Research Article

Kyoo il Kim* and Amil Petrin*

A Generalized Non-Parametric Instrumental Variable-Control Function Approach to Estimation in Nonlinear Settings

<https://doi.org/10.1515/jem-2021-0038>

Received December 17, 2021; accepted December 17, 2021; published online January 4, 2022

Abstract: When the endogenous variables enter non-parametrically into the regression equation standard linear instrumental variables approaches fail. Two existing solutions are the non-parametric instrumental variables (NPIVs) estimators, which are based on a set of conditional moment restrictions (CMRs), and the control function (CF) estimators, which use conditional mean independence (CMI) restrictions. Our first contribution is to show that – similar to CMI – the CMR place shape restrictions on the conditional expectation of the error given the instruments and endogenous variables that are sufficient for identification, and we call our new estimator based on these restrictions the CMR-CF estimator. Our second contribution is to develop an estimator for non-linear and non-parametric settings that can combine both CMR and CMI restrictions, which cannot be done in either the NPIV nor the non-parametric CF setting. This new “Generalized CMR-CF” uses both CMR and CMI restrictions together by allowing the conditional expectation of the structural error to depend on both instruments and control variables. When sieves are used to approximate both the structural function and the CF our estimator reduces to a series of least squares regressions. Our Monte Carlos illustrate that our new estimator performs well across several economic settings.

Keywords: conditional moment restriction, instrumental variables, control function, non-parametric estimation, sieve estimation

JEL Classification: C14, C21, C26

1 Introduction

The problem of endogenous regressors in simultaneous equations models has a long history in econometrics and empirical studies. We consider a triangular non-parametric simultaneous equations model with endogeneity to illustrate our main idea. Let x_i denote the endogenous regressors and z_i denote the instruments. The unknown structural function of interest is $f_0(x_i, z_{if})$, where $z_i = (z_{if}, z_{i,-f})$ with z_{if} the subset of z_i entering the structural function. The model is given as:

$$y_i = f_0(x_i, z_{if}) + \varepsilon_i. \quad (1)$$

$$x_i = \Pi_0(z_i) + v_i, \quad E[v_i|z_i] = 0 \quad (2)$$

*Corresponding authors: **Kyoo il Kim**, Michigan State University, East Lansing, USA, E-mail: kyookim@msu.edu; and **Amil Petrin**, University of Minnesota and NBER, Minneapolis, USA, E-mail: petrin@umn.edu

with ε_i denoting the error (so $E[\varepsilon_i|x_i] \neq 0$) and (2) characterizing the conditional mean decomposition of x_i (so $\Pi_0(z_i) = E[x_i|z_i]$ and $E[v_i|z_i] = 0$ is not restrictive). The econometric problem is further complicated as linear instrumental variable (IV) approaches are no longer consistent because the endogenous variables enter the structural equation non-parametrically.

There are currently two approaches to identification for this case. One is the non-parametric control function (CF) estimator (see e.g. Florens et al. 2008; Kasy 2011; Newey, Powell, and Vella 1999; Pinkse 2000; Su and Ullah 2008). Newey, Powell, and Vella (1999) (NPV) use the conditional mean decomposition from (2) to create “control” variables that are orthogonal to all of the instruments. Reexpressing (1) in “CF” form y_i expressed as a function of (z_i, v_i) is given as

$$y_i = f_0(x_i, z_{if}) + h_0(z_i, v_i) + \eta_i \quad (3)$$

where without loss of generality $h_0(z_i, v_i) = E[\varepsilon_i|z_i, v_i]$ is the CF and $\eta_i = \varepsilon_i - h_0(z_i, v_i)$ (with $E[\eta_i|z_i, v_i] = 0$ by construction). Without further restrictions on $h_0(z_i, v_i)$ the function $f_0(x_i, z_{if})$ is not identified because one cannot separate the effect on y_i of (x_i, z_{if}) via f_0 from their impact on y_i via h_0 . This literature achieves identification by assuming ε_i is conditionally mean independent (CMI) of z_i conditional on v_i so that $h_0(z_i, v_i)$ does not depend on z_i given v_i and can be written as $h_0(v_i)$:

$$(\text{CMI}) \quad E[\varepsilon_i|z_i, v_i] = E[\varepsilon_i|v_i] = h_0(v_i). \quad (4)$$

Given this CMI $f_0(x_i, z_{if})$ is then separately identified from $h_0(v_i)$ as long as an appropriate full rank condition holds.¹

The second approach is the non-parametric instrumental variables (NPIVs) method which uses conditional moment restrictions (CMRs)

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

The structural function in (1) is the solution to the integral equation implied by the CMRs:

$$E[y_i | z_i] = E[f_0(x_i, z_{if}) | z_i].$$

Our first contribution is to show – in the CF setup of Eq. (3) – if we have a sufficient number of CMRs we can identify and consistently estimate $f_0(x_i, z_{if})$ by restricting $h_0(z_i, v_i)$ to satisfy these CMR restrictions. 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

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

In Eq. (3) the CMR shape restrictions imply that the mean of $h_0(z_i, v_i)$ cannot vary by z_i alone, meaning $h_0(z_i, v_i)$ can be distinguished from any function of z_i only. As long as there are a sufficient number of instruments $f_0(x_i, z_{if})$ is then separately identified from $h_0(z_i, v_i)$. We call this new estimator the CMR-CF estimator.

In the NPIV literature it is well known that the above equation on the relationship between the structural function $f_0(x_i, z_{if})$ and the reduced form $E[y_i|z_i]$ is a Fredholm integral equation of the first kind, and solving for the function is subject to the so-called ill-posed inverse problem. Since our approach relies on CMR it is subject to the same ill-posed inverse problem. There is a large literature on restrictions that make the problem well-posed by means of regularization (see e.g. Ai and Chen 2003; Blundell, Chen, and Kristensen 2007; Darolles, Florens, and Renault 2006; Florens 2003; Hall and Horowitz 2005; Horowitz 2011; Newey and Powell 2003).² Our regularization method imposes finite dimensional sieve approximations on the CF – $h_0(z_i, v_i) \equiv$

¹ NPV achieve identification of $f_0(x_i, z_{if})$ by combining this restriction with (i) differentiability of $f_0(x, z_f)$, $h_0(v) = E[\varepsilon|v]$, and $\Pi_0(z)$, (ii) zero mass on the boundary of the support of (z, v) , and (iii) full rank of $\frac{\partial \Pi_0(z)}{\partial z_{-f}}$ with probability one. Identification holds as long as (x, z_f) and a control variate v are measurably separated and imposing NPV's sufficient conditions is one way to achieve this separability (see also Florens et al. (2008) and their use of Matzkin (2003)'s control).

² See e.g. Chen (2007) for related discussions on using a sieve approach for the regularization. Newey and Powell (2003) eliminate the ill-posed inverse problem by assuming f_0 to be in a compact set, and restricting their estimator to belong to this set as the

$E[\varepsilon_i|z_i, v_i] = E[y_i - f_0(x_i, z_{if})|z_i, v_i]$ – and the structural function – f_0 – by restricting both functions to fall within the class of bounded and smooth functions such as the Hölder class, and by imposing a slower growth rate on the number of approximating functions on both the structural and the CF. Our approach is thus similar to Ai and Chen (2003).

Given our regularization approach our estimator shares the strength of both the non-parametric CF and the sieve NPIV estimator in that it is easy to implement (similar to Ai and Chen 2003). For example, in the case of sieve estimation using finite dimensional sieves our estimator reduces to a series of least squares regressions which makes point estimation simple and approximation of standard errors possible by bootstrap methods.³

Our second contribution is to develop an estimator for non-linear and non-parametric settings that can combine both CMR and CMI restrictions, which cannot be done in either the NPIV nor the non-parametric CF setting. This new “generalized CMR-CF” uses both CMR and CMI restrictions together by allowing the conditional expectation of the structural error to depend on both instruments and control variables. For example this estimator would be applicable for a model with two endogenous regressors with one satisfying a CMR while the other satisfies a CMI restriction. Our estimator in this setting continues to maintain the simplicity of the CMR-CF estimator as it reduces to three-stage ordinary least squares estimators in many settings.

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

We demonstrate the performance of all of these estimators in two sets of Monte Carlos. The first set uses parametric examples motivated by standard simultaneity problems where it is hard to motivate the economics of conditional independence relative to CMRs (see discussion in Appendix A). The second set of Monte Carlos uses the non-parametric specifications of Newey and Powell (2003) and non-parametric estimators.

Section 2 discusses the formal identification results and Section 3 covers estimation. Sections 4 and 5 cover convergence rates and asymptotic normality, respectively. Section 6 includes Monte Carlos and Section 7 concludes.

1.1 Identification in a Simple Setting

We provide a simple transparent example to illustrate identification in a setting much less general than what our setup allows where there is only one endogenous regressor x_i and when $h_0(z_i, v_i)$ is exactly equal to its second order polynomial approximation. With parameters $(a_1, a_2, a_3, a_4, a_5)$ the function $h_0(z_i, v_i)$ is then given by

$$h_0(z_i, v_i) = a_1 z_i + a_2 v_i + a_3 z_i v_i + a_4 z_i^2 + a_5 v_i^2 = \varphi(z_i) + a_2 v_i + a_3 z_i v_i + a_5 v_i^2$$

where $\varphi(z_i)$ denotes any arbitrary function of z_i . The CMR condition implies that $h_0(z_i, v_i)$ satisfies

regularization. They achieve this by imposing bounds on integrals of higher-order derivatives of the function f_0 , and for estimation they impose the same bounds on the sieve approximation, which is implemented by restricting the sieve coefficients (see our Monte Carlos in Section 6 for further details).

³ Ai and Chen (2003) assume both f_0 and the conditional moment functions, $m(z_i, f) \equiv E[y_i - f(x_i, z_{if})|z_i]$ in our non-parametric regression context, belong to a Hölder class, and their regularization is achieved by a method of sieve approximation. This sieve NPIV, proposed by Ai and Chen (2003), can be viewed as a nonparametric two-stage least squares (2SLS) estimator. See also Chen and Christensen (2015) where they propose one solution to the continuing challenge of how to choose the optimal number of sieve terms used during estimation.

$$\begin{aligned} 0 &= E[h_0(z_i, v_i)|z_i] = E[\varphi(z_i)|z_i] + a_2 E[v_i|z_i] + a_3 E[z_i v_i|z_i] + a_5 E[v_i^2|z_i] \\ &= \varphi(z_i) + a_5 E[v_i^2|z_i] \end{aligned}$$

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

$$\begin{aligned} h_0(z_i, v_i) &= h_0(z_i, v_i) - E[h_0(z_i, v_i)|z_i] \\ &= (\varphi(z_i) - \varphi(z_i)) + a_2 v_i + a_3 z_i v_i + a_5 (v_i^2 - E[v_i^2|z_i]) = a_2 v_i + a_3 z_i v_i + a_5 \tilde{v}_{2i} \end{aligned}$$

where $\tilde{v}_{2i} = v_i^2 - E[v_i^2|z_i]$. As long as there is no linear functional relationship between any functions of (x_i, z_{if}) and any linear functions of $v_i, \tilde{v}_{2i}$, and $z_i v_i$, an additive regression of y_i on (x_i, z_{if}) and these three controls is sufficient for identification.

1.2 Testing

As in other CF approaches (e.g. Rivers and Vuong 1988, Smith and Blundell 1986, Wooldridge 2005; Wooldridge and Papke 2008) we can test for exogeneity of x_i . If x_i is exogenous we have

$$E[\varepsilon_i|z_i, v_i] = E[\varepsilon_i|z_i, x_i] = 0$$

so $h_0(z_i, v_i) = 0$ is the zero function. In terms of the simple example above, testing for exogeneity is equivalent to testing whether $a_2 = 0, a_3 = 0$, and $a_5 = 0$ versus the alternative that at least one is non-zero.

1.3 Computation

A strength of our non-parametric sieve estimator is that it will typically be a series of non-parametric regressions that proceeds in three steps. First we obtain controls $\hat{v}_i = x_i - \hat{E}[x_i|z_i]$ from a (possibly) non-parametric regression of x_i on z_i for each endogenous variable (Eq. (2)). Second we directly impose the restriction $E[h(z_i, v_i)|z_i] = 0$ on $h(z_i, \hat{v}_i)$ by demeaning – conditional on z_i – each basis control term entering in the approximation to $h(z_i, \hat{v}_i)$. For example one can use regression to construct estimates of control variates given by $\hat{\tilde{v}}_{ki} = \hat{v}_i^k - E[\hat{v}_i^k|z_i]$ for $k \geq 2$ by estimating $E[\hat{v}_i^k|z_i]$ using regression. The final step is a non-parametric regression of y_i on (x_i, z_{if}) and all of the demeaned CF variables entering the approximation to $h(z_i, \hat{v}_i)$. In our 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 non-parametric regression of y_i on (x_i, z_{if}) with three additive regressors $v_i, \tilde{v}_{2i}$, and $z_i v_i$.

2 Identification

We first present the main idea of our identification approach when only CMRs are available. In Section 2.2 we extend the estimator to settings where both CMR and CMI restrictions are available for identification.

2.1 Identification in Control Function Setting with Conditional Moment Restrictions Only

We show that $f_0(x_i, z_{if})$ and $h_0(z_i, v_i)$ are identified by Eq. (3) with restrictions (5) for the model (1) and (2). Our approach to identification closely follows Newey, Powell, and Vella (1999) and Newey and Powell (2003). We consider pairs of the structural function $f(x_i, z_{if})$ and the CF $h(z_i, v_i)$ that satisfy

$$E[y_i|z_i, v_i] = f(x_i, z_{if}) + h(z_i, v_i) \quad (6)$$

and the CMR condition. Because the conditional expectation is unique with probability one, if there exists another pair $\bar{f}(x_i, z_{if})$ and $\bar{h}(z_i, v_i)$ that satisfies (6), it must be that

$$\Pr(f_0(x_i, z_{if}) + h_0(z_i, v_i) = \bar{f}(x_i, z_{if}) + \bar{h}(z_i, v_i)) = 1. \quad (7)$$

Identification of f_0 and h_0 means we must have $f_0 = \bar{f}$ and $h_0 = \bar{h}$ with probability one whenever (7) holds. Working with differences, we let $\delta(x_i, z_{if}) = f_0(x_i, z_{if}) - \bar{f}(x_i, z_{if})$ 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

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

For the model (1) and (2) Theorem 1 shows that both the structural function and the CF are identified using our CMR-CF approach.

Theorem 1. *For the model (1) and (2) assume (5) holds. If for all $\delta(x_i, z_{if})$ with finite expectation $E[\delta(x_i, z_{if})|z_i] = 0$ implies $\delta(x_i, z_{if}) = 0$ a.s. (completeness rank condition), then (f_0, h_0) is identified.*

Proof. Suppose it is not identified. Then there must exist functions $\bar{f}(x_i, z_{if})$ and $\bar{h}(z_i, v_i)$ such that $\delta(x_i, z_{if}) \neq 0$ and $\kappa(z_i, v_i) \neq 0$ but $\Pr(\delta(x_i, z_{if}) + \kappa(z_i, v_i) = 0) = 1$. It follows that $E[\kappa(z_i, v_i)|z_i] = E[h_0(z_i, v_i)|z_i] - E[\bar{h}(z_i, v_i)|z_i] = 0 - 0 = 0$ from the CMR, so $0 = E[\delta(x_i, z_{if}) + \kappa(z_i, v_i)|z_i] = E[\delta(x_i, z_{if})|z_i]$. The completeness condition then implies $\delta(x_i, z_{if}) = 0$ a.s., and $\delta(x_i, z_{if}) = 0$ and $\kappa(z_i, v_i) = 0$ with probability one is a contradiction. $\square$

The completeness condition is the non-parametric analog of the rank condition for identification in the linear setting. We illustrate the implication of this condition for a parametric setting by letting $f_0(x_i, z_{if}) = \beta'_0 x_i + \beta'_1 z_{if}$. An alternative function is $\bar{f}(x_i, z_{if}) = \bar{\beta}'_0 x_i + \bar{\beta}'_1 z_{if}$ and in this case

$$E[\delta(x_i, z_{if})|z_i] = (\beta_0 - \bar{\beta}_0)' E[x_i|z_i] + (\beta_1 - \bar{\beta}_1)' z_{if},$$

so if z_i satisfies the standard rank condition – i.e. it includes a sufficient number of excluded instruments from z_{if} that are correlated with x_i – then $E[\delta(x_i, z_{if})|z_i] = 0$ implies $\delta(x_i, z_{if}) = 0$, which requires $\beta_0 = \bar{\beta}_0$ and $\beta_1 = \bar{\beta}_1$ so $f_0(x_i, z_{if})$ is identified. $h_0(z_i, v_i)$ is also identified because the conditional expectation $E[y_i|z_i, v_i]$ is non-parametrically identified and $h_0(z_i, v_i) = E[y_i|z_i, v_i] - f_0(x_i, z_{if})$.

Non-parametric estimation of structural objects of interest using control variables often requires a full support condition, that is, that the support of the control variable conditional on the endogenous regressor is equal to the marginal support of the control variable (Imbens and Newey 2009). This full support condition rules out the possibility of (z_i, x_i) being discrete random variables. While we also use control variables, we do not require that they satisfy this full support condition, so our estimator is well-defined whether (z_i, x_i) are discrete or continuous. This difference arises because our CMR-CF approach imposes the identifying restriction (5) directly on the CF $h_0(z, v_i)$, and *not* on the control variable v_i .⁴ When both (z_i, x_i) are discrete with finite supports, a standard rank condition such as the one in Theorem 2.4 of Newey and Powell (2003) satisfies the completeness condition and hence is sufficient for identification of $f_0(x_i, z_{if})$ and $h_0(z_i, v_i)$.⁵

⁴ To see this point, rewrite the control function $h_0(z_i, v_i) = E[\varepsilon_i|z_i, v_i] = E[\varepsilon_i|z_i, x_i] \equiv h_0^*(z_i, x_i)$. Then observe the proof of Theorem 1 to identify $f_0(x_i, z_{if})$ goes through with this definition of control function $h^*(z_i, x_i)$ in place of $h(z_i, v_i)$. Once $f_0(x_i, z_{if})$ is identified, we also identify the control function as $h_0^*(z_i, x_i) = E[y_i|z_i, x_i] - f_0(x_i, z_{if})$. Therefore, as it is illustrated here, our identifying restriction is on the control function, not on the control variable. In this sense our result is in line with Newey and Stouli (2021).

⁵ In our notation Theorem 2.4 of Newey and Powell (2003) states that the completeness condition holds when both x and z_{-f} are discrete with finite support $\{x_1, \dots, x_s\}$ and $\{z_{-f1}, \dots, z_{-ft}\}$ such that the rank of the $s \times t$ matrix $P(z_f)$ with $P(z_f)_{jk} = \Pr(x = x_j | z_{-f} = z_{-fk}, z_f)$ is equal to s with probability one.

2.2 Identification with Both CMR and CMI Restrictions

We consider a data generating process with an insufficient number of CMR instruments to achieve identification, but with additional instruments that satisfy CMI that may lead to identification. Neither the NPIV estimator, nor the Newey, Powell, and Vella (1999) (NPV-CF) estimator, nor the CMR-CF estimator is consistent in the face of this data generating process. We show that it is straightforward to extend the CMR-CF insight – and the implied CMR-CF estimator – to accommodate the restrictions implied by the CMI on the CF. We call our new estimator the “generalized CMR-CF estimator”, and we show our approach is identified as long as a standard rank condition holds.

Recall that we divide our instruments into two groups, the z_{if} that enter the structural function and the $z_{i,-f}$ that do not, and all instruments taken together are denoted by $z_i = (z_{if}, z_{i,-f})$. We further divide the instruments that enter into the structural function $f(x_i, z_{if})$ into two groups – $z_{if} = (z_{if}^1, z_{if}^2)$ – with the first set z_{if}^1 satisfying CMR and the second set z_{if}^2 satisfying CMI. We divide the excluded instruments into $z_{i,-f}^1$, which satisfy CMR, and $z_{i,-f}^2$, which satisfy CMI. We write $z_i^1 = (z_{if}^1, z_{i,-f}^1)$ as those instruments satisfying the CMR and $z_i^2 = (z_{if}^2, z_{i,-f}^2)$ as those instruments satisfying CMI.

We now consider the model (1) and (2) with these new restrictions, which we write as

$$E[\varepsilon_i | z_{if}^1, z_{i,-f}^1] = 0, E[\varepsilon_i | z_{if}^2, z_{i,-f}^2] \neq 0, \quad \text{and} \quad E[\varepsilon_i | z_i, v_i] = E[\varepsilon_i | z_{if}^1, z_{i,-f}^1, v_i]. \quad (8)$$

In terms of these restrictions NPIV is not consistent because $E[\varepsilon_i | z_i^2] \neq 0$ and NPV-CF is not consistent because $E[\varepsilon_i | z_i, v_i] \neq E[\varepsilon_i | v_i]$. Letting $h(z_i^1, v_i) = E[\varepsilon_i | z_i^1, v_i]$ we have under (8) the new expression for (6) combined with the CMR:

$$\begin{aligned} E[y_i | z_i, v_i] &= f(x_i, z_{if}) + h(z_i^1, v_i) \\ E[h(z_i^1, v_i) | z_i^1] &= 0. \end{aligned} \quad (9)$$

For identification it must be that if both (f_0, h_0) and $(\bar{f}, \bar{h})$ satisfy (9),

$$\Pr(f_0(x_i, z_{if}) = \bar{f}(x_i, z_{if}), \quad h_0(z_i^1, v_i) = \bar{h}(z_i^1, v_i)) = 1.$$

Theorems 2 and 3 establish that the identifying restrictions (8) under two different full-rank conditions are sufficient to identify (f_0, h_0) .

Theorem 2 shows identification when the full rank condition of the excluded CMI instruments holds for all endogenous regressors x_i .

Theorem 2. For the model (1) and (2) assume (8) holds. Suppose $\frac{\partial \Pi_0(z_i)}{\partial z_{i,-f}^2}$ has the full rank (i.e. $\text{rank}(\frac{\partial \Pi_0(z_i)}{\partial z_{i,-f}^2}) = \dim(x_i)$) with probability one. Then (f_0, h_0) is identified.

Proof. We prove by contradiction. Suppose it is not identified. Then there must exist functions $\bar{f}(x_i, z_{if})$ and $\bar{h}(z_i^1, v_i)$ satisfying (9) such that $\delta(x_i, z_{if}) = f_0(x_i, z_{if}) - \bar{f}(x_i, z_{if}) \neq 0$ and $\kappa(z_i^1, v_i) = h_0(z_i^1, v_i) - \bar{h}(z_i^1, v_i) \neq 0$ but $\delta(x_i, z_{if}) + \kappa(z_i^1, v_i) = 0$ with probability one (wp1). Let subscripts denote partial derivatives with respect to corresponding terms in the function. Then by taking derivatives w.r.t. $z_{i,-f}^2$ to $\delta(x_i, z_{if}) + \kappa(z_i^1, v_i) = 0$, we obtain $\left(\frac{\partial x_i}{\partial z_{i,-f}^2}\right)' \delta_x(x_i, z_{if}) = 0$ wp1 because $\kappa(z_i^1, v_i)$ is not a function of $z_{i,-f}^2$. Then because $0 = \left(\frac{\partial x_i}{\partial z_{i,-f}^2}\right)' \delta_x(x_i, z_{if}) = \left(\frac{\partial \Pi_0(z_i)}{\partial z_{i,-f}^2}\right)' \delta_x(x_i, z_{if})$ and $\frac{\partial \Pi_0(z_i)}{\partial z_{i,-f}^2}$ has the full rank, we find $\delta_x(x_i, z_{if}) = 0$ wp1. Since $\kappa(z_i^1, v_i)$ is not a function of $z_{i,-f}^2$, we have $0 = \left(\frac{\partial x_i}{\partial z_{if}^2}\right)' \delta_x(x_i, z_{if}) + \delta_{z_{if}^2}(x_i, z_{if})$. It then follows that $\delta_{z_{if}^2}(x_i, z_{if}) = 0$ since we have shown $\delta_x(x_i, z_{if}) = 0$ wp1. Therefore $\delta(x_i, z_{if})$ must be a function

of z_{if}^1 only wp1. Also $E[\kappa(z_i^1, v_i) | z_i^1] = 0$ by construction of $h_0(z_i^1, v_i)$ and $\bar{h}(z_i^1, v_i)$. It follows that $0 = E[\delta(x_i, z_{if}) | z_i^1] + E[\kappa(z_i^1, v_i) | z_i^1] = \delta(x_i, z_{if})$ because $z_{if}^1 \subset z_i^1$. We conclude $\delta(x_i, z_{if}) = 0$ and $\kappa(z_i^1, v_i) = 0$ wp1, which is a contradiction. $\square$

Next we show identification when only a subset of endogenous regressors satisfy the full rank condition with the excluded CMI instruments while the other endogenous regressors satisfy the completeness condition of the CMR instruments. Let $x_i = (x_i^1, x_i^2)$. Suppose for the endogenous regressors x_i^1 the CMR holds with the instruments z_i^1 while for x_i^2 the CMI holds with the instruments z_i^2 . Decompose x_i^1 by $x_i^1 = \Pi_0^1(z_i^1) + v_i^1$ where z_i^1 denotes the CMR instruments and decompose $x_i^2 = \Pi_0^2(z_i^2) + v_i^2$, respectively, such that $E[v_i^1 | z_i^1] = 0$ and $E[v_i^2 | z_i^2] = 0$.⁶ Note that the decomposition $x_i^1 = \Pi_0^1(z_i^1) + v_i^1$ does not imply that x_i^1 does not depend on z_i^2 structurally. It only means that we construct the control v_i^1 such that $\Pi_0^1(z_i^1) = E[x_i^1 | z_i^1]$ holds. Let the controls be $v_i = (v_i^1, v_i^2)$. Then the identifying restrictions (8) state that given our construction of the controls $v_i = (v_i^1, v_i^2)$ the instruments z_i^2 satisfy the CMI while z_i^1 satisfies the CMR.

Theorem 3. For the model (1) with $x_i = (x_i^1, x_i^2)$ assume (8) holds given the controls $v_i = (v_i^1, v_i^2) = (x_i^1 - \Pi_0^1(z_i^1), x_i^2 - \Pi_0^2(z_i^2))$. Suppose (a) $\frac{\partial \Pi_0^2(z_i^2)}{\partial z_{i,-f}^2}$ has the full rank (i.e. $\text{rank}(\frac{\partial \Pi_0^2(z_i^2)}{\partial z_{i,-f}^2}) = \dim(x_i^2)$) with probability one (partial full rank condition); (b) for all $\delta^1(x_i^1, z_{if}^1)$ with finite expectation $E[\delta^1(x_i^1, z_{if}^1) | z_i^1] = 0$ implies $\delta^1(x_i^1, z_{if}^1) = 0$ a.s. (partial completeness condition).⁷ Then (f_0, h_0) is identified.

Proof. We prove by contradiction. Suppose it is not identified. Then there must exist functions $\bar{f}(x_i, z_{if})$ and $\bar{h}(z_i^1, v_i)$ satisfying (9) such that $\delta(x_i, z_{if}) = f_0(x_i, z_{if}) - \bar{f}(x_i, z_{if}) \neq 0$ and $\kappa(z_i^1, v_i) = h_0(z_i^1, v_i) - \bar{h}(z_i^1, v_i) \neq 0$ but $\delta(x_i, z_{if}) + \kappa(z_i^1, v_i) = 0$ wp1. Let the subscripts denote partial derivatives with respect to corresponding terms in the function. First, by taking partial derivatives to $\delta(x_i, z_{if}) + \kappa(z_i^1, v_i) = 0$ w.r.t. the excluded CMI instrument $z_{i,-f}^2$ (holding $z_{if} = (z_{if}^1, z_{if}^2)$, $z_{i,-f}^1$, and $v_i = (v_i^1, v_i^2)$ fixed), we obtain $\left(\frac{\partial x_i^2}{\partial z_{i,-f}^2}\right)' \delta_{x^2}(x_i, z_{if}) = 0$ wp1 because $\kappa(z_i^1, v_i)$ is not a function of $z_{i,-f}^2$ and because $x_i^1 = \Pi_0^1(z_i^1) + v_i^1$ is not a function of $z_{i,-f}^2$ by construction given the control v_i^1 . Then because $0 = \left(\frac{\partial x_i^2}{\partial z_{i,-f}^2}\right)' \delta_{x^2}(x_i, z_{if}) = \left(\frac{\partial \Pi_0^2(z_i^2)}{\partial z_{i,-f}^2}\right)' \delta_{x^2}(x_i, z_{if})$ and because $\frac{\partial \Pi_0^2(z_i^2)}{\partial z_{i,-f}^2}$ has the full rank, we conclude $\delta_{x^2}(x_i, z_{if}) = 0$ wp1. Similarly, by taking partial derivatives to $\delta(x_i, z_{if}) + \kappa(z_i^1, v_i) = 0$ w.r.t. z_{if}^2 , we find $0 = \left(\frac{\partial x_i^2}{\partial z_{if}^2}\right)' \delta_{x^2}(x_i, z_{if}) + \delta_{z_f^2}(x_i, z_{if})$ because $\kappa(z_i^1, v_i)$ is not a function of z_{if}^2 and $x_i^1 = \Pi_0^1(z_i^1) + v_i^1$ is not a function of z_{if}^2 given the control v_i^1 . It then follows that $\delta_{z_f^2}(x_i, z_{if}) = 0$ wp1 since we have shown $\delta_{x^2}(x_i, z_{if}) = 0$ wp1. Then, from $\delta_{x^2}(x_i, z_{if}) = 0$ and $\delta_{z_f^2}(x_i, z_{if}) = 0$ wp1, we conclude that $\delta(x_i, z_{if})$ must be a function of (x_i^1, z_{if}^1) only wp1. Therefore we can write $\delta^1(x_i^1, z_{if}^1) \equiv \delta(x_i, z_{if})$. We then find $0 = E[\delta^1(x_i^1, z_{if}^1) | z_i^1] + E[\kappa(z_i^1, v_i) | z_i^1] = E[\delta^1(x_i^1, z_{if}^1) | z_i^1]$ because $E[\kappa(z_i^1, v_i) | z_i^1] = 0$. We then conclude $\delta^1(x_i^1, z_{if}^1) \equiv \delta(x_i, z_{if}) = 0$ by the partial completeness condition (b) and thus $\kappa(z_i^1, v_i) = 0$ wp1, which is a contradiction. $\square$

We now develop the CMR-CF estimator and the generalized CMR-CF estimator.

⁶ The identification result in Theorem 3 holds with the decomposition $x_i^2 = \Pi_0^2(z_i^2) + v_i^2$ and $E[v_i^2 | z_i^2] = 0$ also as long as the full rank condition of the (excluded) CMI instruments $z_{i,-f}^2$ holds w.r.t. x_i^2 .

⁷ If both (z_i^1, x_i^1) are discrete with finite supports, again a standard rank condition like that of Theorem 2.4 in Newey and Powell (2003) trivially satisfies the completeness condition.

3 Estimation

Our estimator is a three-step estimator. The first two steps are ordinary least squares given the sieve approximations and the final step is non-parametric least squares. We first develop the estimator for the model (1) and (2) when all instruments satisfy the CMRs. In Section 3.2 we then show how the estimator is modified when only a subset of instruments satisfy the CMRs as in Section 2.2. For such settings, compared to estimation when all instruments satisfy the CMRs, the only requirement is to drop from the CF those instruments that satisfy conditional independence.

3.1 CMR-CF Estimator

We focus on sieve estimation using finite dimensional sieves 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 $i = 1, \dots, n$ are i.i.d. We let X_i be $d_x \times 1$, Z_{if} be $d_{z_f} \times 1$, Z_{i-f} be $d_{z_{-f}} \times 1$, $d_z = d_{z_f} + d_{z_{-f}}$ and $d = d_z + d_x$, with $d_x = 1$ for ease of exposition. Let $\{p_j(Z), j = 1, 2, \dots\}$ denote a sequence of approximating basis functions (e.g. orthonormal polynomials or splines). Let $p^{k_n} = (p_1(Z), \dots, p_{k_n}(Z))'$, $P = (p^{k_n}(Z_1), \dots, p^{k_n}(Z_n))'$, and $(P'P)^{-}$ denote the Moore–Penrose generalized inverse, where k_n tends to infinity but $k_n/n \rightarrow 0$. Similarly we let $\{\phi_j(X, Z_f), j = 1, 2, \dots\}$ denote a sequence of approximating basis functions, $\phi^{K_n} = (\phi_1(X, Z_f), \dots, \phi_{K_n}(X, Z_f))'$, where K_n tends to infinity but $K_n/n \rightarrow 0$.⁸

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

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

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 $\{\varphi_l(z, v), l = 1, 2, \dots\}$ 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

$$\mathcal{H}_n = \left\{ h : h = \sum_{l \leq L_n} a_l \tilde{\varphi}_l(z, v), \|h\|_{\mathcal{H}} < \bar{c}_h, (z, v) \in \mathcal{W} \right\}$$

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

Because v is not known we use estimates of the approximating basis functions, which we denote as $\hat{\varphi}_l(z, \hat{v}) = \varphi_l(z, \hat{v}) - \hat{\varphi}_l(z)$, where $\hat{\varphi}_l(z) = \hat{E}[\varphi_l(Z, \hat{V})|Z = z]$. We then construct the approximation of $h(z, v)$ as⁹

⁸ 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.

⁹ 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 ease of notation.

$$\begin{aligned}\hat{h}_{L_n}(z, \hat{v}) &= \sum_{l=1}^{L_n} a_l \{ \varphi_l(z, \hat{v}) - \hat{E}[\varphi_l(Z, \hat{V})|Z = z] \} \\ &= \sum_{l=1}^{L_n} a_l \left\{ \varphi_l(z, \hat{v}) - p^{k_n}(z)' (P'P)^{-} \sum_{i=1}^n p^{k_n}(z_i) \varphi_l(z_i, \hat{v}_i) \right\},\end{aligned}\quad (10)$$

with coefficients $(a_1, \dots, a_{L_n})$ to be estimated in the last step. We approximate the sieve space $\mathcal{H}_n$ with $\hat{\mathcal{H}}_n$ using (10), so $\hat{\mathcal{H}}_n$ is given by

$$\hat{\mathcal{H}}_n = \left\{ h: h = \sum_{l \leq L_n} a_l \hat{\varphi}_l(z, \hat{v}), \|h\|_{\mathcal{H}} < \bar{C}_h, (z, \hat{v}) \in \mathcal{W} \right\}. \quad (11)$$

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

$$\mathcal{F}_n = \left\{ f: f = \sum_{l \leq K_n} \beta_l \phi_l(x, z_f), \|f\|_{\mathcal{F}} < \bar{C}_f, (x, z_f) \in \text{supp}(X, Z_f) \right\}$$

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

$$(\hat{f}, \hat{h}) = \underset{(f, h) \in \mathcal{F}_n \times \hat{\mathcal{H}}_n}{\operatorname{arginf}} \sum_{i=1}^n \{y_i - (f(x_i, z_{if}) + h(z_i, \hat{v}_i))\}^2 / n \quad (12)$$

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

Equivalently we can write the last step of estimation as

$$\min_{(\beta_1, \dots, \beta_{K_n}, a_1, \dots, a_{L_n})} \sum_{i=1}^n \left\{ y_i - \left(\sum_{k=1}^{K_n} \beta_k \phi_k(x_i, z_{if}) + \sum_{l=1}^{L_n} a_l \hat{\varphi}_l(z_i, \hat{v}_i) \right) \right\}^2 / n. \quad (13)$$

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 5).

3.2 Generalized CMR-CF Estimator

To estimate the models in Section 2.2 we have only to replace the CF and its approximation with suitable ones that satisfy the underlying identification conditions. We use the CF $h(z^1, v) = h(z_f^1, z_{-f}^1, v)$ and its approximation as

$$\begin{aligned}\hat{h}_{L_n}(z^1, \hat{v}) &= \sum_{l=1}^{L_n} a_l \hat{\varphi}_l(z^1, \hat{v}) \\ &= \sum_{l=1}^{L_n} a_l \left\{ \varphi_l(z^1, \hat{v}) - \hat{E} \left[\varphi_l(Z^1, \hat{V}) | Z^1 = z^1 \right] \right\} \\ &= \sum_{l=1}^{L_n} a_l \left\{ \varphi_l(z^1, \hat{v}) - p_1^{k_n}(z^1)' (P_1' P_1)^{-} \sum_{i=1}^n p_1^{k_n}(z_i^1) \varphi_l(z_i^1, \hat{v}_i) \right\}\end{aligned}$$

where $p_1^{k_n}(z^1) = (p_1(z^1), \dots, p_{k_n}(z^1))'$ and $P_1 = (p_1^{k_n}(z_1^1), \dots, p_1^{k_n}(z_n^1))'$, such that 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 CF as

$$(\hat{\beta}, \hat{a}) = \underset{(\beta_1, \dots, \beta_{K_n}, a_1, \dots, a_{L_n})}{\operatorname{argmin}} \sum_{i=1}^n \left\{ y_i - \left(\sum_{k=1}^{K_n} \beta_k \phi_k(x_i, z_{if}) + \sum_{l=1}^{L_n} a_l \hat{\varphi}_l(z_i^1, \hat{v}_i) \right) \right\}^2 / n$$

such that $\hat{f}(x, z_f) = \sum_{k=1}^{K_n} \hat{\beta}_k \phi_k(x, z_f)$ and $\hat{h}(z^1, v) = \sum_{l=1}^{L_n} \hat{a}_l \hat{\varphi}_l(z^1, v)$ become the estimators of f_0 and h_0 , respectively, for the models in Section 2.2.

3.3 Using Cross-Validation to Choose Specifications

Cross-validation seeks to balance bias and variance in a particular dataset. We consider a cross-validation approach that can be used to select the best specification among a set of alternative specifications and approximations for our estimators. Given the sieve approximating functions this amounts to how to select k_n , K_n , and L_n in the estimation (13). We minimize the mean squared error (MSE) of the final estimator by comparing the out-of-sample fit (on hold-out samples) of candidate models estimated using training samples. Given a candidate specification $c \in C$ where C is the collection of alternative specifications to be considered for the approximations in various stages of our estimation (such as choosing k_n , K_n , and L_n), we can approximate the MSE of the fit using

$$CV_n(c) = \frac{1}{n} \sum_{i=1}^n \{y_i - (\hat{f}_{c,-i}(x_i, z_{if}) + \hat{h}_{c,-i}(z_i, v_i))\}^2$$

where $(\hat{f}_{c,-i}, \hat{h}_{c,-i})$ denotes the so-called “leave-one-out” estimator obtained by estimating (12) under the specification c and using the observations excluding i . Cross-validation’s “optimal” model specification is the one that solves $\min_{c \in C} CV_n(c)$.

4 Convergence Rates

In Sections 4 and 5 we develop asymptotic properties of the estimator for the model (1), (2) and (5). Similar results can be obtained for the estimator of the generalized models in Section 2.2 with minor modifications. We obtain the convergence rates building on Newey, Powell, and Vella (1999). We differ from their approach as we have another non-parametric estimation stage in the middle step of estimation that creates additional terms in the convergence rate results.

We introduce additional notation. We write $g_0(z_i, v_i) \equiv g_0(x_i, z_i, v_i) = f_0(x_i, z_{if}) + h_0(z_i, v_i)$ for ease of notation. For a random matrix D , let $\|D\| = (\text{tr}(D'D))^{1/2}$, and let $\|D\|_\infty$ be the infimum of constants C such that $\Pr(\|D\| < C) = 1$. We derive the convergence rates of the non-parametric 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 2 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 $\text{var}(X|Z)$, $\text{var}(Y|Z, V)$, and $\text{var}(\varphi_l(Z, V)|Z)$ 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) $f_0(X, Z_f)$ is continuously differentiable of order s and all the derivatives of order s are bounded on the support of (X, Z_f) .

Assumptions C1 (iii), (iv), (v), and (vii) ensure that the unknown functions $\Pi_0(Z)$, $\bar{\varphi}_l(Z)$, $h_0(Z, V)$, and $f_0(X, Z_f)$ 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 $O(K_n^{-s/(d_x+d_{z_f})})$ respectively when the finite dimensional sieves such as polynomials or splines approximation are used (see Chen 2007; Newey 1997; Schumaker 1981; Timan 1963). In particular these Assumptions C1 (v) and (vii) that restrict the space of functions for f_0 and h_0 are critical to circumvent

the potential ill-posed problem. Moreover, these assumptions justify the use of the finite dimensional sieves as a regularization method (see Chen (2007) for related discussions).

Assumption C1 (vi) is satisfied for polynomial and spline basis functions with appropriate orders. Assumptions C1 (i) and (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 $\mathbf{L}_n = K_n + L_n$. These rate conditions in Assumption C2 below are derived in the proof of Theorem 4 such that approximation errors of various stages of non-parametric regressions become negligible as the sample size gets large. In particular $\Delta_{n,1}$ controls the approximation error rate of $\hat{\Pi}(z)$ in the first stage and $\Delta_{n,2}$ controls the approximation error rate of $\hat{\varphi}_1(z)$ in the second stage. Assumption C2 implies that for example, for polynomial approximations, if k_n , K_n , and L_n grow at the same rate, they should grow slower than $n^{1/8}$.

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

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

Theorem 4. For the model (1), (2) and (5) suppose Assumptions C1 and C2 are satisfied. Then

$$(a) \left(\int [\hat{g}(z, v) - g_0(z, v)]^2 d\mu_0(z, v) \right)^{1/2} = O_p \left(\sqrt{\mathbf{L}_n/n} + L_n \Delta_n + \mathbf{L}_n^{-s/d} \right), \text{ and}$$

$$(b) |\hat{f}(x, z_f) - f_0(x, z_f)|_\delta = O_p(\zeta_\delta(K_n) \left[\sqrt{\mathbf{L}_n/n} + L_n \Delta_n + \mathbf{L}_n^{-s/d} \right] + K_n^{-s/(d_x + d_{z_f})})$$

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

In the convergence rate results of Theorem 4 the term $L_n \Delta_n$ arises because of the estimation errors in the first and second steps of estimation. With no estimation error from these stages we would obtain standard convergence rates of series estimators.

5 Asymptotic Normality

Following Newey (1997) and Newey, Powell, and Vella (1999) we consider inference for the linear functionals of g , $\theta = \alpha(g)$.¹⁰ 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

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

where we let $\hat{\beta} = (\hat{\beta}_1, \dots, \hat{\beta}_{K_n}, \hat{a}_1, \dots, \hat{a}_{L_n})'$ that solves (13) with some abuse of notation. 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_f . For

¹⁰ This includes θ 's being functionals of f only.

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.¹¹

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 $v^*(Z, V)$ such that

$$\alpha(g) = E[v^*(Z, V)g(Z, V)] \quad (14)$$

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

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

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. Chen 2007; Newey 1997; Newey, Powell, and Vella 1999). Defining $\rho_v(Z) = E[v^*(Z, V)(\frac{\partial h_0(Z, V)}{\partial V} - E[\frac{\partial h_0(Z, V)}{\partial V}|Z])|Z]$ and $\rho_{\varphi_l}(Z) = E[a_l v^*(Z, V)|Z]$, the asymptotic variance of the estimator $\hat{\theta}$ is given by

$$\begin{aligned} \Omega &= E[v^*(Z, V)\text{var}(Y|Z, V)v^*(Z, V)'] + E[\rho_v(Z)\text{var}(X|Z)\rho_v(Z)'] \\ &\quad + \lim_{n \rightarrow \infty} \sum_{l=1}^{L_n} E[\rho_{\varphi_l}(Z)\text{var}(\varphi_l(Z, V)|Z)\rho_{\varphi_l}(Z)']. \end{aligned} \quad (15)$$

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 $\tilde{\varphi}^L(z, v) \equiv (\tilde{\varphi}_1(z, v), \dots, \tilde{\varphi}_L(z, v))'$ and $\psi^L(z, v) \equiv (\phi^K(x, z_f)', \tilde{\varphi}^L(z, v))'$.

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

Assumption R1 restricts the class of linear functionals we consider, which yields the $\sqrt{n}$ -consistency and also requires $v^*(Z, V)$ be well approximated by the approximating basis functions we use to approximate g_0 . To present the theorem, we need additional notation and assumptions. Let $a_L = (a_1, \dots, a_L)'$ with an abuse of notation.

Assumption 4. (N1) (i) There exist δ, γ , and β_L such that $|g_0(z, v) - \beta_L' \psi^L(z, v)|_\delta \leq CL^{-\gamma}$ (which also implies $|h_0(z, v) - a_L' \tilde{\varphi}^L(z, v)|_\delta \leq CL^{-\gamma}$); (ii) $\text{var}(Y|Z, V)$ is bounded away from zero, $E[\eta^4|Z, V]$ and $E[V^4|Z]$ are bounded and $E[\tilde{\varphi}_l(Z, V)^4|Z]$ 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 appropriate central limit theorems. Next we impose the rate conditions that restrict the growth of k_n and $L_n = K_n + L_n$ as n tends to infinity. These rate conditions in Assumption N2 below are derived in the proof of Theorem 5 such that approximation errors of various stages of non-parametric regressions become small enough to justify the stochastic expansions

¹¹ More generally, if $\mathcal{A}$ depends on unknown population objects, we can estimate it using $\hat{\mathcal{A}} = \partial \alpha(\hat{\psi}_i^L \beta) / \partial \beta' |_{\beta = \hat{\beta}}$ where $\hat{\psi}_i^L = (\phi_1(x_i, z_{if}), \dots, \phi_K(x_i, z_{if}), \hat{\varphi}_1(z_i, \hat{v}_i), \dots, \hat{\varphi}_L(z_i, \hat{v}_i))'$, so that $\hat{\theta} = \hat{\mathcal{A}} \hat{\beta}$ (see Newey (1997)).

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 obtain and alternative assumptions on the functionals of interest (see Newey, Powell, and Vella 1999).

to obtain the asymptotic normal distribution. Assumption N2 requires that for example, for polynomial approximations, if k_n , K_n , and L_n grow at the same rate, they should grow slower than $n^{1/12}$.

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

Theorem 5. Suppose Assumptions C1, R1, and N1–N2 are satisfied. Then

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

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

$$\begin{aligned} \hat{\tau} &= \sum_{i=1}^n \hat{\psi}_i^L \hat{\psi}_i^{L'} / n, \hat{\Sigma} = \sum_{i=1}^n (y_i - \hat{g}(z_i, \hat{v}_i))^2 \hat{\psi}_i^L \hat{\psi}_i^{L'} / n \\ \hat{\tau}_1 &= P'P/n, \hat{\Sigma}_1 = \sum_{i=1}^n \hat{v}_i^2 p^k(z_i) p^k(z_i)' / n, \hat{\Sigma}_{2,l} = \sum_{i=1}^n \{ \varphi_l(z_i, \hat{v}_i) - \hat{\varphi}_l(z_i) \}^2 p^k(z_i) p^k(z_i)' / n \\ \hat{H}_{11} &= \sum_{i=1}^n \sum_{l=1}^L \hat{a}_l \frac{\partial \varphi_l(z_i, \hat{v}_i)}{\partial v_i} \hat{\psi}_i^L p^k(z_i)' / n, \\ \hat{H}_{12} &= \sum_{i=1}^n p^k(z_i)' \left((P'P)^{-1} \sum_{j=1}^n p^k(z_j) \frac{\partial \sum_{l=1}^L \hat{a}_l \varphi_l(z_j, \hat{v}_j)}{\partial v_j} \right) \hat{\psi}_i^L p^k(z_i)' / n, \\ \hat{H}_{2,l} &= \sum_{i=1}^n \hat{a}_l \hat{\psi}_i^L p^k(z_i)' / n, \hat{H}_1 = \hat{H}_{11} - \hat{H}_{12}. \end{aligned} \quad (16)$$

Then, we can estimate Ω consistently by

$$\hat{\Omega} = \mathcal{A} \hat{\tau}^{-1} \left[\hat{\Sigma} + \hat{H}_1 \hat{\tau}_1^{-1} \hat{\Sigma}_1 \hat{\tau}_1^{-1} \hat{H}_1' + \sum_{l=1}^{L_n} \hat{H}_{2,l} \hat{\tau}_1^{-1} \hat{\Sigma}_{2,l} \hat{\tau}_1^{-1} \hat{H}_{2,l}' \right] \hat{\tau}^{-1} \mathcal{A}'. \quad (17)$$

Theorem 6. Suppose Assumptions C1, R1, and N1–N2 are satisfied. Then $\hat{\Omega} \rightarrow_p \Omega$.

This $\hat{\Omega}$ is the heteroskedasticity robust variance estimator that accounts for the first and second steps of estimation. The first variance term $\mathcal{A} \hat{\tau}^{-1} \hat{\Sigma} \hat{\tau}^{-1} \mathcal{A}'$ 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 (15)). The third variance term accounts for the estimation of $\hat{\varphi}_l(\cdot)$'s (and corresponds to the third term in (15)). 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. Akerberg, Chen, and Hahn 2012; Newey 1984; Murphy and Topel 1985).

5.1 Discussion

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

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

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

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

where $v^*(z, v) = (E[q(Z, V)q(Z, V)'])^{-1}q(z, v)$ 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_f) 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 $\psi_{-1}^L(z_i, v_i) \equiv (\phi_1(x_{-1i}, z_{if}), \dots, \phi_K(x_{-1i}, z_{if}), \tilde{\varphi}^L(z_i, v_i)')'$, and then use these estimates to approximate $v^*(z, v)$.

Next we consider a weighted average derivative of the form

$$\alpha(g_0) = \int_{\tilde{\mathcal{W}}} \varpi(z, v) \frac{\partial g_0(z, v)}{\partial x} dz dv = \int_{\tilde{\mathcal{W}}} \varpi(z, v) \frac{\partial f_0(x, z_f)}{\partial x} dz dv$$

where the weight function $\varpi(z, v)$ puts zero weights outside $\tilde{\mathcal{W}} \subset \mathcal{W}$. This is a linear functional of g_0 . Following a similar argument to Newey, Powell, and Vella (1999), integration by parts shows that

$$\alpha(g_0) = - \int_{\tilde{\mathcal{W}}} \text{proj} \left(\tau_0(z, v)^{-1} \frac{\partial \varpi(z, v)}{\partial x} | S \right) g_0(z, v) \tau_0(z, v) dz dv = E[v^*(Z, V)g(Z, V)]$$

where $\tau_0(z, v)$ denotes the density of (z, v) and $\text{proj}(\cdot | S)$ denotes the mean-square projection on the space of functions that are additive in functions of (x, z_f) and any $h(z, v)$ such that $E[h(Z, V)|Z] = 0$, and $v^*(z, v) = -\text{proj}(\tau_0(z, v)^{-1} \frac{\partial \varpi(z, v)}{\partial x} | S)$ (so the Riesz representer $v^*(z, v)$ is well-defined). We can then approximate $v^*(z, v)$ using a mean-square projection of $-\tau_0(z, v)^{-1} \frac{\partial \varpi(z, v)}{\partial x}$ on $\psi^L(z, v)$.

6 Monte Carlos

We conduct two sets of Monte Carlos to evaluate the performance of our proposed CMR-CF estimator. The first Monte Carlos have a structural function $f(x)$ that is parametric. The exact specifications are motivated by standard simultaneity problems that arise in economics (see Appendix A). The second set of Monte Carlos uses the nonlinear setup from Newey and Powell (2003) and the structural function is estimated non-parametrically.

6.1 Simultaneity-Based Monte Carlos

In the first setting $f(x)$ is known up to a finite set of parameters. The dependent variable is denoted by y_i , the endogenous regressor is given by x_i , and the instrument is given by z_i . The six parametric designs we consider are

- [1] $y_i = \alpha + \beta x_i + \gamma x_i^2 + \varepsilon_i; \quad x_i = z_i + (3\varepsilon_i + \varsigma_i) \cdot \log(z_i)$
- [2] $y_i = \alpha + \beta x_i + \gamma x_i^2 + \varepsilon_i; \quad x_i = z_i + (3\varepsilon_i + \varsigma_i) / \exp(z_i)$
- [3] $y_i = \alpha + \beta x_i + \gamma \log x_i + \varepsilon_i; \quad x_i = z_i + (3\varepsilon_i + \varsigma_i) / \exp(z_i)$

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

$$[4] \quad y_i = \alpha + \beta x_i + \gamma \log x_i + \varepsilon_i; \quad x_i = z_i + (3\varepsilon_i + \zeta_i + \varepsilon_i \cdot \zeta_i) / \exp(z_i)$$

$$[5] \quad y_i = \alpha + \beta x_i + \varepsilon_i; \quad x_i = z_i + (3\varepsilon_i + \zeta_i) / \exp(z_i)$$

$$[6] \quad y_i = \alpha + \beta x_i + \gamma x_i^2 + \varepsilon_i; \quad x_i = z_i + (3\varepsilon_i + \zeta_i),$$

which are motivated by simultaneity.¹⁴

We generate simulation data based on the following distributions: $\varepsilon_i \sim U_\varepsilon$, $\zeta_i \sim U_\zeta$, $z_i = 2 + 2U_z$, where each U_ε , U_ζ , and U_z independently follows the uniform distribution supported on $[-1/2, 1/2]$, so all three random variables ε_i , ζ_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. We set the true parameter values at $(\alpha_0, \beta_0, \gamma_0) = (1, 1, -1)$ and data are generated with the sample size of $n = 1,000$.

We compare three different estimators: the linear CF estimator that uses the linear control, the Newey, Powell, and Vella (1999)'s control function (NPV-CF) estimator, and our CMR-CF estimator. All estimators we compare are based on a first stage estimation residual $\hat{v}_i = x_i - (\hat{\pi}_0 + \hat{\pi}_1 z_i + \hat{\pi}_2 z_i^2)$ although results are robust to adding higher order terms.¹⁵ The linear CF approach just includes the control in an additive manner. Our analysis begins with this linear CF 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 order 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 the conditional mean function $\hat{E}[\cdot|z_i]$ is estimated using least squares with regressors $(1, z_i, z_i^2)$. Interactions with polynomials of z_i such as $z_i \hat{v}_i$ and $z_i^2 \hat{v}_i$ are defined similarly. In the last step the CMR-CF estimator estimates the parameters as

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

where the approximated CF $h(z_i, \hat{v}_i) = \sum_{l=1}^L a_l \tilde{v}_{li}$ depends on each simulation design.

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. For each design we determine the set of control variables by starting with a smaller set of relatively simple control variables – the first stage residual and its interaction with the instrument – and then adding higher order terms to reduce the bias of the CMR-CF estimator up to the point where the mean-squared error does not noticeably improve.

Our selection of controls for each design is as follows. In the design [1] we use $\tilde{v}_{1i}$ and $z_i \tilde{v}_i$ as the controls. In the designs [2], [5], and [6] we use the controls $\tilde{v}_{1i}$, $\tilde{v}_{2i}$, and $z_i \tilde{v}_i$. In the 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 the design [4] we use $\tilde{v}_{1i}$, $\tilde{v}_{2i}$, $\tilde{v}_{3i}$, $\tilde{v}_{4i}$, $z_i \tilde{v}_i$.

We report the bias and the root mean-squared error (RMSE) based on 200 replications. The results in Tables 1–6 show that our CMR-CF estimator is consistent in all six designs. In contrast the linear CF and

¹⁴ For example, in terms of the notation in Appendix A, 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 $\eta_1 = 3\varepsilon + \zeta$ is a noisy signal of ε . The selection equation (21) is then $x = \frac{\varphi_1 - c_1(z, \eta_1)}{c_2(z, \eta_2) - \varphi_2} = z + (3\varepsilon + \zeta) \cdot \log(z)$. The other designs can be derived in a similar way.

¹⁵ RMSEs 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 specifications we would be indifferent between this simplest specification and the ones with the higher order terms.

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

Nonlinear & the CF condition (4) does not hold				
		Mean	Bias	RMSE
Linear CF	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 2: Design [2], $\alpha_0 = 1, \beta_0 = 1, \gamma_0 = -1$.

Nonlinear & the CF condition (4) does not hold				
		Mean	Bias	RMSE
Linear CF	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 3: Design [3], $\alpha_0 = 1, \beta_0 = 1, \gamma_0 = -1$.

Nonlinear & the CF condition (4) does not hold				
		Mean	Bias	RMSE
Linear CF	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

the NPV-CF are biased in all designs except [5] and [6], the two designs for which theory says they should be consistent. In design [5] all three approaches produce correct estimates because the outcome equation is linear, which is consistent with our discussion in Appendix B. In design [6] all three approaches are consistent because the CMI restriction (4) holds.

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

Nonlinear & the CF condition (4) does not hold				
		Mean	Bias	RMSE
Linear CF	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 5: Design [5], $\alpha_0 = 1, \beta_0 = 1$.

Linear & the CF condition (4) does not hold				
		Mean	Bias	RMSE
Linear CF	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 6: Design [6], $\alpha_0 = 1, \beta_0 = 1, \gamma_0 = -1$.

Nonlinear & the CF condition (4) holds				
		Mean	Bias	RMSE
Linear CF	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

6.2 Monte Carlos Based on Non-parametric Models

In this setting we conduct two simulation studies where we estimate the structural function $f(x)$ non-parametrically. Design [A] has a first stage selection equation that satisfies both the CF restriction (4) and the CMR. Design [B] does not satisfy the CF restriction (4).

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

$$y_i = f(x_i) + \varepsilon_i = \ln(|x_i - 1| + 1) \operatorname{sgn}(x_i - 1) + \varepsilon_i$$

$$[A] \quad x_i = z_i + \eta_i$$

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 $\rho = 0.5$ being the source of endogeneity. This design satisfies the CF 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 both $\rho = 0.5$ and $\rho = 0.9$. The CF 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

$$f(x) \approx x\beta + \sum_{j=1}^J \gamma_j \exp(-x^2)x^{j-1}. \quad (18)$$

We compare four different estimators: non-parametric least squares, Newey, Powell, and Vella (1999)'s control function (NPV-CF) estimator, Newey and Powell (2003)'s NPIV estimator, and our CMR-CF estimator. We fix $J = 5$ for the design [A] and $J = 7$ for the design [B] and we use five different sample sizes ($n = 100, 250, 400, 1000$, and 2000). In all designs we obtain the control variable using the first stage regression residual $\hat{v}_i = x_i - (\hat{\pi}_0 + \hat{\pi}_1 z_i + \hat{\pi}_2 z_i^2)$.¹⁶ We experimented with different choices of approximating functions for $h(z, v)$ for design [B]. To estimate the CF we use the approximation $\sum_{l=1}^5 a_l v^l$ for the NPV-CF estimator and use three different approximations for the CMR-CF estimator: $h(z, v) = a_1 v + a_2 z v$ (for CMR-CF1), $h(z, v) = a_1 v + a_2 z v + a_3 z^2 v$ (for CMR-CF2), and $h(z, v) = a_1 v + a_2 z v + a_3 z^2 v + a_4 \hat{v}_2$ (for CMR-CF3).

The results are summarized in Tables A and B. We report the 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 of non-parametric least squares are larger than those of other estimators that correct for endogeneity. In the design [A] as expected all three estimators NPV-CF, NP-IV, and CMR-CF estimators perform similarly in terms of RMSE.¹⁷

In the design [B] both NP-IV and CMR-CF estimators dominate the NPV-CF estimator in terms of RMSE. The CMR-CF2 shows the smallest RMSE among all CMR-CF estimators. Comparison between the CMR-CF1 and CMR-CF2 suggests that including the term $z^2 v$ significantly reduces RMSE in the larger sample sizes. The RMSEs of non-parametric least squares and NPV-CF estimators tend to increase as the degree of endogeneity increases while RMSEs decrease in NP-IV, CMR-CF2, and CMR-CF3 as the degree of endogeneity increases. For the design [B] the RMSEs of the NP-IV and the CMF-CF estimators are comparable although the NP-IV estimator dominates the CMR-CF estimator with smaller samples while the opposite is true for larger samples.

In our Monte Carlos we varied sample sizes. Although it is hard to gauge a convergence rate of the estimator from these results, the observed patterns of RMSE clearly show the convergence of the CMR-CF estimator as the RMSE decreases with the sample size. The results suggest that when the sample size is quadrupled, the RMSE decreases by less than half, so in terms of convergence rate at a less than $\sqrt{n}$ -rate, as we would normally expect for non-parametric estimators.

¹⁶ We experimented with adding several higher order terms in the first stage and found very similar simulation results across all estimators.

¹⁷ Our design [A] is exactly the same with the one in Newey and Powell (2003), so we can directly compare the RMSEs of our CMR-CF estimator with those of the NP-IV reported in Table 1 of Newey and Powell (2003) for $n = 100$ and 400 . For the comparison we reproduce their Monte Carlos results as close as we can in Table A and then experiment with different sample sizes. The NP-IV is somewhat sensitive to the choice of constraint parameter for regularization. For $n = 1000$ and 2000 , we experiment with this tuning parameter and choose one that tends to minimize the RMSE of the NP-IV estimator. With this selection the RMSEs are comparable between the two estimators as reported in Table A.

Table A: RMSE comparison of non-parametric estimators (Design [A]).

Control function	NP LS	NPV-CF $h(v) \approx a_1 \hat{v}$	NP-IV	CMR-CF $h(z, v) \approx a_1 \hat{v} + a_2 z \hat{v}$
$n = 100$	0.4121	0.2685	0.2610	0.2732
$n = 250$	0.4085	0.1891	0.2224	0.1913
$n = 400$	0.3844	0.1667	0.2054	0.1692
$n = 1000$	0.3788	0.1308	0.1378	0.1317
$n = 2000$	0.3695	0.1149	0.1281	0.1165

We compare the RMSEs based on 500 replications of four estimators: Non-parametric least squares, Newey, Powell, and Vella (1999)'s CF estimator, Newey and Powell (2003)'s non-parametric IV estimator, and the CMR-CF estimator.

Table B: RMSE comparison of non-parametric estimators (Design [B]).

Control function	NP LS	NPV-CF $\sum_{l=1}^5 a_l \hat{v}^l$	NP-IV	CMR-CF1 $a_1 \hat{v} + a_2 z \hat{v}$	CMR-CF2 $a_1 \hat{v} + a_2 z \hat{v} + a_3 z^2 \hat{v}$	CMR-CF3 $a_1 \hat{v} + a_2 z \hat{v} + a_3 z^2 \hat{v} + a_4 \hat{v}_2$
$\rho = 0.5$						
$n = 100$	0.3896	0.3241	0.2905	0.3049	0.3104	0.3231
$n = 250$	0.3130	0.2181	0.1866	0.2036	0.1976	0.2034
$n = 400$	0.2775	0.1671	0.1515	0.1540	0.1422	0.1456
$n = 1000$	0.2511	0.1358	0.1063	0.1190	0.0999	0.1014
$n = 2000$	0.2440	0.1156	0.0892	0.0968	0.0737	0.0745
$\rho = 0.9$						
$n = 100$	0.5277	0.3018	0.2554	0.2833	0.2734	0.2889
$n = 250$	0.4834	0.2303	0.1851	0.2036	0.1811	0.1872
$n = 400$	0.4462	0.2003	0.1422	0.1680	0.1296	0.1322
$n = 1000$	0.4375	0.1865	0.1066	0.1483	0.0941	0.0950
$n = 2000$	0.4311	0.1773	0.0887	0.1356	0.0690	0.0696

We compare the RMSEs based on 500 replications of four estimators: Non-parametric least squares, Newey, Powell, and Vella (1999)'s CF estimator, Newey and Powell (2003)'s non-parametric IV, and the CMR-CF estimator with three different specifications of the CF (CMR-CF1, CMR-CF2, CMR-CF3).

7 Conclusions

When the endogenous variables enter non-parametrically into the regression equation, standard linear IVs approaches fail. The two existing solutions are the NPIVs estimators, which are based on a set of CMRs, and the CF estimators, which use conditional mean independence (CMI) restrictions. Our first contribution is to show that CMR place shape restrictions on the conditional expectation of the error given the instruments and endogenous variables that are sufficient for identification, and we use this insight to develop our new CMR-CF estimator. Our second contribution is to develop an estimator for non-linear and non-parametric settings that can combine both CMR and CMI restrictions, which cannot be done in either the NPIV nor the non-parametric CF setting. This new “Generalized CMR-CF” uses both CMR and CMI restrictions together by allowing the conditional expectation of the structural error to depend on both instruments and control variables. When sieves are used to approximate both the structural function and the CF our estimator reduces to a series of Least Squares regressions. Our Monte Carlos illustrate that our new estimator performs well across several economic settings.

Appendix A: Does $E[\varepsilon_i|z_i, v_i] = E[\varepsilon_i|v_i]$ hold in Models of Demand and Supply?

In this section we show that conditional mean independence can be difficult to motivate common empirical models of demand and supply.

A.1 Returns to Education/Estimation of Production Functions

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 et al. (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

$$x = \arg \max_{\tilde{x}} \{E[f(\tilde{x}) + \varepsilon|z, \eta] - c(\tilde{x}, z, \eta)\}, \quad (19)$$

which leads to the endogeneity problem.

Assuming differentiability the optimal x solves

$$\partial f(x)/\partial x - \partial c(x, z, \eta)/\partial x = 0. \quad (20)$$

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 is independent of ε and η). The optimal x is

$$x = \frac{\varphi_1 - c_1(z, \eta_1)}{c_2(z, \eta_2) - \varphi_2} \quad (21)$$

which is not additively separable in η_1 or η_2 , which means $E[\varepsilon|z, v]$ is not generally equal to $E[\varepsilon|v]$.¹⁸

A.2 Discrete Choice Demand

We consider a buy/not-buy binary choice setting with logit demands and a single product monopolist.¹⁹ We let the latent utilities of consumers be $u_{i0} = \varepsilon_{i0}$ and $u_{i1} = \beta_0 + \beta_1 z_i - \alpha p + \xi + \varepsilon_{i1}$ with $(\varepsilon_{i0}, \varepsilon_{i1})$ i.i.d.

¹⁸ 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 et al. (2008), condition $E[\varepsilon|z, v^*] = E[\varepsilon|v^*]$ does not hold in general unless we restrict $c_1(z, \eta_1)$ and $c_2(z, \eta_2)$.

¹⁹ The idea extends immediately to more general multi-firm and multi-product settings.

extreme value and (z_f, p, ξ) denoting observed characteristics, price, and the unobserved characteristic (to the econometrician). The market share for good 1 is given by

$$s = \frac{\exp(\beta_0 + \beta_1 z_f - \alpha p + \xi)}{1 + \exp(\beta_0 + \beta_1 z_f - \alpha p + \xi)}, \quad (22)$$

which can be linearized as

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

If we let $mc(z_{-f}, \omega)$ denote marginal costs given as a function of a cost shifter z_{-f} and a cost shock ω , then the monopolist chooses price p to maximize the expected profit such that

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

so $E[\xi|p] \neq 0$ in the linearized demand equation. Prices are evidently not separable in ξ and $z = (z_f, z_{-f})$. 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 $E[\xi|z, v] \neq E[\xi|v]$.

Appendix B: Linear Setting with Additive Errors

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

$$y_i = x_i \beta_0 + \varepsilon_i, \quad (23)$$

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

$$E[z_i \varepsilon_i] = 0, \quad E[z_i x_i] \neq 0, \quad (24)$$

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

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

(25) implies the following estimating equation

$$y_i = x_i \beta + \rho v_i + \eta_i, \quad (26)$$

with $\eta_i = \varepsilon_i - E[\varepsilon_i|z_i, x_i]$. The linear CF 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 $E[\varepsilon_i|z_i, v_i]$ being written as 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 linear CF estimator and the two-stage least squares (2SLS) estimator produce numerically identical estimates. Our CMR-CF estimator generally will differ in finite samples from the linear CF/2SLS estimator because of the additional regressors used to approximate $\tilde{h}(z_i, v_i)$. We show that under the IV condition (24) 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. We relegate the proof that the linear CF and

2SLS are numerically equivalent to a footnote below.²⁰ To develop our CMR-CF estimator, write the estimation equation as the generalization of (26) such that²¹

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

where $Y = (y_1, \dots, y_n)'$, $X = (x_1, \dots, x_n)'$, $Z = (z_1, \dots, z_n)'$, $\tilde{H}(Z, \hat{V}) = (\tilde{h}(z_1, \hat{v}_1), \dots, \tilde{h}(z_n, \hat{v}_n))'$ 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

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

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

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

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

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

and therefore consistency holds if

$$\sum_{i=1}^n z_i\tilde{h}(z_i, \hat{v}_i)/n \rightarrow_p 0. \quad (29)$$

Theorem 7 couples (24) with weak regularity conditions which are sufficient for (29) to hold.

Theorem 7. 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 $\sup_{\pi^* \in \Pi_0} E\left[\|z_i\|^2 \left\|\frac{\partial \tilde{h}(z_i, v_i(\pi^*))}{\partial v_i}\right\|\right] < \infty$ for Π_0 some neighborhood of π_0 , (iv) assume $E\left[\|z_i\|^2 \left\|\frac{\partial \tilde{h}(z_i, v_i(\pi))}{\partial v_i}\right\|\right]$ is continuous at $\pi = \pi_0$, and (v) $\hat{\pi} \rightarrow_p \pi_0$. 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 $\sum_{i=1}^n z_i\tilde{h}(z_i, v_i)/n \rightarrow_p E[z_i\tilde{h}(z_i, v_i)]$ by the law of large numbers under (i). Obtain $\|\sum_{i=1}^n z_i(\tilde{h}(z_i, \hat{v}_i) - \tilde{h}(z_i, v_i))/n\| \leq \|\hat{\pi}^* - \pi_0\| \sum_{i=1}^n \|z_i\|^2 \left\|\frac{\partial \tilde{h}(z_i, v_i(\hat{\pi}^*))}{\partial v_i}\right\|/n$ by applying the mean-value expansion, where $\hat{\pi}^*$ lies between

²⁰ We provide a simple proof using projection that shows why they are equivalent. Let $Y = (y_1, \dots, y_n)'$, $X = (x_1, \dots, x_n)'$, $Z = (z_1, \dots, z_n)'$, $\hat{X} = Z\hat{\pi} = (\hat{x}_1, \dots, \hat{x}_n)'$ where $\hat{\pi}$ denotes the least squares estimate in the first stage regression of X on Z , and $\hat{V} = (\hat{v}_1, \dots, \hat{v}_n)'$ below.

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 Y on 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.

²¹ In the linear case “CMR-CF” is a misnomer because – like the linear CF and 2SLS – we only require $E[z_i\epsilon_i] = 0$.

$\hat{\pi}$ and π_0 and $v_i(\pi) = x_i - z_i'\pi$. Then the term $\sum_{i=1}^n z_i(\tilde{h}(z_i, \hat{v}_i) - \tilde{h}(z_i, v_i))/n \rightarrow_p 0$ by the consistency of $\hat{\pi}$ and $\sum_{i=1}^n \|z_i\|^2 \left\| \frac{\partial \tilde{h}(z_i, v_i(\hat{\pi}))}{\partial v_i} \right\| / n \rightarrow_p E \left[\|z_i\|^2 \left\| \frac{\partial \tilde{h}(z_i, v_i(\pi_0))}{\partial v_i} \right\| \right] < \infty$ under (iii) and (iv). Therefore $\sum_{i=1}^n z_i \tilde{h}(z_i, \hat{v}_i)/n \rightarrow_p E[z_i \tilde{h}(z_i, v_i)]$. 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

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

and therefore the conclusion follows. $\square$

The theorem also makes it clear that the classic linear CF 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).²²

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

$$E[\varepsilon_i|z_i, v_i] = \rho_1 v_i + \rho_2 v_i z_i, \quad (30)$$

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

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

which follows from $\hat{v}_i \rightarrow_p v_i$ (because $\hat{\pi} \rightarrow_p \pi_0$) 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 $\hat{\rho}_1 \hat{v}_i + \hat{\rho}_2 \hat{v}_i z_i$ from our CMR-CF estimator would be consistent for $E[\varepsilon_i|z_i, v_i]$.

Appendix C: Proof of Convergence Rates (Theorem 4)

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

Define $h_L(z, v) = a_L' \tilde{\varphi}^L(z, v)$ and $\hat{h}_L(z, v) = a_L' \hat{\varphi}^L(z, v)$ where a_L satisfies Assumption L1 (iv).²³ Define $\psi_i^L(z_i, v_i) = (\phi_1(x_i, z_{if}), \dots, \phi_K(x_i, z_{if}), \tilde{\varphi}^L(z_i, v_i))'$ with $\tilde{\varphi}^L(z_i, v_i) = (\tilde{\varphi}_1(z_i, v_i), \dots, \tilde{\varphi}_L(z_i, v_i))'$ and $\hat{\psi}_i^L(z_i, v_i) = (\phi_1(x_i, z_{if}), \dots, \phi_K(x_i, z_{if}), \hat{\varphi}^L(z_i, v_i))'$ with $\hat{\varphi}^L(z_i, v_i) = (\hat{\varphi}_1(z_i, v_i), \dots, \hat{\varphi}_L(z_i, v_i))'$. We further let $\hat{\psi}_i^L = \hat{\psi}_i^L(z_i, \hat{v}_i)$, $\psi_i^L = \psi_i^L(z_i, v_i)$, and $\hat{\psi}_i^L = \hat{\psi}_i^L(z_i, v_i)$. We further let $\psi^{L,n} = (\psi_1^L, \dots, \psi_n^L)'$, $\hat{\psi}^{L,n} = (\hat{\psi}_1^L, \dots, \hat{\psi}_n^L)'$, and $\hat{\psi}^{L,n} = (\hat{\psi}_1^L, \dots, \hat{\psi}_n^L)'$.

²² 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.

²³ With abuse of notation we write $a_L = (a_1, \dots, a_L)'$.

Let C (also C_1, C_2 , and others) denote a generic positive constant and let $C(Z, V)$ or $C(X, Z_f)$ (also $C_1(\cdot)$, $C_2(\cdot)$, and others) denote a generic bounded positive function of (Z, V) or (X, Z_f) . We often write $C_i = C(x_i, z_{if})$. 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 $p_{B_1}^k(z) = B_1 p^k(z)$, $\phi_{B_2}^K(x, z_f) = B_2 \phi^K(x, z_f)$, $\tilde{\varphi}_{B_3}^L(z, v) = B_3 \tilde{\varphi}^L(z, v)$, and $\psi_B^{\mathbf{L}}(z, v) = B \psi^{\mathbf{L}}(z, v)$, $E[p_{B_1}^k(Z_i) p_{B_1}^k(Z_i)']$, $E[\phi_{B_2}^K(X_i, Z_{if}) \phi_{B_2}^K(X_i, Z_{if})']$, $E[\tilde{\varphi}_{B_3}^L(Z_i, V_i) \tilde{\varphi}_{B_3}^L(Z_i, V_i)']$, 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 $|\phi^K(x, z_f)|_\delta \leq \zeta_\delta(K)$, $|\psi^{\mathbf{L}}(z, v)|_\delta \leq \zeta_\delta(\mathbf{L})$ (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 $\gamma_1, \gamma_2, \gamma_f, \gamma > 0$, and $\beta^K, \alpha_L, \beta_{\mathbf{L}}, \lambda_k^1$, and $\lambda_{l,k}^2$ such that $|\Pi_0(z) - \lambda_k^1 p^k(z)|_\delta \leq Ck^{-\gamma_1}$, $|\tilde{\varphi}_{0l}(z) - \lambda_{l,k}^2 p^k(z)|_\delta \leq Ck^{-\gamma_2}$ for all l , $|f_0(x, z_f) - \beta^{K'} \phi^K(x, z_f)|_\delta \leq CK^{-\gamma_f}$, $|h_0(z, v) - \alpha_L' \tilde{\varphi}^L(z, v)|_\delta \leq CL^{-\gamma}$, and $|g_0(z, v) - \beta_{\mathbf{L}}' \psi^{\mathbf{L}}(z, v)|_\delta \leq C\mathbf{L}^{-\gamma}$; (v) both $\mathcal{Z} = \text{supp}(Z)$ and $\mathcal{X} = \text{supp}(X)$ are compact.

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

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

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

C.1 Proof of Lemma L1

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

First note $(P'P)/n$ becomes nonsingular wpa1 as $\xi_0(k)^2 k/n \rightarrow 0$ 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 Newey, Powell, and Vella (1999), we obtain

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

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

$$\max_{i \leq n} \|\hat{\Pi}_i - \Pi_i\| = O_p(\xi_0(k)\Delta_{n,1}) \quad (32)$$

$$\max_{i \leq n} \|\dot{\hat{\varphi}}_{li} - \bar{\varphi}_{li}\| = O_p(\xi_0(k)\Delta_{n,2}) \quad \text{for all } l. \quad (33)$$

Define $\hat{\mathcal{T}} = (\hat{\psi}^{\mathbf{L},n})' \hat{\psi}^{\mathbf{L},n} / n$ and $\dot{\mathcal{T}} = (\psi^{\mathbf{L},n})' \psi^{\mathbf{L},n} / n$. Our goal is to show that $\hat{\mathcal{T}}$ is nonsingular wpa1. We first note that $\dot{\mathcal{T}}$ is nonsingular wpa1 by Assumption L1 (ii) as $\zeta_0(\mathbf{L})^2 \mathbf{L}/n \rightarrow 0$ 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 C.2 and derive conditions under which all of them are satisfied. Next note that

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

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

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

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

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

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

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

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

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

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

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

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

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

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

It follows that

$$\begin{aligned}
\|\hat{\tau} - \tau\| &\leq \|\hat{\tau} - \dot{\tau}\| + \|\dot{\tau} - \tau\| \\
&= O_p\left((\Delta_n^\varphi)^2 + \mathbf{L}^{1/2} \Delta_n^\varphi + \zeta_0(\mathbf{L})\sqrt{\mathbf{L}/n}\right) \equiv O_p(\Delta_\tau) = o_p(1)
\end{aligned} \tag{41}$$

where we obtain $\|\dot{\tau} - \tau\| = O_p(\zeta_0(\mathbf{L})\sqrt{\mathbf{L}/n})$ by the same proof in Lemma A1 of Newey, Powell, and Vella (1999). Therefore we conclude $\hat{\tau}$ is also nonsingular wpa1. The same conclusion holds even when we instead take $\hat{\tau} = \sum_{i=1}^n C(z_i, v_i) \hat{\psi}_i^L \hat{\psi}_i^{L'}/n$ and $\dot{\tau} = \sum_{i=1}^n C(z_i, v_i) \psi_i^L \psi_i^{L'}/n$ 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 $\eta = (\eta_1, \dots, \eta_n)'$. Let $(\mathbf{Z}, \mathbf{V}) = ((Z_1, V_1), \dots, (Z_n, V_n))$. 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 obtain

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

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 $\|(\hat{\psi}^{L,n} - \psi^{L,n})'\eta/n\|^2 = o_p(n^{-1})$. Also note that $E[\|(\psi^{L,n})'\eta/n\|^2] = \mathbf{CL}/n$ (see proof of Lemma A1 in Newey, Powell, and Vella (1999)). Therefore, by the triangle inequality

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

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

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

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

$$\begin{aligned}
\|\hat{\tau}^{-1}\hat{\psi}^{L,n'}(\tilde{g}_L - \tilde{g}_0)/n\| &\leq O_p(1) \left\{ (\tilde{g}_L - \tilde{g}_0)' \hat{\psi}^{L,n}(\hat{\psi}^{L,n'}\hat{\psi}^{L,n})^{-1}\hat{\psi}^{L,n'}(\tilde{g}_L - \tilde{g}_0)/n \right\}^{1/2} \\
&\leq O_p(1) \{(\tilde{g}_L - \tilde{g}_0)'(\tilde{g}_L - \tilde{g}_0)/n\}^{1/2} = O_p(\mathbf{L}^{-\gamma}).
\end{aligned} \tag{44}$$

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

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

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

$$\begin{aligned}\|\hat{\mathcal{T}}^{-1}\hat{\psi}^{L,n'}(\tilde{\mathbf{g}}_0 - \mathbf{g}_0)/n\| &= O_p(1) \left(\sum_{i=1}^n \|h_0(z_i, \hat{v}_i) - h_0(z_i, v_i)\|^2/n \right)^{1/2} \\ &\leq O_p(1) \left(\sum_{i=1}^n \|\hat{\Pi}_i - \Pi_i\|^2/n \right)^{1/2} = O_p(\Delta_{n,1}) = o_p(1).\end{aligned}\quad (46)$$

Combining (42)–(46) and by $\hat{\mathcal{T}}$ being nonsingular wpa1, we obtain

$$\begin{aligned}\|\hat{\beta} - \beta_L\| &\leq \|\hat{\mathcal{T}}^{-1}\hat{\psi}^{L,n'}\eta/n\| + \|\hat{\mathcal{T}}^{-1}\hat{\psi}^{L,n'}(\hat{\mathbf{g}}_L - \tilde{\mathbf{g}}_L)/n\| + \|\hat{\mathcal{T}}^{-1}\hat{\psi}^{L,n'}(\tilde{\mathbf{g}}_0 - \mathbf{g}_0)/n\| + o_p(1) \\ &= O_p(1) \left\{ \sqrt{\mathbf{L}/n} + L\xi_0(k)\Delta_{n,1}\sqrt{k/n} + L\Delta_{n,2} + \mathbf{L}^{-\gamma} \right\} \equiv O_p(\Delta_{n,\beta}).\end{aligned}\quad (47)$$

Define $\mathbf{g}_{Li}^* = f_K(x_i, z_{if}) + h_L^*(z_i, v_i)$ where $h_L^*(z_i, v_i) = a_L'(\varphi^L(z_i, v_i) - \hat{\varphi}^L(z_i))$. Then applying the triangle inequality, by (37), the Markov inequality, Assumption L1 (iv), and $\hat{\mathcal{T}}$ is nonsingular wpa1 (by Assumption L1 (ii) and (41)), we conclude

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

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

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

C.2 Proof of Theorem 4

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) $\mathbf{L}^{1/2}\Delta_n^\varphi \rightarrow 0$ from (40) where we define $\Delta_n^\varphi = (\zeta_1(L) + L^{1/2}\xi_0(k)\sqrt{k/n} + L^{1/2})\Delta_n$, (ii) $\zeta_0(\mathbf{L})^2\mathbf{L}/n \rightarrow 0$ (such

that $\mathcal{T}$ is nonsingular wpa1), and (iii) $\xi_0(k)^2 k/n \rightarrow 0$ (such that $P'P/n$ is nonsingular wpa1). The other rate conditions are dominated by these three conditions.

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) $\mathbf{L}^{1/2}(L^3 + L^{1/2}k^{3/2}/\sqrt{n} + L^{1/2})\Delta_n \rightarrow 0$, (ii) $\mathbf{L}^3/n \rightarrow 0$, and (iii) $k^3/n \rightarrow 0$ and for the spline approximations, they become (i) $\mathbf{L}^{1/2}(L^{3/2} + L^{1/2}k/\sqrt{n} + L^{1/2})\Delta_n \rightarrow 0$, (ii) $\mathbf{L}^2/n \rightarrow 0$, and (iii) $k^2/n \rightarrow 0$. Also note that $\Delta_{n,\beta} \equiv \|\hat{\beta} - \beta_L\| = \sqrt{\mathbf{L}/n} + L\xi_0(k)\Delta_{n,1}\sqrt{k/n} + L\Delta_{n,2} + \mathbf{L}^{-\gamma} = \sqrt{\mathbf{L}/n} + L\Delta_n + \mathbf{L}^{-\gamma}$ since $\xi_0(k)\sqrt{k/n} = o(1)$. Here we can take $\gamma_f = s/(d_x + d_{z_f})$ 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 Chen 2007; Newey 1997; Schumaker 1981; Timan 1963). Therefore, the conclusion of Theorem 4 (a) follows from Lemma L1 applying the dominated convergence theorem since $\hat{g}_i$ and g_{0i} are bounded.

For Theorem 4 (b) note that for $\hat{\beta} = (\hat{\beta}^{K'}, \hat{a}_L')'$ and $\beta_L = (\beta^{K'}, a_L')'$,

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

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

Appendix D: Proof of Asymptotic Normality (Theorems 5 and 6)

D.1 Rate Conditions

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

$$\begin{aligned} \Delta_n^\varphi &= (\zeta_1(L) + L^{1/2}\xi_0(k)\sqrt{k/n} + L^{1/2})\Delta_n, \quad \Delta_{n,\beta} = \sqrt{\mathbf{L}/n} + L\Delta_n + \mathbf{L}^{-\gamma} \\ \Delta_\tau &= (\Delta_n^\varphi)^2 + \mathbf{L}^{1/2}\Delta_n^\varphi + \zeta_0(\mathbf{L})\sqrt{\mathbf{L}/n}, \quad \Delta_{\tau_1} = \xi_0(k)\sqrt{k/n} \\ \Delta_H &= \zeta_0(\mathbf{L})k^{1/2}/\sqrt{n} + k^{1/2}\Delta_n^\varphi + L^{-\gamma}\zeta_0(\mathbf{L})\sqrt{k} \\ \Delta_{d\varphi} &= \zeta_0(\mathbf{L})L\Delta_{n,2}, \quad \Delta_g = \zeta_0(\mathbf{L})\Delta_{n,\beta} \\ \Delta_\Sigma &= \Delta_\tau + \zeta_0(\mathbf{L})^2\mathbf{L}/n, \quad \Delta_{\hat{H}} = (\zeta_1(L)\Delta_{n,\beta} + \xi_0(k)\Delta_{n,1})\mathbf{L}^{1/2}\xi_0(k) \end{aligned}$$

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

$$\begin{aligned} \sqrt{n}\mathbf{L}^{-\gamma} &\rightarrow 0, \quad \sqrt{nk}^{1/2}L^{-\gamma} \rightarrow 0, \quad \sqrt{nk}^{-\gamma_1} \rightarrow 0, \quad \sqrt{nk}^{-\gamma_2} \rightarrow 0 \\ k^{1/2}(\Delta_{\tau_1} + \Delta_H) + \mathbf{L}^{1/2}\Delta_\tau &\rightarrow 0, \quad n^{-1}(\zeta_0(\mathbf{L})^2\mathbf{L} + \xi_0(k)^2k + \xi_0(k)^2kL^4) \rightarrow 0, \\ k^{1/2}(\Delta_{\tau_1} + \Delta_H) + \mathbf{L}^{1/2}\Delta_\tau + \Delta_{d\varphi} &\rightarrow 0, \quad \Delta_g \rightarrow 0, \quad \Delta_\Sigma \rightarrow 0, \quad \Delta_{\hat{H}} \rightarrow 0. \end{aligned}$$

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

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

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

D.2 Asymptotic Variance Terms

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

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

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

$$E[v_{Li}^* \text{var}(Y_i | Z_i, V_i) v_{Li}^{*'}] = \mathcal{A} \mathcal{T}^{-1} E[\psi_i^L \text{var}(Y_i | Z_i, V_i) \psi_i^{L'}] \mathcal{T}^{-1} \mathcal{A}' \rightarrow E[v_i^* \text{var}(Y_i | Z_i, V_i) v_i^{*'}].$$

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

$$b_{Li} = E\left[v_{Li}^* \left(\frac{\partial h_{0i}}{\partial V_i} - E\left[\frac{\partial h_{0i}}{\partial V_i} | Z_i\right]\right) p_i^{k'}\right] p_i^k$$

and $b_i = E\left[v_i^* \left(\frac{\partial h_{0i}}{\partial V_i} - E\left[\frac{\partial h_{0i}}{\partial V_i} | Z_i\right]\right) p_i^{k'}\right] p_i^k$. Note that because $(\mathcal{T}_1)^{-1} = I$, b_{Li} and b_i are least squares mean projections of $v_{Li}^* \left(\frac{\partial h_{0i}}{\partial V_i} - E\left[\frac{\partial h_{0i}}{\partial V_i} | Z_i\right]\right)$ on p_i^k and $v_i^* \left(\frac{\partial h_{0i}}{\partial V_i} - E\left[\frac{\partial h_{0i}}{\partial V_i} | Z_i\right]\right)$ on p_i^k , respectively. Then $E[\|b_{Li} - b_i\|^2] \leq CE[\|v_{Li}^* - v_i^*\|^2] \rightarrow 0$ 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 $E[\|\rho_v(Z_i) - b_i\|^2] \rightarrow 0$ as $k \rightarrow \infty$ because b_i is a least squares projection of $v_i^* \left(\frac{\partial h_{0i}}{\partial V_i} - E\left[\frac{\partial h_{0i}}{\partial V_i} | Z_i\right]\right)$ on p_i^k and it converges to the conditional mean as $k \rightarrow \infty$. Finally note that

$$E[b_{Li} \text{var}(V_i | Z_i) b_{Li}'] = \mathcal{A} \mathcal{T}^{-1} E\left[\psi_i^L \left(\frac{\partial h_{0i}}{\partial V_i} - E\left[\frac{\partial h_{0i}}{\partial V_i} | Z_i\right]\right) p_i^{k'}\right] E[\text{var}(V_i | Z_i) p_i^k p_i^{k'}]$$

$$\begin{aligned} & \times E \left[p_i^k \left(\frac{\partial h_{0i}}{\partial V_i} - E \left[\frac{\partial h_{0i}}{\partial V_i} | Z_i \right] \right) \psi_i^L \right] \tau^{-1} \mathcal{A}' \\ & = \mathcal{A} \tau^{-1} H_1 \Sigma_1 H_1' \tau^{-1} \mathcal{A}' \end{aligned}$$

and therefore we conclude $\mathcal{A} \tau^{-1} H_1 \Sigma_1 H_1' \tau^{-1} \mathcal{A}'$ 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

$$\mathcal{A} \tau^{-1} H_{2,l} \Sigma_{2,l} H_{2,l}' \tau^{-1} \mathcal{A}' \rightarrow E [\rho_{\bar{\varphi}_l}(Z) \text{var}(\varphi_l(Z, V) | Z) \rho_{\bar{\varphi}_l}(Z)'] .$$

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

D.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 (16) to the corresponding terms in (50). First we establish some results here to be used for deriving the asymptotic normality below. Note $\|\hat{\tau} - \tau\| = O_p(\Delta_\tau) = o_p(1)$ and $\|\hat{\tau}_1 - \tau_1\| = O_p(\Delta_{\tau_1}) = o_p(1)$. We also have $\|\Gamma \mathcal{A}(\hat{\tau}^{-1} - \tau^{-1})\| = o_p(1)$ and $\|\Gamma \mathcal{A} \hat{\tau}^{-1/2}\|^2 = O_p(1)$ (see proof in Lemma A1 of Newey, Powell, and Vella (1999)). We next show $\|\bar{H}_{11} - H_{11}\| = o_p(1)$. Let $H_{11L} = E \left[\sum_{l=1}^L a_l \frac{\partial \varphi_l(Z_i, V_i)}{\partial V_i} \psi_i^L p_i^{k'} \right]$ and $\bar{H}_{11L} = \sum_{i=1}^n \sum_{l=1}^L a_l \frac{\partial \varphi_l(Z_i, V_i)}{\partial V_i} \psi_i^L p_i^{k'} / n$. Similarly define $H_{12L} = E \left[E \left[\sum_{l=1}^L a_l \frac{\partial \varphi_l(Z_i, V_i)}{\partial V_i} | Z_i \right] \psi_i^L p_i^{k'} \right]$ and $\bar{H}_{12L} = \sum_{i=1}^n E \left[\sum_{l=1}^L a_l \frac{\partial \varphi_l(Z_i, V_i)}{\partial V_i} | Z_i \right] \psi_i^L p_i^{k'} / n$ and let $H_{1L} = H_{11L} - H_{12L}$. By Assumption N1 (i), L1 (iii) and the Cauchy–Schwarz inequality,

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

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

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

where the first equality holds because $\frac{\partial h_{Li}}{\partial V_i} = \sum_{l=1}^L a_l \frac{\partial \bar{\varphi}_l(Z_i, V_i)}{\partial V_i} = \sum_{l=1}^L a_l \frac{\partial \varphi_l(Z_i, V_i)}{\partial V_i}$ 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

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

Therefore, we have $\|\bar{H}_{11} - H_{11}\| = O_p(\zeta_0(\mathbf{L}) k^{1/2} / \sqrt{n} + k^{1/2} \Delta_n^\varphi + L^{-\gamma} \zeta_0(\mathbf{L}) \sqrt{k}) \equiv O_p(\Delta_H) = o_p(1)$. 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 $\hat{g}_{Li} = f_K(x_i, z_{if}) + \tilde{h}_L(z_i, \hat{v}_i)$ where $f_K(x_i, z_{if}) = \sum_{j=1}^K \beta_j \phi_j(x_i, z_{if})$ 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

$g_{Li} = f_K(x_i, z_{if}) + h_L(z_i, v_i)$. Recall $\beta_L = (\beta_1, \dots, \beta_K, a'_L)'$ and let this β_L satisfy Assumption N1 (i). Then from the first order condition, we obtain the expansion similar to (43) as ²⁴

$$\begin{aligned} 0 &= \hat{\psi}^{L,n'}(y - \hat{g})/\sqrt{n} \\ &= \hat{\psi}^{L,n'}(\eta - (\hat{g} - \hat{g}_L) - (\hat{g}_L - \hat{g}_L) - (\hat{g}_L - g_L) - (g_L - g_0))/\sqrt{n} \\ &= \hat{\psi}^{L,n'}(\eta - \hat{\psi}^{L,n}(\hat{\beta} - \beta_L) - (\hat{g}_L - \hat{g}_L) - (\hat{g}_L - g_L) - (g_L - g_0))/\sqrt{n}. \end{aligned} \quad (51)$$

Similar to (44), we obtain

$$\|\hat{\mathcal{T}}^{-1}\hat{\psi}^{L,n'}(g_L - g_0)/\sqrt{n}\| = O_p(\sqrt{n}L^{-\gamma}). \quad (52)$$

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

$$\begin{aligned} \sqrt{n}\|\Gamma(\alpha(g_L) - \alpha(g_0))\| &= \sqrt{n}\|\Gamma\| \cdot \|\alpha(g_L - g_0)\| \leq C\sqrt{n}\|\Gamma\| \cdot |\psi^{L'}(\cdot)\beta_L - g_0(\cdot)|_\delta \\ &= O_p(\sqrt{n}L^{-\gamma}) = o_p(1). \end{aligned} \quad (53)$$

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

$$\begin{aligned} \sqrt{n}\Gamma(\hat{\theta} - \theta_0) &= \sqrt{n}\Gamma(\alpha(\hat{g}) - \alpha(g_0)) = \sqrt{n}\Gamma(\alpha(\hat{g}) - \alpha(g_L)) + \sqrt{n}\Gamma(\alpha(g_L) - \alpha(g_0)) \\ &= \sqrt{n}\Gamma\mathcal{A}(\hat{\beta} - \beta_L) + \sqrt{n}\Gamma\{\alpha(g_L) - \alpha(g_0)\} \\ &= \Gamma\mathcal{A}\hat{\mathcal{T}}^{-1}\hat{\psi}^{L,n'}(\eta - (\hat{g}_L - \hat{g}_L) - (\hat{g}_L - g_L))/\sqrt{n} + o_p(1). \end{aligned} \quad (54)$$

D.3.1 Influence Function for the First Stage

Now we derive the stochastic expansion of $\Gamma\mathcal{A}\hat{\mathcal{T}}^{-1}\hat{\psi}^{L,n'}(\hat{g}_L - g_L)/\sqrt{n}$. 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 $h_{Li} = h_L(z_i, v_i)$),

$$\begin{aligned} \Gamma\mathcal{A}\hat{\mathcal{T}}^{-1}\sum_{i=1}^n \hat{\psi}_i^L(\hat{g}_{Li} - g_{Li})/\sqrt{n} &= \Gamma\mathcal{A}\hat{\mathcal{T}}^{-1}\sum_{i=1}^n \hat{\psi}_i^L(\tilde{h}_{Li} - h_{Li})/\sqrt{n} \\ &= \Gamma\mathcal{A}\hat{\mathcal{T}}^{-1}\sum_{i=1}^n \hat{\psi}_i^L\left(\frac{dh_{Li}}{dv_i} - E\left[\frac{dh_{Li}}{dv_i}|Z_i\right]\right)(\hat{\Pi}_i - \Pi_i)/\sqrt{n} + \hat{\zeta} \\ &= \Gamma\mathcal{A}\hat{\mathcal{T}}^{-1}\bar{H}_1\hat{\mathcal{T}}_1^{-1}\sum_{i=1}^n p_i^k v_i/\sqrt{n} + \Gamma\mathcal{A}\hat{\mathcal{T}}^{-1}\bar{H}_1\hat{\mathcal{T}}_1^{-1}\sum_{i=1}^n p_i^k (\Pi_i - p_i^{k'}\lambda_k^1)/\sqrt{n} \\ &\quad + \Gamma\mathcal{A}\hat{\mathcal{T}}^{-1}\sum_{i=1}^n \hat{\psi}_i^L\left(\frac{dh_{Li}}{dv_i} - E\left[\frac{dh_{Li}}{dv_i}|Z_i\right]\right)(p_i^{k'}\lambda_k^1 - \Pi_i)/\sqrt{n} + \hat{\zeta}. \end{aligned} \quad (55)$$

The remainder term $\|\hat{\zeta}\| \leq C\sqrt{n}\|\Gamma\mathcal{A}\hat{\mathcal{T}}^{-1/2}\|\zeta_0(L)\sum_{i=1}^n C_i\|\hat{\Pi}_i - \Pi_i\|^2/n = O_p(\sqrt{n}\zeta_0(L)\Delta_{n,1}^2) = o_p(1)$. 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 $\sqrt{nk^{-s_1/d_z}} \rightarrow 0$ (so that $\sqrt{n}|\Pi_0(z) - \lambda_k^{1'}p^k(z)|_0 \rightarrow 0$ by Assumption L1 (iv)) and under $k^{1/2}(\Delta_{\mathcal{T}_1} + \Delta_H) + L^{1/2}\Delta_{\mathcal{T}} \rightarrow 0$ (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

²⁴ If there exists an estimation error due to tolerance in minimization, take the error arbitrary small to justify this asymptotic expansion.

$$\Gamma \mathcal{A} \hat{\mathcal{T}}^{-1} \hat{\psi}^{\mathbf{L}, n'} (\hat{\mathbf{g}}_{\mathbf{L}} - \mathbf{g}_{\mathbf{L}}) / \sqrt{n} = \Gamma \mathcal{A} \mathcal{T}^{-1} H_1 \sum_{i=1}^n p_i^k v_i / \sqrt{n} + o_p(1). \quad (56)$$

This derives the influence function due to estimating v_i in the first step.

D.3.2 Influence Function for the Second Stage

Next we derive the stochastic expansion of $\Gamma \mathcal{A} \hat{\mathcal{T}}^{-1} \hat{\psi}^{\mathbf{L}, n'} (\hat{\mathbf{g}}_{\mathbf{L}} - \mathbf{g}_{\mathbf{L}}) / \sqrt{n}$ as

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

where $\rho_i = \sum_l a_l \left\{ p_i^{k'} \hat{\mathcal{T}}_1^{-1} \sum_{j=1}^n p_j^k (\varphi_l(z_j, \hat{v}_j) - \varphi_l(z_j, v_j)) / n - (E[\varphi_l(Z_i, \hat{V}_i) | z_i] - \tilde{\varphi}_l(z_i)) \right\}$. We first focus on the last term in (57). Note that $p_i^{k'} \hat{\mathcal{T}}_1^{-1} \sum_{j=1}^n p_j^k (\varphi_l(z_i, \hat{v}_i) - \varphi_l(z_i, v_i)) / n$ 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 $E[\varphi_l(Z_i, \hat{V}_i) | z_i] - \tilde{\varphi}_l(z_i)$. 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 $E[\rho_i | Z_1, \dots, Z_n] = 0$ and therefore

$$E[\|\rho_i\|^2 | Z_1, \dots, Z_n] \leq E\left[L \sum_{l=1}^L \|\rho_{il}\|^2 | Z_1, \dots, Z_n\right] \leq L^2 O_p(\Delta_{n,2}^2)$$

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,

$$E\left[\left\|\sum_{i=1}^n \hat{\psi}_i^{\mathbf{L}} \rho_i / \sqrt{n}\right\| | Z_1, \dots, Z_n\right] \leq \left(E[\|\hat{\psi}_i^{\mathbf{L}}\|^2 \|\rho_i\|^2 | Z_1, \dots, Z_n]\right)^{1/2} \leq C \zeta_0(\mathbf{L}) L \Delta_{n,2}.$$

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

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 $\sqrt{nk}^{-s_2/d_z} \rightarrow 0$ (so that $\sqrt{n}|\tilde{\varphi}_l(z) - \lambda_{l,k}' p^k(z)|_0 \rightarrow 0$ for all l by Assumption L1 (iv)), under $\sqrt{nk}^{1/2} L^{-s/d} \rightarrow 0$ (so that $\sqrt{nk}^{1/2} |h_0(z, v) - a'_L \tilde{\varphi}^L(z, v)|_0 \rightarrow 0$ by Assumption L1 (iv)), and under $k^{1/2}(\Delta_{\mathcal{T}_1} + \Delta_H) + \mathbf{L}^{1/2} \Delta_{\mathcal{T}} + \Delta_{d\varphi} \rightarrow 0$ (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

$$\Gamma \mathcal{A} \hat{\mathcal{T}}^{-1} \hat{\psi}^{\mathbf{L}, n'} (\hat{\mathbf{g}}_{\mathbf{L}} - \mathbf{g}_{\mathbf{L}}) / \sqrt{n} = \Gamma \mathcal{A} \mathcal{T}^{-1} \sum_l H_{2,l} \sum_{i=1}^n p_i^k \tilde{\varphi}_{li} / \sqrt{n} + o_p(1). \quad (58)$$

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 $\hat{\psi}_i^{\mathbf{L}}$ with $\psi_i^{\mathbf{L}}$ does not influence the stochastic expansion by (39). Therefore by (54), (56), and (58), we obtain the stochastic expansion,

$$\sqrt{n} \Gamma(\hat{\theta} - \theta_0) = \Gamma \mathcal{A} \mathcal{T}^{-1} \left(\psi^{\mathbf{L}, n'} \eta - H_1 \sum_{i=1}^n p_i^k v_i / \sqrt{n} - \sum_l H_{2,l} \sum_{i=1}^n p_i^k \tilde{\varphi}_{li} / \sqrt{n} \right) + o_p(1).$$

To apply the Lindeberg–Feller central limit theorem, we check the Lindeberg condition. For any vector q with $\|q\| = 1$, let $W_{in} = q' \Gamma \mathcal{A} \mathcal{T}^{-1} (\psi_i^L \eta_i - H_1 p_i^k v_i - \sum_l H_{2,l} p_i^k \tilde{\varphi}_{li}) / \sqrt{n}$. 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 $CI - H_j H_j'$ being positive semi-definite for $j = 1, (2, 1), \dots, (2, L)$. Also note that $\left(\sum_{l=1}^L \tilde{\varphi}_{li}\right)^4 \leq L^2 \left(\sum_{l=1}^L \tilde{\varphi}_{li}^2\right)^2 \leq L^3 \sum_{l=1}^L \tilde{\varphi}_{li}^4$. It follows that for any $\varepsilon > 0$,

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

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

D.4 Consistency of the Estimate of the Asymptotic Variance

Now we show the convergence of the each term in (16) 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 $\max_{i \leq n} |\hat{g}_i - g_{0i}| = O_p(\zeta_0(\mathbf{L})\Delta_{n,\beta}) = o_p(1)$ by (49). Let $\hat{D} = \Gamma \mathcal{A} \hat{\mathcal{T}}^{-1} \hat{\psi}^{L,n'} \text{diag}\{1 + |\eta_i|, \dots, 1 + |\eta_n|\} \hat{\psi}^{L,n} \hat{\mathcal{T}}^{-1} \mathcal{A}' \Gamma'$ and note that $\hat{\psi}^{L,n}$ and $\hat{\mathcal{T}}$ only depend on $(Z_1, V_1), \dots, (Z_n, V_n)$ and thus $E[\hat{D} | (Z_1, V_1), \dots, (Z_n, V_n)] \leq C \Gamma \mathcal{A} \hat{\mathcal{T}}^{-1} \mathcal{A}' \Gamma' = O_p(1)$. Therefore, $\|\hat{D}\| = O_p(1)$ as well. Next let $\tilde{\Sigma} = \sum_{i=1}^n \hat{\psi}_i^L \hat{\psi}_i^{L'} \eta_i^2 / n$. Then,

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

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

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

It also follows that $\|\hat{\Sigma} - \Sigma\| = o_p(1)$ because $\|\Gamma \mathcal{A} \hat{\mathcal{T}}^{-1}\| = O_p(1)$. Then, by (59) and (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 $j = 1, (2, 1), \dots, (2, L)$,

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

As we have shown $\|\hat{\Sigma} - \Sigma\| = o_p(1)$, similarly we can show $\|\hat{\Sigma}_j - \Sigma_j\| = o_p(1)$, $j = 1, (2, 1), \dots, (2, L)$. We focus on showing $\|\hat{H}_j - \bar{H}_j\| = o_p(1)$ for $j = 1, (2, 1), \dots, (2, L)$. First note that $\|\hat{H}_{11} - \bar{H}_{11}\| = \left\| \sum_{i=1}^n \left(\sum_{l=1}^L \hat{a}_l \frac{\partial \varphi_l(z_i, \hat{v}_i)}{\partial v_i} - a_l \frac{\partial \varphi_l(z_i, v_i)}{\partial v_i} \right) \hat{\psi}_i^L p^k(z_i)' / n \right\|$. By the Cauchy–Schwarz inequality, (38), and Assumption L1 (iii), we have $\sum_{i=1}^n \|\hat{\psi}_i^L p_i^{k'}\|^2 / n \leq \sum_{i=1}^n \|\hat{\psi}_i^L\|^2 \|p_i^k\|^2 / n = O_p(\mathbf{L} \xi_0(k)^2)$. 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 ,

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

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, $\|\hat{H}_{11} - \bar{H}_{11}\| \leq O_p((\zeta_1(L) \Delta_{n,\beta} + \xi_0(k) \Delta_{n,1}) \mathbf{L}^{1/2} \xi_0(k)) = O_p(\Delta_{\hat{H}}) = o_p(1)$. Similarly we can show that $\|\hat{H}_{12} - \bar{H}_{12}\| = o_p(1)$ and $\|\hat{H}_{2,l} - \bar{H}_{2,l}\| = o_p(1)$ for $l = 1, \dots, L$. We have shown that $\|\bar{H}_j - H_j\| = o_p(1)$ for $j = 1, (2, 1), \dots, (2, L)$ previously. Therefore, $\|\hat{H}_j - H_j\| = o_p(1)$ for $j = 1, (2, 1), \dots, (2, L)$. Then by a similar proof to (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

- Ackerberg, D., X. Chen, and J. Hahn. 2012. “A Practical Asymptotic Variance Estimator for Two-Step Semiparametric Estimators.” *The Review of Economics and Statistics* 94: 481–98.
- Ai, C., and X. Chen. 2003. “Efficient Estimation of Models with Conditional Moment Restrictions Containing Unknown Functions.” *Econometrica* 71 (6): 1795–843.
- Blundell, R., X. Chen, and D. Kristensen. 2007. “Semi-Nonparametric IV Estimation of Shape-Invariant Engel Curves.” *Econometrica* 75: 1613–69.
- Chen, X. 2007. “Large Sample Sieve Estimation of Semi-Nonparametric Models.” In *Handbook of Econometrics*, Vol. 6. Amsterdam: North Holland Publishing Co.
- Chen, X., and T. Christensen. 2015. “Optimal Sup-Norm Rates, Adaptivity and Inference in Nonparametric Instrumental Variables Estimation.” In *Cowles Foundation Discussion Paper No. 1923R*.
- Darolles, S., J. Florens, and E. Renault. 2006. “Nonparametric Instrumental Regression.” In *Working Paper*. GREMAQ, University of Toulouse.
- Florens, J. 2003. “Inverse Problems and Structural Econometrics: The Example of Instrumental Variables.” In *Advances in Economics and Econometrics: Theory and Applications, Eight World Congress, Econometric Society Monographs*, 1–25. Cambridge: Cambridge University Press.
- Florens, J., J. Heckman, C. Meghir, and E. Vytlacil. 2008. “Identification of Treatment Effects Using Control Functions in Models with Continuous Endogenous Treatment and Heterogeneous Effects.” *Econometrica* 76: 1191–206.
- Hall, P., and J. Horowitz. 2005. “Nonparametric Methods for Inference in the Presence of Instrumental Variables.” *Annals of Statistics* 33: 2904–29.
- Horowitz, J. 2011. “Applied Nonparametric Instrumental Variables Estimation.” *Econometrica* 79: 347–94.
- Imbens, G., and W. Newey. 2009. “Identification and Estimation of Triangular Simultaneous Equations Models without Additivity.” *Econometrica* 77: 1481–512.
- Kasy, M. 2011. “Identification in Triangular Systems Using Control Functions.” *Econometric Theory* 27: 663–71.
- Matzkin, R. 2003. “Nonparametric Estimation of Nonadditive Random Functions.” *Econometrica* 71: 1339–75.
- Murphy, K., and R. Topel. 1985. “Estimation and Inference in Two-Step Econometric Models.” *Journal of Business & Economic Statistics* 3: 370–9.
- Newey, W. 1984. “A Method of Moments Interpretation of Sequential Estimators.” *Economics Letters* 14: 201–6.
- Newey, W. 1997. “Convergence Rates and Asymptotic Normality for Series Estimators.” *Journal of Econometrics* 79: 147–68.
- Newey, W., and J. Powell. 2003. “Instrumental Variable Estimation of Nonparametric Models.” *Econometrica* 71: 1565–78.
- Newey, W., J. Powell, and F. Vella. 1999. “Nonparametric Estimation of Triangular Simultaneous Equations Models.” *Econometrica* 67: 565–603.

- Newey, W., and S. Stouli. 2021. "Control Variables, Discrete Instruments, and Identification of Structural Functions." *Journal of Econometrics* 222: 73–88.
- Pinkse, J. 2000. "Nonparametric Two-Step Regression Functions When Regressors and Error are Dependent." *Canadian Journal of Statistics* 28: 289–300.
- Rivers, D., and Q. Vuong. 1988. "Limited Information Estimators and Exogeneity Tests for Simultaneous Probit Models." *Journal of Econometrics* 39: 347–66.
- Schumaker, L. 1981. *Spline Functions: Basic Theory*. Cambridge: Cambridge University Press.
- Smith, R., and R. Blundell. 1986. "An Exogeneity Test for a Simultaneous Equation Tobit Model with an Application to Labor Supply." *Econometrica* 54: 679–85.
- Su, L., and A. Ullah. 2008. "Local Polynomial Estimation of Nonparametric Simultaneous Equations Models." *Journal of Econometrics* 144: 193–218.
- Timan, A. 1963. *Theory of Approximation of Functions of a Real Variable*. Oxford, New York: Pergamon Press.
- Wooldridge, J. 2005. "Unobserved Heterogeneity and Estimation of Average Partial Effects." In *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*, edited by J. H. Stock, and D. W. K. Andrews, 27–55. Cambridge: Cambridge University Press.
- Wooldridge, J., and L. Papke. 2008. "Panel Data Methods for Fractional Response Variables with an Application to Test Pass Rates." *Journal of Econometrics* 145: 121–33.