sum_{i=1}^{n} ||\tilde{\varphi}_i^L||^2/n = o_p(1) + O_p(L). \tag{38}$$

Let $\triangle_n^{\varphi} = (\zeta_1(L) + L^{1/2}\xi_0(k)\sqrt{k/n} + L^{1/2})\triangle_n$. It also follows that

$$\sum_{i=1}^{n} \left\| \hat{\tilde{\psi}}_i^{\mathbf{L}} - \psi_i^{\mathbf{L}} \right\|^2 /n \leq \sum_{i=1}^{n} \left\| \hat{\tilde{\varphi}}_i^L - \tilde{\varphi}_i^L \right\|^2 /n = O_p((\triangle_n^{\varphi})^2) = o_p(1). \tag{39}$$

This also implies $\sum_{i=1}^{n} ||\hat{\tilde{\psi}}_i^{\mathbf{L}}||^2/n = O_p(\mathbf{L})$ because $\sum_{i=1}^{n} \left\| \hat{\tilde{\psi}}_i^{\mathbf{L}} \right\|^2 /n \leq 2\sum_{i=1}^{n} \left\| \hat{\tilde{\psi}}_i^{\mathbf{L}} - \psi_i^{\mathbf{L}} \right\|^2 /n + 2\sum_{i=1}^{n} \left\| \psi_i^{\mathbf{L}} \right\|^2 /n = O_p(\mathbf{L})$.

Then applying (39) and applying the triangle inequality and Cauchy-Schwarz inequality and by Assumption L1 (iii) , we obtain

$$\begin{aligned} ||\hat{\mathcal{T}} - \dot{\mathcal{T}}|| &\leq \sum_{i=1}^{n} \left\| \hat{\tilde{\psi}}_i^{\mathbf{L}} - \psi_i^{\mathbf{L}} \right\|^2 /n + 2\sum_{i=1}^{n} \left\| \psi_i^{\mathbf{L}} \right\| \left\| \hat{\tilde{\psi}}_i^{\mathbf{L}} - \psi_i^{\mathbf{L}} \right\| /n \\ &\leq O_p((\triangle_n^{\varphi})^2) + 2\left(\sum_{i=1}^{n} \left\| \psi_i^{\mathbf{L}} \right\|^2 /n \right)^{1/2} \left(\sum_{i=1}^{n} \left\| \hat{\tilde{\psi}}_i^{\mathbf{L}} - \psi_i^{\mathbf{L}} \right\|^2 /n \right)^{1/2} \\ &= O_p((\triangle_n^{\varphi})^2) + O_p(\mathbf{L}^{1/2}\triangle_n^{\varphi}) = o_p(1). \end{aligned} \tag{40}$$

It follows that

$$\begin{aligned} ||\hat{\mathcal{T}} - \mathcal{T}|| &\leq ||\hat{\mathcal{T}} - \dot{\mathcal{T}}|| + ||\dot{\mathcal{T}} - \mathcal{T}|| \\ &= O_p((\triangle_n^{\varphi})^2 + \mathbf{L}^{1/2}\triangle_n^{\varphi} + \zeta_0(\mathbf{L})\sqrt{\mathbf{L}/n}) \equiv O_p(\triangle_{\mathcal{T}}) = o_p(1) \end{aligned} \tag{41}$$

where we obtain $||\dot{\mathcal{T}} - \mathcal{T}|| = O_p(\zeta_0(\mathbf{L})\sqrt{\mathbf{L}/n})$ by the same proof in Lemma A1 of Newey, Powell, and Vella (1999). Therefore we conclude $\hat{\mathcal{T}}$ is also nonsingular w.p.a.1. The same conclusion holds even when instead we take $\hat{\mathcal{T}} = \sum_{i=1}^{n} C(z_i, v_i)\hat{\tilde{\psi}}_i^{\mathbf{L}}\hat{\tilde{\psi}}_i^{\mathbf{L}\prime}/n$ and $\dot{\mathcal{T}} = \sum_{i=1}^{n} C(z_i, v_i)\psi_i^{\mathbf{L}}\psi_i^{\mathbf{L}\prime}/n$ for some positive bounded function $C(z_i, v_i)$ and this helps to derive the consistency of the heteroskedasticity robust variance estimator later.

Let $\eta_i = y_i - g_0(z_i, v_i)$ and let $\eta = (\eta_1, \ldots, \eta_n)'$. Let $(\mathbf{Z}, \mathbf{V}) = ((Z_1, V_1), \ldots, (Z_n, V_n))$. Then we have $E[\eta_i|\mathbf{Z}, \mathbf{V}] = 0$ and by the independence assumption of the observations, we have $E[\eta_i\eta_j|\mathbf{Z}, \mathbf{V}] = 0$ for $i \neq j$. We also have $E[\eta_i^2|\mathbf{Z}, \mathbf{V}] < \infty$. Then by (39) and the triangle inequality, we bound

$$\begin{aligned} E\big[||(\hat{\tilde{\psi}}^{\mathbf{L},n} - \psi^{\mathbf{L},n})'\eta/n||^2|\mathbf{Z}, \mathbf{V}\big] &\leq Cn^{-2}\sum_{i=1}^{n} E[\eta_i^2|\mathbf{Z}, \mathbf{V}] \left\| \hat{\tilde{\psi}}_i^{\mathbf{L}} - \psi_i^{\mathbf{L}} \right\|^2 \\ &\leq n^{-1}O_p(L(\triangle_n^{\varphi})^2) = o_p(n^{-1}). \end{aligned}$$

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 $||(\hat{\tilde{\psi}}^{\mathbf{L},n} - \psi^{\mathbf{L},n})'\eta/n||^2 = o_p(n^{-1})$. Also note that $E[\left\|(\psi^{\mathbf{L},n})'\eta/n\right\|^2] = C\mathbf{L}/n$ (see proof of Lemma A1 in

Newey, Powell, and Vella (1999)). Therefore, by the triangle inequality

$$\begin{aligned} ||(\hat{\hat{\psi}}^{\mathbf{L},n})'\eta/n||^2 &\leq 2||(\hat{\hat{\psi}}^{\mathbf{L},n}-\psi^{\mathbf{L},n})'\eta/n||^2+2||(\psi^{\mathbf{L},n})'\eta/n||^2. \\ &= o_p(1)+O_p(\mathbf{L}/n)=O_p(\mathbf{L}/n). \end{aligned} \tag{42}$$

Define $\hat{g}_i = \hat{f}(x_i,z_{1i})+\hat{h}(z_i,\hat{v}_i)$, $\hat{\tilde{g}}_{\mathbf{L}i} = f_K(x_i,z_{1i})+\hat{h}_L(z_i,\hat{v}_i)$, $\tilde{g}_{\mathbf{L}i} = f_K(x_i,z_{1i})+h_L(z_i,\hat{v}_i)$, $\tilde{g}_{0i}=f_0(x_i,z_{1i})+h_0(z_i,\hat{v}_i)$, and $g_{0i}=f_0(x_i,z_{1i})+h_0(z_i,v_i)$ where $f_K(x_i,z_{1i})=\sum_{l=1}^K \beta_l\phi_l(x_i,z_{1i})$, $\hat{h}(z_i,\hat{v}_i)=\hat{a}_L'\hat{\tilde{\varphi}}(z_i,\hat{v}_i)$, $\hat{h}_L(z_i,\hat{v}_i)=a_L'\hat{\tilde{\varphi}}(z_i,\hat{v}_i)$, and $h_L(z_i,\hat{v}_i)=a_L'(\varphi(z_i,\hat{v}_i)-\bar{\varphi}^L(z_i))$ and let $\hat{g},\hat{\tilde{g}}_{\mathbf{L}}$, $\tilde{g}_{\mathbf{L}}$, and $\tilde{g}_0$ stack the n observations of $\hat{g}_i,\hat{\tilde{g}}_{\mathbf{L}i}$, $\tilde{g}_{\mathbf{L}i}$, and $\tilde{g}_{0i}$, respectively. Recall $\beta_{\mathbf{L}}=(\beta^{K\prime},a_L')'$ where $\beta^K=(\beta_1,\ldots,\beta_K)'$ and let this $\beta_{\mathbf{L}}$ satisfies Assumption L1 (iv). From the first order condition of the last step least squares we obtain

$$\begin{aligned} 0 &= \hat{\hat{\psi}}^{\mathbf{L},n\prime}(y-\hat{g})/n \\ &= \hat{\hat{\psi}}^{\mathbf{L},n\prime}(\eta-(\hat{g}-\hat{\tilde{g}}_{\mathbf{L}})-(\hat{\tilde{g}}_{\mathbf{L}}-\tilde{g}_{\mathbf{L}})-(\tilde{g}_{\mathbf{L}}-\tilde{g}_0))/n \\ &= \hat{\hat{\psi}}^{\mathbf{L},n\prime}(\eta-\hat{\hat{\psi}}^{\mathbf{L},n}(\hat{\beta}-\beta_{\mathbf{L}})-(\hat{\tilde{g}}_{\mathbf{L}}-\widetilde{g}_{\mathbf{L}})-(\tilde{g}_{\mathbf{L}}-\tilde{g}_0)-(\tilde{g}_0-g_0))/n. \end{aligned} \tag{43}$$

Note that by $\hat{\hat{\psi}}^{\mathbf{L},n}(\hat{\hat{\psi}}^{\mathbf{L},n\prime}\hat{\hat{\psi}}^{\mathbf{L},n})^{-1}\hat{\hat{\psi}}^{\mathbf{L},n\prime}$ idempotent and by Assumption L1 (iv),

$$\begin{aligned} ||\hat{\mathcal{T}}^{-1}\hat{\hat{\psi}}^{\mathbf{L},n\prime}(\tilde{g}_{\mathbf{L}}-\tilde{g}_0)/n|| &\leq O_p(1)\{(\tilde{g}_{\mathbf{L}}-\tilde{g}_0)'\hat{\hat{\psi}}^{\mathbf{L},n}(\hat{\hat{\psi}}^{\mathbf{L},n\prime}\hat{\hat{\psi}}^{\mathbf{L},n})^{-1}\hat{\hat{\psi}}^{\mathbf{L},n\prime}(\tilde{g}_{\mathbf{L}}-\tilde{g}_0)/n\}^{1/2} \\ &\leq O_p(1)\{(\tilde{g}_{\mathbf{L}}-\tilde{g}_0)'(\tilde{g}_{\mathbf{L}}-\tilde{g}_0)/n\}^{1/2}=O_p(\mathbf{L}^{-\gamma}). \end{aligned} \tag{44}$$

Similarly we obtain by $\hat{\hat{\psi}}^{\mathbf{L},n}(\hat{\hat{\psi}}^{\mathbf{L},n\prime}\hat{\hat{\psi}}^{\mathbf{L},n})^{-1}\hat{\hat{\psi}}^{\mathbf{L},n\prime}$ idempotent, Assumption L1 (iv), and (37),

$$\begin{aligned} &||\hat{\mathcal{T}}^{-1}\hat{\hat{\psi}}^{\mathbf{L},n\prime}(\hat{\tilde{g}}_{\mathbf{L}}-\tilde{g}_{\mathbf{L}})/n|| = O_p(1)\{(\hat{\tilde{g}}_{\mathbf{L}}-\tilde{g}_{\mathbf{L}})'(\hat{\tilde{g}}_{\mathbf{L}}-\tilde{g}_{\mathbf{L}})/n\}^{1/2} \\ \leq\ & O_p(1)(\textstyle\sum_{i=1}^n ||\hat{h}_L(z_i,\hat{v}_i)-\tilde{h}_L(z_i,\hat{v}_i)||^2/n)^{1/2} \\ \leq\ & O_p(1)(\textstyle\sum_{i=1}^n ||a_L||^2||\hat{\tilde{\varphi}}^L(z_i)-\bar{\varphi}^L(z_i)||^2/n)^{1/2}=O_p(L\xi_0(k)\triangle_{n,1}\sqrt{k/n}+L\triangle_{n,2}). \end{aligned} \tag{45}$$

Similarly also by $\hat{\hat{\psi}}^{\mathbf{L},n}(\hat{\hat{\psi}}^{\mathbf{L},n\prime}\hat{\hat{\psi}}^{\mathbf{L},n})^{-1}\hat{\hat{\psi}}^{\mathbf{L},n\prime}$ idempotent and (31) and applying the mean value expansion to $h_0(z_i,v_i)$, we have

$$\begin{aligned} &||\hat{\mathcal{T}}^{-1}\hat{\hat{\psi}}^{\mathbf{L},n\prime}(\tilde{g}_0-g_0)/n|| = O_p(1)(\sum_{i=1}^n ||h_0(z_i,\hat{v}_i)-h_0(z_i,v_i)||^2/n)^{1/2} \\ \leq\ & O_p(1)(\textstyle\sum_{i=1}^n ||\hat{\Pi}_i-\Pi_i||^2/n)^{1/2}=O_p(\triangle_{n,1})=o_p(1). \end{aligned} \tag{46}$$

Combining (42), (43), (44), (45), (46) and by $\hat{\mathcal{T}}$ is nonsingular w.p.a.1, we obtain

$$\begin{aligned} ||\hat{\beta}-\beta_{\mathbf{L}}|| &\leq ||\hat{\mathcal{T}}^{-1}\hat{\hat{\psi}}^{\mathbf{L},n\prime}\eta/n||+||\hat{\mathcal{T}}^{-1}\hat{\hat{\psi}}^{\mathbf{L},n\prime}(\hat{\tilde{g}}_{\mathbf{L}}-\tilde{g}_{\mathbf{L}})/n||+||\hat{\mathcal{T}}^{-1}\hat{\hat{\psi}}^{\mathbf{L},n'}(\tilde{g}_{\mathbf{L}}-\tilde{g}_0)/n||+o_p(1) \\ &= O_p(1)\{\sqrt{\mathbf{L}/n}+L\xi_0(k)\triangle_{n,1}\sqrt{k/n}+L\triangle_{n,2}+\mathbf{L}^{-\gamma}\}\equiv O_p(\triangle_{n,\beta}). \end{aligned} \tag{47}$$

Define $g^*_{\mathbf{L}i}=f_K(x_i,z_{1i})+h^*_L(z_i,v_i)$ where $h^*_L(z_i,v_i)=a_L'(\varphi^L(z_i,v_i)-\hat{\tilde{\varphi}}^L(z_i))$. Then applying the

triangle inequality, by (37), (47), the Markov inequality, Assumption L1 (iv), and $\hat{\mathcal{T}}$ is nonsingular w.p.a.1 (by Assumption L1 (ii) and (41)), we conclude

$$\begin{aligned}
& \sum_{i=1}^{n} \left(\hat{g}(z_i, v_i) - g_0(z_i, v_i)\right)^2 / n \qquad (48)\\
\leq\ & 3\sum_{i=1}^{n} \left(\hat{g}(z_i, v_i) - g^*_{\mathbf{L}i}\right)^2 / n + 3\sum_{i=1}^{n} \left(g^*_{\mathbf{L}i} - g_{\mathbf{L}i}\right)^2 / n + 3\sum_{i=1}^{n} \left(g_{\mathbf{L}i} - g_0(z_i, v_i)\right)^2 / n \\
\leq\ & O_p(1)||\hat{\beta} - \beta_{\mathbf{L}}||^2 \\
& + C_1 \sum_{i=1}^{n} ||a_L||^2 ||\hat{\bar{\varphi}}^L(z_i) - \bar{\varphi}^L(z_i)||^2 / n + C_2 \sup_{\mathcal{W}} ||\beta'_{\mathbf{L}} \psi^{\mathbf{L}}(z, v) - g_0(z, v)||^2 \\
\leq\ & O_p(\triangle^2_{n,\beta}) + LO_p(L\xi_0(k)^2 \triangle^2_{n,1} k/n + L\triangle^2_{n,2}) + O_p(\mathbf{L}^{-2\gamma}) = O_p(\triangle^2_{n,\beta}).
\end{aligned}$$

This also implies that by a similar proof to Theorem 1 of Newey (1997)

$$\max_{i \leq n} |\hat{g}_i - g_{0i}| = O_p(\zeta_0(\mathbf{L})\triangle_{n,\beta}). \qquad (49)$$

B.2 Proof of Theorem 3

Under Assumption C1, all the assumptions in Assumption L1 are satisfied. We can take $\gamma_1 = s_1/d_z$ and $\gamma_2 = s_2/d_z$ as discussed in Assumption C1. For the consistency, we require the following rate conditions: (i) $\mathbf{L}^{1/2}\triangle^{\varphi}_n \to 0$ from (40), (ii) $\zeta_0(\mathbf{L})^2\mathbf{L}/n \to 0$ (such that $\dot{\mathcal{T}}$ is nonsingular w.p.a.1), and (iii) $\xi_0(k)^2 k/n \to 0$ (such that $P'P/n$ is nonsingular w.p.a.1). The other rate conditions are dominated by these three. From the definition of $\triangle^{\varphi}_n = (\zeta_1(L) + L^{1/2}\xi_0(k)\sqrt{k/n} + L^{1/2})\triangle_n$, we have (i) : $\mathbf{L}^{1/2}(\zeta_1(L) + L^{1/2}\xi_0(k)\sqrt{k/n} + L^{1/2})\triangle_n$.

For the polynomial approximations, we have $\zeta_\delta(L) \leq CL^{1+2\delta}$ and $\xi_0(k) \leq Ck$ and for the spline approximations, we have $\zeta_\delta(L) \leq CL^{0.5+\delta}$ and $\xi_0(k) \leq Ck^{0.5}$. Therefore for the polynomial approximations, the rate conditions become (i) $\mathbf{L}^{1/2}(L^3 + L^{1/2}k^{3/2}/\sqrt{n} + L^{1/2})\triangle_n \to 0$, (ii) $\mathbf{L}^3/n \to 0$, and (iii) $k^3/n \to 0$ and for the spline approximations, they become (i) $\mathbf{L}^{1/2}(L^{3/2} + L^{1/2}k/\sqrt{n} + L^{1/2})\triangle_n \to 0$, (ii) $\mathbf{L}^2/n \to 0$, and (iii) $k^2/n \to 0$. Also note that $\triangle_{n,\beta} \equiv ||\hat{\beta} - \beta_{\mathbf{L}}|| = \sqrt{\mathbf{L}/n} + L\xi_0(k)\triangle_{n,1}\sqrt{k/n} + L\triangle_{n,2} + \mathbf{L}^{-\gamma} = \sqrt{\mathbf{L}/n} + L\triangle_n + \mathbf{L}^{-\gamma}$ since $\xi_0(k)\sqrt{k/n} = o(1)$. Here we can take $\gamma_f = s/(d_x + d_1)$ and $\gamma = s/d$ because f_0 and h_0 belong to 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 3 (a) follows from Lemma L1 applying the dominated convergence theorem by $\hat{g}_i$ and g_{0i} are bounded.

For Theorem 3 (b) note that for $\hat{\beta} = (\hat{\beta}^{K\prime}, \hat{a}'_L)'$ and $\beta_{\mathbf{L}} = (\beta^{K\prime}, a'_L)'$,

$$\begin{aligned}
|\hat{f} - f_0|_\delta &\leq |\phi^K(x, z_1)'(\hat{\beta}^K - \beta^K)|_\delta + |\phi^K(x, z_1)'\beta^K - f_0(x, z_1)|_\delta \\
&\leq \zeta_\delta(K)||\hat{\beta}^K - \beta^K|| + O(K^{-s/(d_x+d_1)}) \leq \zeta_\delta(K)||\hat{\beta} - \beta_{\mathbf{L}}|| + O(K^{-s/(d_x+d_1)}) \\
&= O_p(\zeta_\delta(K)[\sqrt{\mathbf{L}/n} + L\triangle_n + \mathbf{L}^{-s/d}] + K^{-s/(d_x+d_1)})
\end{aligned}$$

where the second inequality holds by Assumption C1 (vii) and the last equality holds by (47). This completes the proof.

C Proof of asymptotic normality (Theorem 4 and 5)

C.1 Rate conditions

Along the proof, we obtain rate conditions to bound terms. We collect them here. Define

$$\begin{aligned}
\triangle_n^{\varphi} &= (\zeta_1(L) + L^{1/2}\xi_0(k)\sqrt{k/n} + L^{1/2})\triangle_n,\ \triangle_{n,\beta} = \sqrt{\mathbf{L}/n} + L\triangle_n + \mathbf{L}^{-\gamma} \\
\triangle_{\mathcal{T}} &= (\triangle_n^{\varphi})^2 + \mathbf{L}^{1/2}\triangle_n^{\varphi} + \zeta_0(\mathbf{L})\sqrt{\mathbf{L}/n}, \triangle_{\mathcal{T}_1} = \xi_0(k)\sqrt{k/n} \\
\triangle_H &= \zeta_0(\mathbf{L})k^{1/2}/\sqrt{n} + k^{1/2}\triangle_n^{\varphi} + L^{-\gamma}\zeta_0(\mathbf{L})\sqrt{k} \\
\triangle_{d\varphi} &= \zeta_0(\mathbf{L})L\triangle_{n,2}, \triangle_g = \zeta_0(\mathbf{L})\triangle_{n,\beta} \\
\triangle_{\Sigma} &= \mathbf{\Delta}_{\mathcal{T}} + \zeta_0(\mathbf{L})^2\mathbf{L}/n, \triangle_{\hat{H}} = (\zeta_1(L)\triangle_{n,\beta} + \xi_0(k)\triangle_{n,1})\mathbf{L}^{1/2}\xi_0(k)
\end{aligned}$$

and we need the following rate conditions for the $\sqrt{n}$-consistency and the consistency of the variance matrix estimator $\hat{\Omega}$:

$$\begin{gathered}
\sqrt{n}\mathbf{L}^{-\gamma} \to 0, \sqrt{n}k^{1/2}L^{-\gamma} \to 0, \sqrt{n}k^{-\gamma_1} \to 0, \sqrt{n}k^{-\gamma_2} \to 0 \\
k^{1/2}(\triangle_{\mathcal{T}_1} + \triangle_H) + \mathbf{L}^{1/2}\triangle_{\mathcal{T}} \to 0, n^{-1}(\zeta_0(\mathbf{L})^2\mathbf{L} + \xi_0(k)^2k + \xi_0(k)^2kL^4) \to 0, \\
k^{1/2}(\triangle_{\mathcal{T}_1} + \triangle_H) + \mathbf{L}^{1/2}\triangle_{\mathcal{T}} + \triangle_{d\varphi} \to 0, \triangle_g \to 0, \triangle_{\Sigma} \to 0, \triangle_{\hat{H}} \to 0.
\end{gathered}$$

Dropping the dominated ones and assuming $\sqrt{n}\mathbf{L}^{-\gamma}$, $\sqrt{n}k^{-\gamma_1}$, and $\sqrt{n}k^{-\gamma_2}$ are small enough, under the following all the rate conditions are satisfied:

$$\frac{\zeta_0(\mathbf{L})k + \zeta_1(L)k^{3/2} + \zeta_0(\mathbf{L})\mathbf{L} + \mathbf{L}\zeta_1(L)\xi_0(k) + \mathbf{L}^{1/2}\zeta_1(L)L\xi_0(k)k^{1/2} + \mathbf{L}^{1/2}\xi_0(k)^2k^{1/2}}{\sqrt{n}} \to 0.$$

For the polynomial approximations it becomes $\frac{\mathbf{L}^2+\mathbf{L}L^3k+\mathbf{L}^{1/2}(L^4k^{3/2}+k^{5/2})}{\sqrt{n}} \to 0$ and for the spline approximations it becomes $\frac{\mathbf{L}^{3/2}+\mathbf{L}L^{3/2}k^{1/2}+\mathbf{L}^{1/2}(L^{5/2}k+k^{3/2})+L^{3/2}k^{3/2}}{\sqrt{n}} \to 0$.

C.2 Asymptotic variance terms

Let $p_i^k = p^k(Z_i)$ and $p_i^k = (p_{1i}, \ldots, p_{ki})'$. We start with introducing additional notation:

$$\begin{aligned}
\Sigma &= E[\psi_i^{\mathbf{L}}\psi_i^{\mathbf{L}\prime}\mathrm{var}(Y_i|Z_i, V_i)],\ \mathcal{T} = E[\psi_i^{\mathbf{L}}\psi_i^{\mathbf{L}\prime}], \mathcal{T}_1 = E[p_i^k p_i^{k\prime}], \\
\Sigma_1 &= E[V_i^2 p_i^k p_i^{k\prime}],\ \Sigma_{2,l} = E[(\varphi_l(Z_i, V_i) - \bar{\varphi}_l(Z_i))^2 p_i^k p_i^{k\prime}], \\
H_{11} &= E[\frac{\partial h_{0i}}{\partial V_i}\psi_i^{\mathbf{L}} p_i^{k\prime}],\ \bar{H}_{11} = \sum_{i=1}^n \frac{\partial h_{0i}}{\partial V_i}\psi_i^{\mathbf{L}} p_i^{k\prime}/n \\
H_{12} &= E[E[\frac{\partial h_{0i}}{\partial V_i}|Z_i]\psi_i^{\mathbf{L}} p_i^{k\prime}],\ \bar{H}_{12} = \sum_{i=1}^n E[\frac{\partial h_{0i}}{\partial V_i}|Z_i]\psi_i^{\mathbf{L}} p_i^{k\prime}/n \\
H_{2,l} &= E[a_l\psi_i^{\mathbf{L}} p_i^{k\prime}],\ \bar{H}_{2,l} = \sum_{i=1}^n a_l\psi_i^{\mathbf{L}} p_i^{k\prime}/n,\ H_1 = H_{11} - H_{12},\ \bar{H}_1 = \bar{H}_{11} - \bar{H}_{12} \\
\bar{\Omega} &= \mathcal{A}\mathcal{T}^{-1}[\Sigma + H_1\mathcal{T}_1^{-1}\Sigma_1\mathcal{T}_1^{-1}H_1' + \sum\nolimits_{l=1}^L H_{2,l}\mathcal{T}_1^{-1}\Sigma_{2,l}\mathcal{T}_1^{-1}H_{2,l}']\mathcal{T}^{-1}\mathcal{A}'.
\end{aligned} \tag{50}$$

Let $\mathcal{T}_1 = I$ without loss of generality. Then $\bar{\Omega} = \mathcal{A}\mathcal{T}^{-1}\left[\Sigma + H_1\Sigma_1 H_1' + \sum_{l=1}^{L} H_{2,l}\Sigma_{2,l}H_{2,l}'\right]\mathcal{T}^{-1}\mathcal{A}'$. Let Γ be a symmetric square root of $\bar{\Omega}^{-1}$. Because $\mathcal{T}$ is nonsingular and $\text{var}(Y_i|Z_i, V_i)$ is bounded away from zero, $C\Sigma - I$ is positive semidefinite for some positive constant C. It follows that

$$\begin{aligned} ||\Gamma\mathcal{A}\mathcal{T}^{-1}|| &= \{\text{tr}(\Gamma\mathcal{A}\mathcal{T}^{-1}\mathcal{T}^{-1}\mathcal{A}'\Gamma')\}^{1/2} \leq \{\text{tr}(\Gamma\mathcal{A}\mathcal{T}^{-1}C\Sigma\mathcal{T}^{-1}\mathcal{A}'\Gamma')\}^{1/2} \\ &\leq \{\text{tr}(C\Gamma\bar{\Omega}\Gamma')\}^{1/2} \leq C \end{aligned}$$

and therefore $||\Gamma\mathcal{A}\mathcal{T}^{-1}||$ is bounded. Next we show $\bar{\Omega} \to \Omega$ as $k, K, L \to \infty$. Under Assumption R1, we have $\mathcal{A} = E[\nu^*(Z,V)\psi_i^{\mathbf{L}\prime}]$. Take $\nu_{\mathbf{L}}^*(Z,V) = \mathcal{A}\mathcal{T}^{-1}\psi_i^{\mathbf{L}}$. Then note $E[||\nu^*(Z,V) - \nu_{\mathbf{L}}^*(Z,V)||^2] \to 0$ because (i) $\nu_{\mathbf{L}}^*(Z,V) = E[\nu^*(Z,V)\psi_i^{\mathbf{L}\prime}]\mathcal{T}^{-1}\psi_i^{\mathbf{L}}$ is a mean-squared projection of $\nu^*(z_i, v_i)$ on $\psi_i^{\mathbf{L}}$; (ii) $\nu^*(z_i, v_i)$ is smooth and the second moment of $\nu^*(z_i, v_i)$ is bounded, so it is well-approximated in the mean-squared error as assumed in Assumption R1. Let $\nu_i^* = \nu^*(Z_i, V_i)$ and $\nu_{\mathbf{L}i}^* = \nu_{\mathbf{L}}^*(Z_i, V_i)$. It follows that

$$E[\nu_{\mathbf{L}i}^*\text{var}(Y_i|Z_i,V_i)\nu_{\mathbf{L}i}^{*\prime}] = \mathcal{A}\mathcal{T}^{-1}E[\psi_i^{\mathbf{L}}\text{var}(Y_i|Z_i,V_i)\psi_i^{\mathbf{L}\prime}]\mathcal{T}^{-1}\mathcal{A}' \to E[\nu_i^*\text{var}(Y_i|Z_i,V_i)\nu_i^{*\prime}].$$

It concludes that $\mathcal{A}\mathcal{T}^{-1}\Sigma\mathcal{T}^{-1}\mathcal{A}'$ converges to $E[\nu_i^*\text{var}(Y_i|Z_i,V_i)\nu_i^{*\prime}]$ (the first term in Ω) as $k, K, L \to \infty$. Next let

$$b_{\mathbf{L}i} = E[\nu_{\mathbf{L}i}^*\left(\frac{\partial h_{0i}}{\partial V_i} - E[\frac{\partial h_{0i}}{\partial V_i}|Z_i]\right)p_i^{k\prime}]p_i^k$$

and $b_i = E\left[\nu_i^*\left(\frac{\partial h_{0i}}{\partial V_i} - E[\frac{\partial h_{0i}}{\partial V_i}|Z_i]\right)p_i^{k\prime}\right]p_i^k$. Note that because $(\mathcal{T}_1)^{-1} = I$, $b_{\mathbf{L}i}$ and b_i are least squares mean projections of $\nu_{\mathbf{L}i}^*\left(\frac{\partial h_{0i}}{\partial V_i} - E[\frac{\partial h_{0i}}{\partial V_i}|Z_i]\right)$ on p_i^k and $\nu_i^*\left(\frac{\partial h_{0i}}{\partial V_i} - E[\frac{\partial h_{0i}}{\partial V_i}|Z_i]\right)$ on p_i^k, respectively. Then $E[||b_{\mathbf{L}i} - b_i||^2] \leq CE[||\nu_{\mathbf{L}i}^* - \nu_i^*||^2] \to 0$ where the first inequality holds because the mean square error of a least squares projection cannot be larger than the MSE of the variable being projected. Also note that $E[||\rho_v(Z_i) - b_i||^2] \to 0$ as $k \to \infty$ because b_i is a least squares projection of $\nu_i^*\left(\frac{\partial h_{0i}}{\partial V_i} - E\left[\frac{\partial h_{0i}}{\partial V_i}|Z_i\right]\right)$ on p_i^k and it converges to the conditional mean as $k \to \infty$. Finally note that

$$\begin{aligned} E[b_{\mathbf{L}i}\text{var}(V_i|Z_i)b_{\mathbf{L}i}'] &= \mathcal{A}\mathcal{T}^{-1}E\left[\psi_i^{\mathbf{L}}\left(\frac{\partial h_{0i}}{\partial V_i} - E\left[\frac{\partial h_{0i}}{\partial V_i}|Z_i\right]\right)p_i^{k\prime}\right]E[\text{var}(V_i|Z_i)p_i^kp_i^{k\prime}] \\ &\quad \times E\left[p_i^k\left(\frac{\partial h_{0i}}{\partial V_i} - E\left[\frac{\partial h_{0i}}{\partial V_i}|Z_i\right]\right)\psi_i^{\mathbf{L}\prime}\right]\mathcal{T}^{-1}\mathcal{A}' \\ &= \mathcal{A}\mathcal{T}^{-1}H_1\Sigma_1H_1'\mathcal{T}^{-1}\mathcal{A}' \end{aligned}$$

and therefore we conclude $\mathcal{A}\mathcal{T}^{-1}H_1\Sigma_1H_1'\mathcal{T}^{-1}\mathcal{A}'$ converges to $E[\rho_v(Z)\text{var}(X|Z)\rho_v(Z)']$ (the second term in Ω). Similarly we can show that for all l

$$\mathcal{A}\mathcal{T}^{-1}H_{2,l}\Sigma_{2,l}H_{2,l}'\mathcal{T}^{-1}\mathcal{A}' \to E[\rho_{\bar{\varphi}_l}(Z)\text{var}(\varphi_l(Z,V)|Z)\rho_{\bar{\varphi}_l}(Z)'].$$

We then conclude $\bar{\Omega} \to \Omega$ as $k, K, L \to \infty$. This also implies $\Gamma \to \Omega^{-1/2}$ and Γ is bounded.

C.3 Influence functions and asymptotic normality

Next we derive the asymptotic normality of $\sqrt{n}(\hat{\theta} - \theta_0)$. After we establish the asymptotic normality, we will show the convergence of the each term in (20) to the corresponding terms in (50). We show some of them here to be used for deriving the asymptotic normality. Note $||\hat{\mathcal{T}} - \mathcal{T}|| = O_p(\triangle_{\mathcal{T}}) = o_p(1)$ and $||\hat{\mathcal{T}}_1 - \mathcal{T}_1|| = O_p(\triangle_{\mathcal{T}_1}) = o_p(1)$. We also have $||\Gamma\mathcal{A}(\hat{\mathcal{T}}^{-1} - \mathcal{T}^{-1})|| = o_p(1)$ and $||\Gamma\mathcal{A}\hat{\mathcal{T}}^{-1/2}||^2 = O_p(1)$ (see proof in Lemma A1 of Newey, Powell, and Vella (1999)). We next show $||\bar{H}_{11} - H_{11}|| = o_p(1)$. Let $H_{11\mathbf{L}} = E[\sum_{l=1}^{L} a_l \frac{\partial\varphi_l(Z_i,V_i)}{\partial V_i}\psi_i^{\mathbf{L}} p_i^{k\prime}]$ and $\bar{H}_{11\mathbf{L}} = \sum_{i=1}^{n}\sum_{l=1}^{L} a_l \frac{\partial\varphi_l(Z_i,V_i)}{\partial V_i}\psi_i^{\mathbf{L}} p_i^{k\prime}/n$. Similarly define $H_{12\mathbf{L}} = E[E[\sum_{l=1}^{L} a_l \frac{\partial\varphi_l(Z_i,V_i)}{\partial V_i}|Z_i]\psi_i^{\mathbf{L}} p_i^{k\prime}]$ and $\bar{H}_{12\mathbf{L}} = \sum_{i=1}^{n} E[\sum_{l=1}^{L} a_l \frac{\partial\varphi_l(Z_i,V_i)}{\partial V_i}|Z_i]\psi_i^{\mathbf{L}} p_i^{k\prime}/n$ and let $H_{1\mathbf{L}} = H_{11\mathbf{L}} - H_{12\mathbf{L}}$. By Assumption N1 (i), L1 (iii) and the Cauchy-Schwarz inequality,

$$
\begin{aligned}
& ||H_1 - H_{1\mathbf{L}}||^2 \\
\leq\ & CE[||\{(\frac{\partial h_{0i}}{\partial V_i} - E[\frac{\partial h_{0i}}{\partial V_i}|Z_i]) - \sum_l a_l(\frac{\partial\varphi_l(Z_i,V_i)}{\partial V_i} - E[\frac{\partial\varphi_l(Z_i,V_i)}{\partial V_i}|Z_i])\}\psi_i^{\mathbf{L}} p_i^{k\prime}||^2] \\
\leq\ & CL^{-2\gamma}E[||\psi_i^{\mathbf{L}}||^2 \sum_{j=1}^{k} p_{ji}^2] = O(L^{-2\gamma}\zeta_0(\mathbf{L})^2 k).
\end{aligned}
$$

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

$$
\begin{aligned}
E[\sqrt{n}||\bar{H}_{11\mathbf{L}} - H_{11\mathbf{L}}||] &\leq C(E[(\sum_{l=1}^{L} a_l \frac{\partial\varphi_l(Z_i,V_i)}{\partial V_i})^2||\psi_i^{\mathbf{L}}||^2 \sum_{j=1}^{k} p_{ji}^2])^{1/2} \\
&= C(E[(\frac{\partial h_{Li}}{\partial V_i})^2||\psi_i^{\mathbf{L}}||^2 \sum_{j=1}^{k} p_{ji}^2])^{1/2} \leq C\zeta_0(\mathbf{L})k^{1/2}
\end{aligned}
$$

where the first equality holds because $\frac{\partial h_{Li}}{\partial V_i} = \sum_{l=1}^{L} a_l \frac{\partial\tilde{\varphi}_l(Z_i,V_i)}{\partial V_i} = \sum_{l=1}^{L} a_l \frac{\partial\varphi_l(Z_i,V_i)}{\partial V_i}$ and the last result holds because $h_{Li} \in \mathcal{H}_n$ (i.e. $|h_{Li}|_1$ is bounded). Similarly by (39), the Cauchy-Schwarz inequality, and the Markov inequality, we obtain

$$
\begin{aligned}
\left\|\bar{H}_{11} - \bar{H}_{11\mathbf{L}}\right\| &\leq Cn^{-1}\sum_{i=1}^{n} |\sum_{l=1}^{L} a_l \frac{\partial\varphi_l(Z_i,V_i)}{\partial V_i}| \cdot ||\hat{\psi}_i^{\mathbf{L}} - \psi_i^{\mathbf{L}}|| \cdot ||p_i^k|| \\
&\leq C(\sum_{i=1}^{n} C_i||\hat{\psi}_i^{\mathbf{L}} - \psi_i^{\mathbf{L}}||^2/n)^{1/2} \cdot \left(\sum_{i=1}^{n} ||p_i^k||^2/n\right)^{1/2} \leq O_p(k^{1/2}\triangle_n^{\varphi}).
\end{aligned}
$$

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

Now we derive the asymptotic expansion to obtain the influence functions. Further define $\hat{g}_{\mathbf{L}i} = f_K(x_i, z_{1i}) + \tilde{h}_L(z_i, \hat{v}_i)$ where $f_K(x_i, z_{1i}) = \sum_{j=1}^{K} \beta_j \phi_j(x_i, z_{1i})$ and $\tilde{h}_L(z_i, \hat{v}_i) = a_L'(\varphi^L(z_i, \hat{v}_i) - E[\varphi^L(Z_i, \hat{V}_i)|z_i])$ and $g_{\mathbf{L}i} = f_K(x_i, z_{1i}) + h_L(z_i, v_i)$. Recall $\beta_{\mathbf{L}} = (\beta_1, \ldots, \beta_K, a_L')'$ and let this $\beta_{\mathbf{L}}$ satisfy Assumption N1 (i). Then from the first order condition, we obtain the expansion similar to

(43) as [21]

$$\begin{aligned} 0 &= \hat{\bar{\psi}}^{\mathbf{L},n\prime}(y-\hat{g})/\sqrt{n} \\ &= \hat{\bar{\psi}}^{\mathbf{L},n\prime}(\eta-(\hat{g}-\hat{g}_{\mathbf{L}})-(\hat{\hat{g}}_{\mathbf{L}}-\hat{g}_{\mathbf{L}})-(\hat{g}_{\mathbf{L}}-g_{\mathbf{L}})-(g_{\mathbf{L}}-g_0))/\sqrt{n} \\ &= \hat{\bar{\psi}}^{\mathbf{L},n\prime}(\eta-\hat{\bar{\psi}}^{\mathbf{L},n}(\hat{\beta}-\beta_{\mathbf{L}})-(\hat{\hat{g}}_{\mathbf{L}}-\hat{g}_{\mathbf{L}})-(\hat{g}_{\mathbf{L}}-g_{\mathbf{L}})-(g_{\mathbf{L}}-g_0))/\sqrt{n}. \end{aligned} \tag{51}$$

Similar to (44), we obtain

$$||\hat{\mathcal{T}}^{-1}\hat{\bar{\psi}}^{\mathbf{L},n\prime}(g_{\mathbf{L}}-g_0)/\sqrt{n}|| = O_p(\sqrt{n}\mathbf{L}^{-\gamma}). \tag{52}$$

Also note that because $\alpha(\cdot)$ is a linear functional and by Assumption N1 (i),

$$\begin{aligned} \sqrt{n}||\Gamma(\alpha(g_{\mathbf{L}})-\alpha(g_0))|| &= \sqrt{n}||\Gamma||\cdot||\alpha(g_{\mathbf{L}}-g_0)|| \leq C\sqrt{n}\,\|\Gamma\|\cdot|\psi^{\mathbf{L}\prime}(\cdot)\beta_{\mathbf{L}}-g_0(\cdot)|_\delta \\ &= O_p(\sqrt{n}\mathbf{L}^{-\gamma}) = o_p(1). \end{aligned} \tag{53}$$

Then from the linearity of $\alpha(\cdot)$, (51), (52), and (53) we have

$$\begin{aligned} \sqrt{n}\Gamma(\hat{\theta}-\theta_0) &= \sqrt{n}\Gamma(\alpha(\hat{g})-\alpha(g_0)) = \sqrt{n}\Gamma(\alpha(\hat{g})-\alpha(g_{\mathbf{L}}))+\sqrt{n}\Gamma(\alpha(g_{\mathbf{L}})-\alpha(g_0)) \\ &= \sqrt{n}\Gamma\mathcal{A}(\hat{\beta}-\beta_{\mathbf{L}})+\sqrt{n}\Gamma\{a(g_{\mathbf{L}})-a(g_0)\} \\ &= \Gamma\mathcal{A}\hat{\mathcal{T}}^{-1}\hat{\bar{\psi}}^{\mathbf{L},n\prime}(\eta-(\hat{\hat{g}}_{\mathbf{L}}-\hat{g}_{\mathbf{L}})-(\hat{g}_{\mathbf{L}}-g_{\mathbf{L}}))/\sqrt{n}+o_p(1). \end{aligned} \tag{54}$$

C.3.1 Influence function for the first stage

Now we derive the stochastic expansion of $\Gamma\mathcal{A}\hat{\mathcal{T}}^{-1}\hat{\bar{\psi}}^{\mathbf{L},n\prime}(\hat{g}_{\mathbf{L}}-g_{\mathbf{L}})/\sqrt{n}$. Note that by a second order mean-value expansion of each $\tilde{h}_{Li}=\tilde{h}_L(z_i,\hat{v}_i)$ around v_i (also write $h_{Li}=h_L(z_i,v_i)$),

$$\begin{aligned} &\Gamma\mathcal{A}\hat{\mathcal{T}}^{-1}\textstyle\sum_{i=1}^n\hat{\bar{\psi}}_i^{\mathbf{L}}(\hat{g}_{\mathbf{L}i}-g_{\mathbf{L}i})/\sqrt{n} = \Gamma\mathcal{A}\hat{\mathcal{T}}^{-1}\sum_{i=1}^n\hat{\bar{\psi}}_i^{\mathbf{L}}(\tilde{h}_{Li}-h_{Li})/\sqrt{n} \\ =\ &\Gamma\mathcal{A}\hat{\mathcal{T}}^{-1}\textstyle\sum_{i=1}^n\hat{\bar{\psi}}_i^{\mathbf{L}}(\frac{dh_{Li}}{dv_i}-E[\frac{dh_{Li}}{dV_i}|Z_i])(\hat{\Pi}_i-\Pi_i)/\sqrt{n}+\hat{\varsigma} \\ =\ &\Gamma\mathcal{A}\hat{\mathcal{T}}^{-1}\bar{H}_1\hat{\mathcal{T}}_1^{-1}\textstyle\sum_{i=1}^n p_i^k v_i/\sqrt{n}+\Gamma\mathcal{A}\hat{\mathcal{T}}^{-1}\bar{H}_1\hat{\mathcal{T}}_1^{-1}\sum_{i=1}^n p_i^k(\Pi_i-p_i^{k\prime}\lambda_k^1)/\sqrt{n} \\ &+\Gamma A\hat{\mathcal{T}}^{-1}\textstyle\sum_{i=1}^n\hat{\bar{\psi}}_i^{\mathbf{L}}(\frac{dh_{Li}}{dv_i}-E[\frac{dh_{Li}}{dV_i}|Z_i])(p_i^{k\prime}\lambda_k^1-\Pi_i)/\sqrt{n}+\hat{\varsigma}. \end{aligned} \tag{55}$$

The remainder term $||\hat{\varsigma}|| \leq C\sqrt{n}||\Gamma\mathcal{A}\hat{\mathcal{T}}^{-1/2}||\zeta_0(L)\sum_{i=1}^n C_i||\hat{\Pi}_i-\Pi_i||^2/n = O_p(\sqrt{n}\zeta_0(L)\triangle_{n,1}^2) = o_p(1)$. 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 (55) are $o_p(1)$ under $\sqrt{n}k^{-s_1/d_z}\to 0$ (so that $\sqrt{n}\left|\Pi_0(z)-\lambda_k^{1\prime}p^k(z)\right|_0\to 0$ by Assumption L1 (iv)) and under $k^{1/2}(\triangle_{\mathcal{T}_1}+\triangle_H)+\mathbf{L}^{1/2}\triangle_{\mathcal{T}}\to$

[21] If there exists an estimation error due to tolerance in minimization, take the error arbitrary small to justify this asymptotic expansion.

0 (so that we can replace $\hat{\mathcal{T}}_1$ with $\mathcal{T}_1$, $\bar{H}_1$ with H_1, and $\hat{\mathcal{T}}$ with $\mathcal{T}$ respectively). We therefore obtain

$$\Gamma\mathcal{A}\hat{\mathcal{T}}^{-1}\hat{\psi}^{\mathbf{L},n\prime}(\hat{g}_{\mathbf{L}}-g_{\mathbf{L}})/\sqrt{n}=\Gamma\mathcal{A}\mathcal{T}^{-1}H_1\sum\nolimits_{i=1}^{n}p_i^k v_i/\sqrt{n}+o_p(1). \tag{56}$$

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

C.3.2 Influence function for the second stage

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

$$\begin{aligned}
&\Gamma\mathcal{A}\hat{\mathcal{T}}^{-1}\sum_{i=1}^{n}\hat{\hat{\psi}}_i^{\mathbf{L}}(\hat{\hat{g}}_{\mathbf{L}i}-\hat{g}_{\mathbf{L}i})/\sqrt{n}=\Gamma\mathcal{A}\hat{\mathcal{T}}^{-1}\sum_{i=1}^{n}\hat{\hat{\psi}}_i^{\mathbf{L}}a_L'(\hat{\hat{\varphi}}^L(z_i)-E[\varphi^L(Z_i,\hat{V}_i)|z_i])/\sqrt{n}\\
=\;&\Gamma\mathcal{A}\hat{\mathcal{T}}^{-1}\{\sum\nolimits_l\bar{H}_{2,l}\hat{\mathcal{T}}_1^{-1}\sum\nolimits_{i=1}^{n}p_i^k\tilde{\varphi}_{li}+\sum\nolimits_l\bar{H}_{2,l}\hat{\mathcal{T}}_1^{-1}\sum\nolimits_{i=1}^{n}p_i^k(\bar{\varphi}_l(z_i)-p_i^{k\prime}\lambda_{l,k}^2)\}/\sqrt{n}\\
&+\Gamma\mathcal{A}\hat{\mathcal{T}}^{-1}\sum\nolimits_{i=1}^{n}\hat{\hat{\psi}}_i^{\mathbf{L}}\sum\nolimits_l a_l(p_i^{k\prime}\lambda_{l,k}^2-\bar{\varphi}_l(z_i))/\sqrt{n}+\Gamma\mathcal{A}\hat{\mathcal{T}}^{-1}\sum\nolimits_{i=1}^{n}\hat{\hat{\psi}}_i^{\mathbf{L}}\rho_i/\sqrt{n}
\end{aligned} \tag{57}$$

where $\rho_i=\sum_l a_l\{p_i^{k\prime}\hat{\mathcal{T}}_1^{-1}\sum_{j=1}^{n}p_j^k(\varphi_l(z_j,\hat{v}_j)-\varphi_l(z_j,v_j))/n-(E[\varphi_l(Z_i,\hat{V}_i)|z_i]-\bar{\varphi}_l(z_i))\}$. We first focus on the last term in (57). Note that $p_i^{k\prime}\hat{\mathcal{T}}_1^{-1}\sum_{i=1}^{n}p_i^k(\varphi_l(z_i,\hat{v}_i)-\varphi_l(z_i,v_i))/n$ is a least squares projection of $\varphi_l(z_i,\hat{v}_i)-\varphi_l(z_i,v_i)$ on p_i^k and it converges to the conditional mean $E[\varphi_l(Z_i,\hat{V}_i)|z_i]-\bar{\varphi}_l(z_i)$. Therefore we can write $\rho_i=\sum_{l=1}^{L}a_l\rho_{il}$ and ρ_{il} is the projection residual from the least squares projection of $\varphi_l(z_i,\hat{v}_i)-\varphi_l(z_i,v_i)$ on p_i^k for each l. It follows that $E[\rho_i|Z_1,\ldots,Z_n]=0$ and therefore

$$E[||\rho_i||^2|Z_1,\ldots,Z_n]\leq E[L\sum\nolimits_{l=1}^{L}||\rho_{il}||^2|Z_1,\ldots,Z_n]\leq L^2O_p(\triangle_{n,2}^2)$$

where the first inequality holds by the Cauchy-Schwarz inequality and the second inequality holds by a similar proof to (31) and by the Markov inequality. It follows that by Assumption L1 (iii) and the Cauchy-Schwarz inequality,

$$E[\|\sum\nolimits_{i=1}^{n}\hat{\hat{\psi}}_i^{\mathbf{L}}\rho_i/\sqrt{n}\||Z_1,\ldots,Z_n]\leq(E[||\hat{\hat{\psi}}_i^{\mathbf{L}}||^2||\rho_i||^2|Z_1,\ldots,Z_n])^{1/2}\leq C\zeta_0(\mathbf{L})L\triangle_{n,2}.$$

This implies that $\sum_{i=1}^{n}\hat{\hat{\psi}}_i^{\mathbf{L}}\rho_i/\sqrt{n}=O_p(\zeta_0(\mathbf{L})L\triangle_{n,2})\equiv O_p(\triangle_{d\varphi})=o_p(1)$.

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 (57) are $o_p(1)$ under $\sqrt{n}k^{-s_2/d_z}\to 0$ (so that $\sqrt{n}|\bar{\varphi}_l(z)-\lambda_{l,k}^{2\prime}p^k(z)|_0\to 0$ for all l by Assumption L1 (iv)), under $\sqrt{n}k^{1/2}L^{-s/d}\to 0$ (so that $\sqrt{n}k^{1/2}|h_0(z,v)-a_L'\tilde{\varphi}^L(z,v)|_0\to 0$ by Assumption L1 (iv)), and under $k^{1/2}(\triangle_{\mathcal{T}_1}+\triangle_H)+\mathbf{L}^{1/2}\triangle_{\mathcal{T}}+\triangle_{d\varphi}\to 0$ (so that we can replace $\hat{\mathcal{T}}_1$ with $\mathcal{T}_1$, $\bar{H}_{2,l}$ with $H_{2,l}$, and $\hat{\mathcal{T}}$ with $\mathcal{T}$ respectively). We therefore obtain

$$\Gamma\mathcal{A}\hat{\mathcal{T}}^{-1}\hat{\psi}^{\mathbf{L},n\prime}(\hat{\hat{g}}_{\mathbf{L}}-\hat{g}_{\mathbf{L}})/\sqrt{n}=\Gamma\mathcal{A}\mathcal{T}^{-1}\sum\nolimits_l H_{2,l}\sum\nolimits_{i=1}^{n}p_i^k\tilde{\varphi}_{li}/\sqrt{n}+o_p(1). \tag{58}$$

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

We can also show that replacing $\hat{\hat{\psi}}_i^{\mathbf{L}}$ with $\psi_i^{\mathbf{L}}$ does not influence the stochastic expansion by

(39). Therefore by (54), (56), and (58), we obtain the stochastic expansion,

$$\sqrt{n}\Gamma(\hat{\theta}-\theta_0) = \Gamma\mathcal{A}\mathcal{T}^{-1}(\psi^{\mathbf{L},n\prime}\eta - H_1\sum_{i=1}^n p_i^k v_i/\sqrt{n} - \sum_l H_{2,l}\sum_{i=1}^n p_i^k\tilde{\varphi}_{li}/\sqrt{n}) + o_p(1).$$

To apply the Lindeberg-Feller theorem, we check the Lindeberg condition. For any vector q with $||q|| = 1$, let $W_{in} = q'\Gamma\mathcal{A}\mathcal{T}^{-1}(\psi_i^{\mathbf{L}}\eta_i - H_1 p_i^k v_i - \sum_l H_{2,l}p_i^k\tilde{\varphi}_{li})/\sqrt{n}$. Note that W_{in} is i.i.d, given n and by construction, $E[W_{in}] = 0$ and $\text{var}(W_{in}) = O(1/n)$. Also note that $||\Gamma\mathcal{A}\mathcal{T}^{-1}|| \leq C$, $||\Gamma\mathcal{A}\mathcal{T}^{-1}H_j|| \leq C||\Gamma\mathcal{A}\mathcal{T}^{-1}|| \leq C$ by $CI - H_jH_j'$ being positive semidefinite for $j = 1,(2,1),\ldots,(2,L)$. Also note that $(\sum_{l=1}^L\tilde{\varphi}_{li})^4 \leq L^2(\sum_{l=1}^L\tilde{\varphi}_{li}^2)^2 \leq L^3\sum_{l=1}^L\tilde{\varphi}_{li}^4$. It follows that for any $\varepsilon > 0$,

$$\begin{aligned}
& nE[1(|W_{in}| > \varepsilon)W_{in}^2] = n\varepsilon^2 E[1(|W_{in}| > \varepsilon)(W_{in}/\varepsilon)^2] \leq n\varepsilon^{-2}E[|W_{in}|^4] \\
\leq\ & Cn\varepsilon^{-2}\{E[||\psi_i^{\mathbf{L}}||^4 E[\eta_i^4|Z_i,V_i]] + E[||p_i^k||^4 E[V_i^4|Z_i]] + L^3\sum_l E[||p_i^k||^4 E[\tilde{\varphi}_{li}^4|Z_i]]\}/n^2 \\
\leq\ & Cn^{-1}(\zeta_0(\mathbf{L})^2\mathbf{L} + \xi_0(k)^2 k + \xi_0(k)^2 kL^4) = o(1).
\end{aligned}$$

Therefore, $\sqrt{n}\Gamma(\hat{\theta}-\theta_0) \to_d N(0,I)$ by the Lindeberg-Feller central limit theorem. We have shown that $\bar{\Omega} \to \Omega$ and Γ is bounded. We therefore also conclude $\sqrt{n}(\hat{\theta}-\theta_0) \to_d N(0,\Omega^{-1})$.

C.4 Consistency of the estimate of the asymptotic variance

Now we show the convergence of the each term in (20) to the corresponding terms in (50). Let $\hat{\eta}_i = y_i - \hat{g}(z_i,\hat{v}_i)$. Note that $\hat{\eta}_i^* \equiv \hat{\eta}_i^2 - \eta_i^2 = -2\eta_i(\hat{g}_i - g_{0i}) + (\hat{g}_i - g_{0i})^2$ and that $\max_{i\leq n}|\hat{g}_i - g_{0i}| = O_p(\zeta_0(\mathbf{L})\triangle_{n,\beta}) = o_p(1)$ by (49). Let $\hat{D} = \Gamma\mathcal{A}\hat{\mathcal{T}}^{-1}\hat{\psi}^{\mathbf{L},n\prime}\text{diag}\{1+|\eta_i|,\ldots,1+|\eta_n|\}\hat{\psi}^{\mathbf{L},n}\hat{\mathcal{T}}^{-1}\mathcal{A}'\Gamma'$ and note that $\hat{\psi}^{\mathbf{L},n}$ and $\hat{\mathcal{T}}$ only depend on $(Z_1,V_1),\ldots,(Z_n,V_n)$ and thus $E[\hat{D}|(Z_1,V_1),\ldots,(Z_n,V_n)] \leq C\Gamma\mathcal{A}\hat{\mathcal{T}}^{-1}\mathcal{A}'\Gamma' = O_p(1)$. Therefore, $||\hat{D}|| = O_p(1)$ as well. Next let $\tilde{\Sigma} = \sum_{i=1}^n\hat{\psi}_i^{\mathbf{L}}\hat{\psi}_i^{\mathbf{L}\prime}\eta_i^2/n$. Then,

$$\begin{aligned}
||\Gamma\mathcal{A}\hat{\mathcal{T}}^{-1}(\hat{\Sigma}-\tilde{\Sigma})\hat{\mathcal{T}}^{-1}\mathcal{A}'\Gamma'|| &= ||\Gamma\mathcal{A}\hat{\mathcal{T}}^{-1}\hat{\psi}^{\mathbf{L},n\prime}\text{diag}\{\hat{\eta}_1^*,\ldots,\hat{\eta}_n^*\}\hat{\psi}^{\mathbf{L},n}\hat{\mathcal{T}}^{-1}\mathcal{A}'\Gamma'|| \\
&\leq C\text{tr}(\hat{D})\max_{i\leq n}|\hat{g}_i - g_{0i}| = O_p(1)o_p(1).
\end{aligned} \tag{59}$$

Also by the essentially same proofs in Lemma A2 of Newey, Powell, and Vella (1999),

$$\begin{aligned}
||\tilde{\Sigma}-\Sigma|| &= O_p(\Delta_{\mathcal{T}} + \zeta_0(\mathbf{L})^2\mathbf{L}/n) \equiv O_p(\triangle_\Sigma) = o_p(1), \\
||\Gamma\mathcal{A}\hat{\mathcal{T}}^{-1}(\hat{\Sigma}-\Sigma)\hat{\mathcal{T}}^{-1}\mathcal{A}'\Gamma'|| &= o_p(1),\ ||\Gamma\mathcal{A}(\hat{\mathcal{T}}^{-1}\Sigma\hat{\mathcal{T}}^{-1} - \mathcal{T}^{-1}\Sigma\mathcal{T}^{-1})\mathcal{A}'\Gamma'|| = o_p(1).
\end{aligned} \tag{60}$$

It also follows that $||\hat{\Sigma}-\Sigma|| = o_p(1)$ because $||\Gamma\mathcal{A}\hat{\mathcal{T}}^{-1}|| = O_p(1)$. Then, by (59-60) and the triangle inequality, we find $||\Gamma\mathcal{A}\hat{\mathcal{T}}^{-1}\hat{\Sigma}\hat{\mathcal{T}}^{-1}\mathcal{A}'\Gamma' - \Gamma\mathcal{A}\mathcal{T}^{-1}\Sigma\mathcal{T}^{-1}\mathcal{A}'\Gamma'|| = o_p(1)$. It remains to show that for $j = 1,(2,1),\ldots,(2,L)$,

$$\Gamma\mathcal{A}(\hat{\mathcal{T}}^{-1}\hat{H}_j\hat{\mathcal{T}}_1^{-1}\hat{\Sigma}_j\hat{\mathcal{T}}_1^{-1}\hat{H}_j'\hat{\mathcal{T}}^{-1} - \mathcal{T}^{-1}H_j\Sigma_jH_j'\mathcal{T}^{-1})\mathcal{A}'\Gamma' = o_p(1). \tag{61}$$

As we have shown $||\hat{\Sigma}-\Sigma|| = o_p(1)$, similarly we can show $||\hat{\Sigma}_j-\Sigma_j|| = o_p(1)$, $j = 1,(2,1),\ldots,(2,L)$. We focus on showing $||\hat{H}_j - \bar{H}_j|| = o_p(1)$ for $j = 1,(2,1),\ldots,(2,L)$. First note that $||\hat{H}_{11} - \bar{H}_{11}|| =$

$||\sum_{i=1}^n(\sum_{l=1}^L \hat{a}_l\frac{\partial\varphi_l(z_i,\hat{v}_i)}{\partial v_i} - a_l\frac{\partial\varphi_l(z_i,v_i)}{\partial v_i})\hat{\hat{\psi}}_i^{\mathbf{L}}p^k(z_i)'/n||$. By the Cauchy-Schwarz inequality, (38), and Assumption L1 (iii), we have $\sum_{i=1}^n||\hat{\hat{\psi}}_i^{\mathbf{L}}p_i^{k\prime}||^2/n \leq \sum_{i=1}^n||\hat{\hat{\psi}}_i^{\mathbf{L}}||^2||p_i^k||^2/n = O_p(\mathbf{L}\xi_0(k)^2)$. Also note that by the triangle inequality, the Cauchy-Schwarz inequality, and by Assumption C1 (vi) and (32), applying a mean value expansion to $\frac{\partial\varphi_l(z_i,v_i)}{\partial v_i}$ w.r.t v_i,

$$
\begin{aligned}
& \sum_{i=1}^n||\sum_{l=1}^L(\hat{a}_l\frac{\partial\varphi_l(z_i,\hat{v}_i)}{\partial v_i} - a_l\frac{\partial\varphi_l(z_i,v_i)}{\partial v_i})||^2/n \\
\leq\ & 2\sum_{i=1}^n||\sum_{l=1}^L(\hat{a}_l - a_l)\frac{\partial\varphi_l(z_i,v_i)}{\partial v_i}||^2/n + 2\sum_{i=1}^n||\sum_{l=1}^L\hat{a}_l(\frac{\partial\varphi_l(z_i,\hat{v}_i)}{\partial v_i} - \frac{\partial\varphi_l(z_i,v_i)}{\partial v_i})||^2/n \\
\leq\ & C||\hat{a}-a_L||^2\sum_{i=1}^n||\frac{\partial\tilde{\varphi}^L(z_i,v_i)}{\partial v_i}||^2/n + C_1\sum_{i=1}^n||\sum_{l=1}^L\hat{a}_l\frac{\partial^2\varphi_l(z_i,\tilde{v}_i)}{\partial v_i^2}(\hat{\Pi}_i-\Pi_i)||^2/n \\
\leq\ & C||\hat{a}-a_L||^2\sum_{i=1}^n||\frac{\partial\tilde{\varphi}^L(z_i,v_i)}{\partial v_i}||^2/n + C_1\max_{1\leq i\leq n}||\hat{\Pi}_i-\Pi_i||^2\sum_{i=1}^n||\sum_{l=1}^L\hat{a}_l\frac{\partial^2\varphi_l(z_i,\tilde{v}_i)}{\partial v_i^2}||^2/n \\
=\ & O_p(\zeta_1^2(L)\triangle_{n,\beta}^2 + \xi_0^2(k)\triangle_{n,1}^2)
\end{aligned}
$$

where $\tilde{v}_i$ lies between $\hat{v}_i$ and v_i, which may depend on l. We therefore conclude by the triangle inequality and the Cauchy-Schwarz inequality, $||\hat{H}_{11}-\bar{H}_{11}|| \leq O_p((\zeta_1(L)\triangle_{n,\beta}+\xi_0(k)\triangle_{n,1})\mathbf{L}^{1/2}\xi_0(k)) = O_p(\triangle_{\hat{H}}) = o_p(1)$. Similarly we can show that $||\hat{H}_{12}-\bar{H}_{12}|| = o_p(1)$ and $||\hat{H}_{2,l}-\bar{H}_{2,l}|| = o_p(1)$ $l = 1,\ldots,L$. We have shown that $||\bar{H}_j - H_j|| = o_p(1)$ for $j = 1,(2,1),\ldots,(2,L)$ previously. Therefore, $||\hat{H}_j - H_j|| = o_p(1)$ for $j = 1,(2,1),\ldots,(2,L)$. Then by the similar proof like (59) and (60), the conclusion (61) follows. From (59-61) finally note that by Γ is bounded, $||\hat{\Omega}-\bar{\Omega}|| \leq C||\Gamma\hat{\Omega}\Gamma' - \Gamma\bar{\Omega}\Gamma'|| = o_p(1)$.

References

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

Blundell, R., X. Chen, and D. Kristensen (2007): "Semi-Nonparametric IV Estimation of Shape-Invariant Engel Curves," *Econometrica*, 75, 1613–1669.

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

Chen, X., and D. Pouzo (2009): "Efficient Estimation of Semiparametric Conditional Moment Models with Possibly Nonsmooth Residuals," *Journal of Econometrics*, 152, 46–60.

——— (2011): "Estimation of Nonparametric Conditional Moment Models with Possibly Nonsmooth Generalized Residuals," *Econometrica forthcoming*.

Chen, X., and M. Reiss (2011): "On Rate Optimality for Ill-posed Inverse Problems in Econometrics," *Econometric Theory*, 75, 1–25.

Chernozhukov, V., and C. Hansen (2005): "An IV Model of Quantile Treatment Effects," *Econometrica*, 73, 245–262.

Darolles, S., Y. Fan, J. Florens, and E. Renault (2011): "Nonparametric Instrumental Regression," *Econometrica forthcoming*.

Darolles, S., J. Florens, and E. Renault (2006): "Nonparametric Instrumental Regression," working paper, GREMAQ, University of Toulouse.

Engl, H., M. Hanke, and A. Neubauer (1996): *Regularization of Inverse Problems*Dordrecht: Kluwer Academic.

Florens, J. (2003): "Inverse Problems and Structural Econometrics: The Example of Instrumental Variables," *Advances in Economics and Econometrics: Theory and Applications, Eight World Congress,Econometric Society Monographs*, pp. 1–25.

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.

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

Hausman, J. (1978): "Specification Tests in Econometrics," *Econometrica*, 46, 1251–1272.

Heckman, J. (1978): "Specification Tests in Econometrics," *Econometrica*, 46, 1261–1272.

Horowitz, J. (2007): "Asymptotic Normality of a Nonparametric Instrumental Variables Estimator," *International Economic Review*, 48, 1329–1349.

——— (2011a): "Applied Nonparametric Instrumental Variables Estimation," *Econometrica*, 79, 347–394.

——— (2011b): "Specification Testing in Nonparametric Instrumental Variables Estimation," *Journal of Econometrics forthcoming*.

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

Johannes, J., S. V. Bellegem, and A. Vanhems (2011): "Convergence Rates for Ill-posed Inverse Problems with an Unknown Operator," *Econometric Theory*, 27, 522–545.

Kress, R. (1999): *Linear Integral Equations*New York: Springer-Verlag.

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

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.

Rivers, D., and Q. Vuong (1988): "Limited Information Estimators and Exogeneity Tests for Simultaneous Probit Models," *Journal of Econometrics*, 39, 347–366.

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

Smith, R., and R. Blundell (1986): "An Exogeneity Test for a Simultaneous Equation Tobit Model with an Application to Labor Supply," *Econometrica*, 54, 679–685.

Telser, L. (1964): "Iterative Estimation of a Set of Linear Regression Equations," *Journal of the American Statistical Association*, 59, 845–862.

Tikhonov, A., A. Goncharsky, V. Stepanov, and A. Yagola (1995): *Numerical Methods for the Solution of Ill-Posed Problems*Mathematics and Its Applications, New York: Springer.

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

Wooldridge, J. (2005): "Unobserved heterogeneity and estimation of average partial effects," in *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*, ed. by S. J. E. Andrews, D.W.K., pp. 27–55. Cambridge.

Wooldridge, J., and L. Papke (2008): "Panel Data Methods for Fractional Response Variables with an Application to Test Pass Rates," *Journal of Econometrics*, 145, 121–133.