NBER WORKING PAPER SERIES

REVISITING INSTRUMENTAL VARIABLES AND THE CLASSIC CONTROL FUNCTION APPROACH, WITH IMPLICATIONS FOR PARAMETRIC AND NON-PARAMETRIC REGRESSIONS

Kyoo il Kim
Amil Petrin

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

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

Revisiting Instrumental Variables and the Classic Control Function Approach, with Implications for Parametric and Non-Parametric Regressions
Kyoo il Kim and Amil Petrin
NBER Working Paper No. 16679
January 2011
JEL No. C14

ABSTRACT

It is well-known from projection theory that two-stage least squares (2SLS) and the classic control function (CF) estimator in the linear simultaneous equations models are numerically equivalent. Yet the classic CF approach assumes that the regression error in the outcome equation is mean independent of the instruments conditional on the CF control while 2SLS does not. We resolve this puzzle by showing that the classic CF approach omits a generalized control function that may depend on the instruments and control. This term is (asymptotically) uncorrelated with the endogenous regressors given the control under the unconditional moment restrictions of 2SLS. We also show that imposing the 2SLS unconditional moment restrictions in the classic CF setup allows the mean of the error to depend on the instruments and control. In contrast to the linear setting, the non-linear and non-parametric control function setting of Newey, Powell, and Vella (1999) (NPVCF) is no longer consistent if the classic CF condition is violated. This dependence can occur in many economic settings including returns to education, production functions, and demand or supply with non-separable reduced forms for equilibrium prices. We use our results to develop an estimator for this setting that is consistent when the structural error may depend on the instruments given the CF control. Our approach achieves identification by augmenting the NPVCF setting with conditional moment restrictions. Our estimator is a multi-step least squares estimator and thus maintains the simplicity of the NPVCF estimator. Our monte carlos are motivated by our economic examples and they show that our new estimator performs well while the classical CF estimator and the non-parametric analog of NPVCF can be biased in non-linear or nonparametric settings when the conditional mean of the error depends on the instruments given the control.

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

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

1 Introduction

The problem of endogenous regressors in simultaneous equations models has a long history in econometrics and empirical studies. In linear models with additively separable errors researchers have used both two-stage least squares (2SLS) and the classic control function (CF) approach to correct for the bias induced by the correlation between the error and the regressor(s).[1] While it is well known from projection theory that these two estimators are numerically equivalent, they require different exclusion restrictions (or order conditions) to hold for identification. In the case of the classic CF estimator, the first moment of the error in the structural equation cannot depend on the exogenous regressors or instruments conditional on the classic CF control (i.e. the mean projection residual obtained from regressing the endogenous variable on the instruments). If it did the control function would have to include both the regressors and the classic CF control and one would not be able to separately identify the impact of the regressors on the dependent variable from their impact on the control function.

A weakness of the CF restriction is that it can be violated in economic settings where endogeneity is a first-order concern. These include estimation of returns to education, production functions, and demand or supply with non-separable reduced forms for equilibrium prices. Yet the classic CF estimator must be consistent in these settings because it is equal to the 2SLS estimator.

Our first result resolves this puzzle. We show that the classic CF approach omits a *generalized* control function term that may depend on the instruments and control. This term is (asymptotically) uncorrelated with the endogenous regressor(s) given the classic CF control under the unconditional moment restrictions of the 2SLS.[2] We then show that the classic CF estimator can be generalized to allow the conditional expectation of the error to depend on both the classic CF control and instruments by adding the moment restrictions used by 2SLS for identification.

We then turn to the non-linear and the non-parametric setting with additive errors. We build on the non-parametric CF estimator of Newey, Powell, and Vella (1999)(NPVCF) which achieves identification using the classic CF restriction.[3] In contrast to the linear setting, the NPVCF estimator is no longer consistent if the classic CF condition is violated. We show

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

[2]The estimated control function in the classic CF approach is no longer consistent for the expected value of the error conditional on the control and instruments although this is typically viewed as a nuisance parameter.

[3]We use “nonlinear model” to refer to a regression model that is nonlinear in regressors but linear in parameters.

how to use our insights from the linear case to develop an estimator for non-linear and non-parametric models that is consistent even when the structural error may depend on the instruments given the control.

Our approach is to add the conditional moment restrictions to the NPVCF setting to loosen the classic CF restriction. We cast our estimator as a multi-step sieve estimator and develop convergence rates and consistent estimators for the standard errors. An advantage of our estimator is that it maintains the simplicity of implementation of the NPVCF estimator.

Our monte carlos are motivated by our economic examples. They illustrate the ease of implementing our estimator. They also show that our new estimator performs well while the classic CF estimator and the non-parametric analog of NPVCF can be biased in non-linear settings.

The paper proceeds as follows. In the next section we consider the linear additive model. Section 3 provides economic examples where the classic CF restriction may not hold and then uses the results from Section 2 to formulate our new estimator for the non-linear or non-parametric setting. Section 4 discusses identification and Section 5 develops the details of our estimator. Section 6 addresses convergence rates and Section 7 provides conditions under which asymptotic normality holds for several structural objects often of interest. Section 8 provides monte carlos and Section 9 concludes.

2 The Linear Setting with Additive Errors

In this section we revisit the implication of the well-known numerical equivalence of the classic CF estimator and the 2SLS estimator in the linear simultaneous equations models and find that the classic CF approach omits a *generalized* control function term that is asymptotically irrelevant in this setting. We then show that the classic CF estimator can be generalized to allow the conditional expectation of the error in the outcome equation to depend on both the classic CF control and instruments.

We work in the linear simultaneous equations model in mean-deviated form,

$$y_i \;=\; x_i\beta_0 + \varepsilon_i,$$

with y_i the dependent variable and x_i a scalar explanatory variable that is potentially correlated with ε_i. We let z_i denote an instrument vector satisfying

$$E[z_i\,\varepsilon_i] = 0, \quad E[z_i\,x_i] \neq 0. \tag{1}$$

Defining $v_i = x_i - E[x_i|z_i]$ (we further let $E[x_i|z_i] = z_i'\pi_0$, the linear projection in this linear setting), the classic CF estimator posits:

$$y_i = x_i\beta_0 + \rho v_i + \eta_i, \tag{2}$$

and regresses y_i on $E[x_i|z_i, v_i] = x_i$ and $E[\varepsilon_i|z_i, v_i]$, where conditioning the error on (z_i, v_i) controls for its correlation with x_i. The classic CF estimator thus imposes

$$(\textit{Classic}\ \textbf{CF Restriction})\ E[\varepsilon_i|z_i, x_i] = E[\varepsilon_i|z_i, v_i] = E[\varepsilon_i|v_i] = \rho v_i. \tag{3}$$

The first equality in (3) requires that v_i be chosen such that, conditional on it and z_i, x_i is known. This assumption is not restrictive given the way that v_i is constructed. The second equality requires the conditional mean of ε_i to not depend on the instruments z_i conditional on the control v_i. This latter restriction is not innocuous and can be violated in common economic settings (see Section 3.1 for examples). This raises a puzzle as it is well known that 2SLS and the classic CF estimator are numerically equivalent but 2SLS does not require the classic CF restriction to hold.

We resolve the puzzle by using the unconditional moment restrictions from 2SLS given in (1). Consider an unrestricted general specification for the conditional expectation of the error

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

with the function characterizing $E[\varepsilon_i|z_i, v_i]$ having a leading term in v_i and a remaining term denoted by the function $\tilde{h}(z_i, v_i)$. Under the moment restriction of (1) using the law of iterated expectations we have

$$\begin{aligned} 0 &= E[z_i\,\varepsilon_i] = E[z_i\,E[E[\varepsilon_i|z_i, v_i]|z_i]\,] = E[z_i(\tilde{\rho} E[v_i|z_i] + E[\tilde{h}(z_i, v_i)|z_i])] \\ &= E[z_i E[\tilde{h}(z_i, v_i)|z_i]] = E[z_i\,\tilde{h}(z_i, v_i)]. \end{aligned} \tag{4}$$

The result shows that x_i is also uncorrelated with $\tilde{h}(z_i, v_i)$ given v_i when $E[z_i\,\varepsilon_i] = 0$ and v_i is constructed as the classic control function variable, $v_i = x_i - z_i'\pi_0$. It suggests that omitting the term $\tilde{h}(z_i, v_i)$ that satisfies (4) in the classic CF estimation does not create omitted variable bias as long as the classic control v_i is included in the estimation. We elaborate on this point below.

Letting $Y = (y_1, \ldots, y_n)'$, $X = (x_1, \ldots, x_n)'$, $Z = (z_1, \ldots, z_n)'$, and $\hat{V} = (\hat{v}_1, \ldots, \hat{v}_n)'$ we rewrite equation (2) as

$$Y = X\beta_0 + \tilde{\rho}\hat{V} + \tilde{H}(Z, \hat{V}) + \hat{\eta} \tag{5}$$

where $\tilde{H}(Z, \hat{V}) = (\tilde{h}(z_1, \hat{v}_1), \ldots, \tilde{h}(z_n, \hat{v}_n))'$, $\hat{v}_i = x_i - z_i'\hat{\pi}$ is the estimated classic CF control (the fitted residual from the linear projection of x_i on z_i), and $\hat{\eta}$ is the remaining error term.

Then due to the partitioned regression theory, defining $M_{\hat{V}} = I - \hat{V}(\hat{V}'\hat{V})^{-1}\hat{V}'$ and rewriting (5), estimation of β_0 is numerically equivalent to the estimation of β_0 from

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

By coupling (1) with weak regularity conditions we can show $Z'M_{\hat{V}}\tilde{H}(Z, \hat{V})/n$ converges to zero as the sample size increases which proves that the classic CF estimator that omits the function $\tilde{H}(Z, \hat{V})$ in the regression is consistent as long as we include the control $\hat{V}$ in (5) (see Appendix A).

If the classic CF estimator is modified to include the new regressors associated with $\tilde{H}(Z, \hat{V})$ then 2SLS and this generalized CF estimator for β_0 are no longer numerically equivalent although asymptotically they both converge to β_0.[4] In this generalized CF case one would also recover a consistent estimate $E[\varepsilon_i|z_i, v_i]$.[5]

3 The Non-Linear or Non-Parametric Setting with Additive Errors

We consider a nonparametric simultaneous equations model with additivity:

$$\begin{aligned} x_i &= \Pi_0(z_i) + v_i, \qquad E[v_i|z_i] = 0 &\quad (6) \\ y_i &= f_0(x_i, z_{1i}) + \varepsilon_i &\quad (7) \end{aligned}$$

where the instruments z_i includes z_{1i} and $f(x_i, z_{1i})$ can be parametric as $f(x_i, z_{1i}) \equiv f(x_i, z_{1i}; \theta)$ or nonparametric. (6) is a conditional mean decomposition of x_i with $\Pi_0(z_i)$ denoting $E[x_i|z_i]$, so $E[v_i|z_i] = 0$ is not restrictive and (6) does not need to be the true decision equation (or selection equation). We write the true decision equation as $x_i = r_0(z_i, v_i^*)$ where v_i^* is possibly a vector. The second equation is the outcome equation and it specifies how the decision variable affects the outcome of interest. $f_0(x_i, z_{1i})$ is our parameter of

[4] On the other hand the numerical equivalence of the 2SLS and the classical CF estimators follows from projection theory. Let $\hat{X} = (\hat{x}_1, \ldots, \hat{x}_n)'$ where $\hat{x}_i$ is the fitted regressor from the linear projection of x_i on z_i. Then in matrix formulation $\hat{\beta}_{2SLS} = (\hat{X}'\hat{X})^{-1}\hat{X}'Y$ and $(\hat{\beta}_{CF}, \hat{\rho}_{CF}) = ((X, \hat{V})'(X, \hat{V}))^{-1}(X, \hat{V})'Y$. 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 $\hat{V}$. The estimators are then identical 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 $\hat{V}'\hat{X} = 0$.

[5] For example, if z_i is a scalar and

$$E[\varepsilon_i|z_i, v_i] = \rho_1 v_i + \rho_2 v_i z_i,$$

then including $\hat{v}_i z_i$ in the regression would yield an estimate for $E[\varepsilon_i|z_i, v_i]$ of $\hat{\rho}_1\hat{v}_i + \hat{\rho}_2\hat{v}_i z_i$ which would be consistent. Although this is not typically the object of interest, an exception is when one tests for endogeneity based on the estimate of ρ in (2) (See e.g. Smith and Blundell (1986)).

interest and endogeneity arises because there is dependence between v_i^* and ε_i.

We introduce our estimator for this setup in Section 3.2. It is based on the non-parametric control function estimator of Newey, Powell, and Vella (1999) (NPVCF). They use the orthogonal decomposition from equation (6) and maintain $E[\varepsilon_i|z_i, v_i] = E[\varepsilon_i|v_i]$ to achieve identification of $f_0(x_i, z_{1i})$. The role of the classic CF assumption in their setting is that it rules out the possibility that the control function $E[\varepsilon_i|v_i]$ has an additive functional relationship with (x_i, z_{1i}).

Unlike the linear setting, in the nonlinear/non-parametric setting of Newey, Powell, and Vella (1999) the classic CF assumption is necessary for identification of the structural function $f_0(x_i, z_{1i})$. This assumption can be restrictive because even if ε_i is independent of z_i given the *true* control v_i^*, ε_i needs not be mean independent of z_i conditional on the *pseudo* control v_i from (6). For example, in the simple case when $v_i^* = \varepsilon_i$, if $v_i = \psi(z_i)v_i^*$ with $\psi(z_i) \neq 0$, then $\varepsilon_i = v_i/\psi(z_i)$ and $E[\varepsilon_i|z_i, v_i] = v_i/\psi(z_i) \neq E[\varepsilon_i|v_i]$ unless $\psi(z_i)$ is constant.

3.1 Economic Examples Where the Classic CF Assumption May Not Hold

There are several economic settings where endogeneity is a first-order concern and where ε_i is not necessarily mean independent of the instruments once the classic CF control is conditioned upon. These include estimation of returns to education, production functions, and demand or supply with non-separable reduced forms for equilibrium prices.

We borrow the setup from Imbens and Newey (2009) and Florens, Heckman, Meghir, and Vytlacil (2008) and consider the returns to education and the production function examples together. We let y denote the outcome variable - individual lifetime earnings or firm revenue - and we let x be the agent's choice variable, which is either individual schooling or firm's input into production. ε is the input into production that is unobserved by the econometrician but partially observed by the agent in the sense that she sees a noisy signal η of ε, with η possibly a vector.

We write the output function as $y = f(x) + \varepsilon$ and we let the cost function be given as $c(x, z, \eta)$ where z denotes a cost shifter. The agent optimally chooses x by maximizing the expected profit given the information available to her so the observed x is the solution to

$$x = \arg\max_{\tilde{x}}\{E[f(\tilde{x}) + \varepsilon|z, \eta] - c(\tilde{x}, z, \eta)\}. \tag{8}$$

Assuming differentiability the optimal x solves

$$\partial f(x)/\partial x - \partial c(x, z, \eta)/\partial x = 0, \tag{9}$$

so we have $x = k(z, \eta)$ for some function $k(\cdot)$. By the implicit function theorem we have

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

Without further restrictions on $f(\cdot)$ and $c(\cdot)$, $x = k(z, \eta)$ is neither additively separable in z and η nor is it necessarily monotonic in η when η is a scalar.

We illustrate by considering a simple example where the (educational) production function is given as

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

and the cost function is

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

where ε and $\eta = (\eta_0, \eta_1, \eta_2)$ are unobserved heterogeneity production and cost. Endogeneity arises because of dependence between ε and η. We assume the instruments z are independent of ε and η. From (9) the optimal educational choice x is

$$x = \frac{\varphi_1 - c_1(z, \eta_1)}{c_2(z, \eta_2) - \varphi_2}. \tag{10}$$

In the special case when $c_1(z, \eta_1) = c_{1z}(z) + \eta_1$ and $c_2(z, \eta_2)$ is constant the CF restriction holds with the control $v = x - E[x|z] = -\frac{\eta_1}{c_2 - \varphi_2}$. More generally, if η_1 is not additively separable from z in $c_1(z, \eta_1)$ or if $c_2(\cdot)$ depends on η_2 the CF restriction will not hold.

In our last example we consider a single product monopolistic pricing model in a binary choice setting with logit demands. If $u_{i0} = \epsilon_{i0}$ and $u_{i1} = \beta_0 + \beta_1' X - \alpha p + \xi + \epsilon_{i1}$ with $(\epsilon_{i0,}\epsilon_{i1})$ i.i.d. extreme value and (X, p, ξ) denoting observed characteristics, price, and the unobserved characteristic (to the econometrician), then the market share for good 1 is given by $s = \frac{\exp(\beta_0 + \beta_1' X - \alpha p + \xi)}{1 + \exp(\beta_0 + \beta_1' X - \alpha p + \xi)}$. Let $mc(\cdot)$ denote marginal costs and assume the practitioner observes a cost shifter z that does not enter demand. The monopolist chooses price p such that

$$p = \text{argmax}_p \ (p - mc(\cdot)) \frac{\exp(\beta_0 + \beta_1' X - \alpha p + \xi)}{1 + \exp(\beta_0 + \beta_1' X - \alpha p + \xi)}.$$

While demands can be linearized as

$$\ln s - \ln(1 - s) = \beta_0 + \beta_1' X - \alpha p + \xi,$$

prices will not generally either be separable or necessarily monotonic in ξ. Thus, with $v = p - E[p|z]$, $E[\xi|z, v]$ will not necessarily equal $E[\xi|v]$.

3.2 The Conditional Moment Restriction-Control Function (CM-RCF) estimator

We now describe our estimator. We consider a regression based on our generalized version of the classic CF estimator from Section 2 given as

$$y_i = f_0(x_i, z_{1i}) + h_0(z_i, v_i) + \eta_i \quad \text{with} \quad E[\eta_i|z_i, v_i] = 0 \tag{11}$$

where v_i is given as in (6) and $h_0(z_i, v_i) = E[\varepsilon_i|z_i, v_i]$. Without further restrictions on $h_0(z_i, v_i)$, $f_0(x_i, z_{1i})$ is not identified because $h_0(z_i, v_i)$ can be a function of (x_i, z_{1i}).

We achieve identification by adding the conditional moment restrictions (CMR)

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

which strengthens the unconditional moment restrictions from the linear setting as we must in the non-parametric setting for identification. **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

$$0 = E[\varepsilon_i|z_i] = E[E[\varepsilon_i|z_i, v_i]|z_i] = E[h_0(z_i, v_i)|z_i]. \tag{12}$$

We prove that this restriction suffices for identification of $f_0(x_i, z_{1i})$ in Section 4 and develop the properties of a sieve estimator that can be used to recover $f_0(x_i, z_{1i})$ in Sections 5-7. Our approach loosens the classic **CF** restriction in (3) by combining the generalized CF moment in (11) with the commonly used **CMR** restriction.[6] We refer to our estimator as the **CMRCF** estimator.

We provide a simple example that shows how we can identify $f_0(x_i, z_{1i})$ from an additive regression of y_i on (x_i, z_{1i}) and the control function when $h_0(z_i, v_i)$ satisfies the **CMR** condition. Conditional on (z_i, v_i), the expectation of y_i (from (7)) is equal to

$$E[y_i|z_i, v_i] = f_0(x_i, z_{1i}) + E[\varepsilon_i|z_i, v_i] \equiv f_0(x_i, z_{1i}) + h_0(z_i, v_i) \tag{13}$$

because x_i is known given z_i and v_i. For this example we assume

$$h_0(z_i, v_i) = a_1(\Pi_0(z_i) + v_i) + a_2 v_i^2 + a_3' z_i v_i + \varphi(z_i) = a_1 x_i + a_2 v_i^2 + a_3' z_i v_i + \varphi(z_i)$$

[6]However note that the classic CF restriction does not imply the CMR restriction and vice versa.

where $\varphi(z_i)$ denotes any arbitrary function of z_i. Then the **CMR** condition implies that

$$\begin{aligned}0 &= E[h_0(z_i, v_i)|z_i] = a_1 E[x_i|z_i] + a_2 E[v_i^2|z_i] + a_3' E[z_i v_i|z_i] + E[\varphi(z_i)|z_i] \\ &= a_1 \Pi_0(z_i) + a_2 E[v_i^2|z_i] + \varphi(z_i)\end{aligned}$$

since $E[v_i|z_i] = 0$. It follows that

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

where $\tilde{v}_{2i} = v_i^2 - E[v_i^2|z_i]$. Thus the **CMR** condition puts shape restrictions on $h_0(z_i, v_i)$ so it is not a function of x_i and it does not contain functions of z_i only. Identification in this example is then equivalent to the non-existence of a linear functional relationship among $x_i, z_{1i}, v_i, \tilde{v}_{2i}$, and $z_i v_i$.

Estimation proceeds in three steps. In the first step we obtain the control $\hat{v}_i = x_i - \hat{E}[x_i|z_i]$ from the first stage nonparametric regression (e.g., series estimation in Newey (1997) or sieve estimation in Chen (2007)). In the second step we construct an approximation of $h(z_i, \hat{v}_i)$ using (e.g.) polynomial approximations while imposing the restriction $E[h(z_i, v_i)|z_i] = 0$. For example, we can take

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

where $\varphi_{l_2}(z_i)$ denotes functions of z_i, $L_1, L \to \infty$, $L_1/n, L/n \to 0$ as $n \to \infty$, and we approximate $E[\hat{v}_i^{l_1}|z_i]$ using (possibly nonparametric) regressions. In the last step we estimate $f(x_i, z_{1i})$ by including $h(z_i, \hat{v}_i)$ in the regression, estimating $f(x_i, z_{1i})$ and $h(z_i, \hat{v}_i)$ simultaneously.

An alternative to the control function approach is the non-parametric IV (NPIV) estimator that solves the integral equation implied by the **CMR** condition

$$E[y|z] = E[f_0(x, z_1)|z] = \int f_0(x, z_1)\mu(dx|z)$$

where μ denotes the conditional c.d.f. of x given z (see (e.g.) Newey and Powell (2003), Hall and Horowitz (2005), Darolles, Florens, and Renault (2006), Blundell, Chen, and Kristensen (2007), and Gagliardini and Scaillet (2009), to name only a few). This approach imposes regularity conditions on f_0 and the conditional expectation operator to achieve identification. The control function (CF) approaches do not impose these restrictions but they must impose restrictions on h_0 because the CF approaches estimate both f_0 and h_0.

4 Identification

We ask whether $f_0(x_i, z_{1i})$ is identified by equation (11) with restrictions (12). Our approach to identification closely follows Newey, Powell, and Vella (1999) and Newey and Powell (2003). We consider pairs of functions $\bar{f}(x_i, z_{1i})$ and $\bar{h}(z_i, v_i)$ that satisfy the conditional expectation in (13) and (12). Because conditional expectations are unique with probability one, if there is such a pair $\bar{f}(x_i, z_{1i})$ and $\bar{h}(z_i, v_i)$, it must be that

$$\Pr(f_0(x_i, z_{1i}) + h_0(z_i, v_i) = \bar{f}(x_i, z_{1i}) + \bar{h}(z_i, v_i)) = 1. \tag{14}$$

Identification of $f_0(x_i, z_{1i})$ means we must have $f_0(x_i, z_{1i}) = \bar{f}(x_i, z_{1i})$ whenever (14) holds. Working with differences, we let $\delta(x_i, z_{1i}) = f_0(x_i, z_{1i}) - \bar{f}(x_i, z_{1i})$ and $\kappa(z_i, v_i) = h_0(z_i, v_i) - \bar{h}(z_i, v_i)$, with $E[\kappa(z_i, v_i)|z_i] = 0$ by (12). Identification of $f_0(x_i, z_{1i})$ is then equivalent to

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

Theorem 1 (Identification with CMR). *If equations (11) and (12) are satisfied, then $f_0(x_i, z_{1i})$ is identified if for all $\delta(x_i, z_{1i})$ with finite expectation, $E[\delta(x_i, z_{1i})|z_i] = 0$ implies $\delta(x_i, z_{1i}) = 0$ a.s.*

Proof. Suppose it is not identified. Then we must find functions $\delta(x_i, z_{1i}) \neq 0$ and $\kappa(z_i, v_i) \neq 0$ with $E[\kappa(z_i, v_i)|z_i] = 0$ such that $\Pr(\delta(x_i, z_{1i}) + \kappa(z_i, v_i) = 0) = 1$. But this is not possible because $0 = E[\delta(x_i, z_{1i}) + \kappa(z_i, v_i)|z_i] = E[\delta(x_i, z_{1i})|z_i]$ and $E[\delta(x_i, z_{1i})|z_i] = 0$ implies $\delta(x_i, z_{1i}) = 0$ a.s., so $\Pr(\delta(x_i, z_{1i}) = 0, \kappa(z_i, v_i) = 0) = 1$. $\square$

The result implies that $h_0(z_i, v_i)$ is also identified because the conditional expectation $E[y_i|z_i, v_i]$ is nonparametrically identified and $h_0(z_i, v_i) = E[y_i|z_i, v_i] - f_0(x_i, z_{1i})$.

We consider several cases, with the regressors demeaned in each example. For the simple model $f_0(x_i, z_{1i}) = \beta_0 x_i$, we have the alternative function $\tilde{f}(x_i, z_{1i}) = \tilde{\beta} x_i \neq \beta_0 x_i$. We have $\delta(x_i, z_{1i}) = (\beta_0 - \tilde{\beta})x_i$, so $E[\delta(x_i, z_{1i})|z_i] = 0$ implies $\delta(x_i, z_{1i}) = 0$ (or $\beta_0 = \tilde{\beta}$) as long as $E[x_i|z_i] \neq 0$. Identification is then equivalent to z_i being correlated x_i, the standard instrumental variable condition.

The general case is given by $f_0(x_i, z_{1i}) = \beta_0' x_i + \beta_{10}' z_{1i}$. An alternative function is $\tilde{f}(x_i, z_{1i}) = \tilde{\beta}' x_i + \tilde{\beta}_1' z_{1i} \neq \beta_0' x_i + \beta_{10}' z_{1i}$, so $E[\delta(x_i, z_{1i})|z_i] = (\beta_0 - \tilde{\beta})' E[x_i|z_i] + (\beta_{10} - \tilde{\beta}_1)' z_{1i}$. Therefore $E[\delta(x_i, z_{1i})|z_i] = 0$ implies $\delta(x_i, z_{1i}) = 0$ - or $\beta_0 = \tilde{\beta}$ and $\beta_{10} = \tilde{\beta}_1$ - if z_i satisfies the standard rank condition (e.g., it includes excluded instruments from z_{1i} that are correlated with x_i).

For the general non-parametric case, a sufficient condition for identification is that the conditional distribution of x_i given z_i satisfies the completeness condition (see Newey and

Powell (2003) or Hall and Horowitz (2005)). The condition implies that $E[\delta(x_i, z_{1i})|z_i] = 0$ implies $\delta(x_i, z_{1i}) = 0$ for any $\delta(x_i, z_{1i})$ with finite expectation. In this sense the completeness condition is the nonparametric analog of the rank condition for identification in the linear setting.

5 Estimation

Our estimator is obtained in three steps. We focus on sieve estimation because it is convenient to impose the restriction (12). We use capital letters to denote random variables and lower case letters to denote their realizations. We assume the tuple $\{(Y_i, X_i, Z_i)\}$ for $i = 1, \ldots, n$ are i.i.d. We let X_i be $d_x \times 1$, Z_{1i} be $d_1 \times 1$, Z_{2i} be $d_2 \times 1$, $d_z = d_1 + d_2$ and $d = d_z + d_x$, with $d_x = 1$ for ease of exposition. Let $\{p_j(Z), j = 1, 2, \ldots\}$ denote a sequence of approximating basis functions (e.g. orthonormal polynomials or splines). Let $p^{k_n} = (p_1(Z), \ldots, p_{k_n}(Z))'$, $P = (p^{k_n}(Z_1), \ldots, p^{k_n}(Z_n))'$, and $(P'P)^-$ denote the Moore-Penrose generalized inverse, where k_n tends to infinity but $k_n/n \to 0$. Similarly we let $\{\phi_j(X, Z_1), j = 1, 2, \ldots\}$ denote a sequence of approximating basis functions, $\phi^{Kn} = (\phi_1(X, Z_1), \ldots, \phi_{K_n}(X, Z_1))'$, where K_n tends to infinity but $K_n/n \to 0$.[7]

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

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

and obtain the control variable as $\hat{v} = x - \hat{\Pi}(z)$.

In the second step we construct approximating basis functions using $\hat{v}$ and z, where we impose the CMR condition (12) by subtracting out the conditional means (conditional on Z). We start by assuming v is known and then show how the setup changes when $\hat{v}$ replaces v. We write basis functions when v is known as

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

where $\bar{\varphi}_l(z) = E[\varphi_l(Z, V)|Z = z]$ and $\{\varphi_l(z, v), l = 1, 2, \ldots\}$ denotes a sequence of approximating basis functions generated using $(z, v) \in \mathcal{Z} \times \mathcal{V} \equiv \mathcal{W}$, 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 $\mathcal{H}$.

[7] 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.

We define the sieve space $\mathcal{H}_n$ as the collection of functions

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

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

Because v is not known we use instead estimates of the approximating basis functions, which we denote as $\hat{\tilde{\varphi}}_l(z, \hat{v}) = \varphi_l(z, \hat{v}) - \hat{\bar{\varphi}}_l(z)$, where $\hat{\bar{\varphi}}_l(z) = \hat{E}[\varphi_l(Z, \hat{V})|Z = z]$. We then construct the approximation of $h(z, v)$ as [8]

$$\begin{aligned} \hat{h}_{L_n}(z, \hat{v}) &= \sum\nolimits_{l=1}^{Ln} a_l \{\varphi_l(z, \hat{v}) - \hat{E}[\varphi_l(Z, \hat{V})|Z = z]\} \\ &= \sum\nolimits_{l=1}^{Ln} a_l \{\varphi_l(z, \hat{v}) - p^{k_n}(z)'(P'P)^- \sum\nolimits_{i=1}^{n} p^{k_n}(z_i)\varphi_l(z_i, \hat{v}_i)\}, \end{aligned} \tag{15}$$

with coefficients, $(a_1, \ldots, a_{L_n})$ to be estimated in the last step. We approximate the sieve space $\mathcal{H}_n$ with $\hat{\mathcal{H}}_n$ using (15), so $\hat{\mathcal{H}}_n$ is given by

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

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

$$\mathcal{F}_n = \{f : f = \sum\nolimits_{l \le K_n} \beta_l \phi_l(x, z_1), \|f\|_{\mathcal{F}} < \bar{C}_f, (x, z_1) \in \mathcal{X} \times \mathcal{Z}_1\}$$

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

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

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

Equivalently we can write the estimation problem as

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

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

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

6 Convergence rates

We obtain the convergence rates building on Newey, Powell, and Vella (1999). We differ from their approach as we have another nonparametric estimation stage in the middle step of estimation that creates additional terms in the convergence rate results. We derive the mean-squared error convergence rates of the nonparametric estimator $\hat{f}(\cdot)$ and $\hat{h}(\cdot)$, which we later use to obtain the $\sqrt{n}$-consistency and the asymptotic normality of the linear functionals of (f_0, h_0).

We introduce additional notation. We let $g_0(z_i, v_i) = f_0(x_i, z_{1i}) + h_0(z_i, v_i)$ be a function of (z_i, v_i) (x_i is fixed given (z_i, v_i)). For a random matrix D, let $\|D\| = (\mathrm{tr}(D'D))^{1/2}$, and let $\|D\|_\infty$ be the infimum of constants C such that $\Pr(||D|| < C) = 1$. Assumptions C1 and C2 together ensure that we obtain the mean-squared error convergence of $\hat{g} = \hat{f} + \hat{h}$ to g_0, and so that of $\hat{f}$ to f_0, too.

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

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

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

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

Theorem 2. *Suppose Assumptions C1-C2 are satisfied. Then*

$$\left(\int (\hat{g}(z,v) - g(z,v))^2 d\mu_0(z,v)\right)^{1/2} = O_p(\sqrt{\mathbf{L}_n/n} + L_n\triangle_n + \mathbf{L}_n^{-s/d}).$$

where $\mu_0(z,v)$ *denotes the distribution function of* (z,v).

In Theorem 2 the term $L_n\triangle_n$ arises because of the estimation error from the first and second steps of estimation. With no estimation error from these stages we would obtain the convergence rate of $O_p(\sqrt{\mathbf{L}_n/n} + \mathbf{L}_n^{-s/d})$, which is a standard convergence rate of series estimators.

7 Asymptotic Normality

Following Newey (1997) and Newey, Powell, and Vella (1999) we consider inference for the linear functions of g, $\theta = \alpha(g)$ where we also need to account for the multi-stage estimation of g as described in Section 5. The estimator $\hat{\theta} = \alpha(\hat{g})$ of $\theta_0 = \alpha(g_0)$ is a well-defined "plug-in" estimator, and because of the linearity of $\alpha(g)$ we have

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

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

We focus on conditions that provide for $\sqrt{n}$-asymptotics and allow for a straightforward consistent estimator for the standard errors of $\hat{\theta}$.[9] If there exists a Riesz representer $\nu^*(Z,V)$ such that

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

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

$$\begin{aligned} \Omega \;=\; & E[\nu^*(Z,V)\mathrm{var}(Y|Z,V)\nu^*(Z,V)'] + E[\rho_v(Z)\mathrm{var}(X|Z)\rho_v(Z)'] \\ & + \lim_{n\to\infty} \sum_{l=1}^{L_n} E[\rho_{\bar{\varphi}_l}(Z)\mathrm{var}(\varphi_l(Z,V)|Z)\rho_{\bar{\varphi}_l}(Z)']. \end{aligned} \tag{16}$$

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.

Assumption C1, R1, N1, and N2 below are sufficient for us to characterize the asymptotic normality of $\hat{\theta}$ and also a consistent estimator for the asymptotic variance of $\hat{\theta}$. Let $\psi^{\mathbf{L}}(z_i,v_i) \equiv (\phi_1(x_i,z_{1i}),\ldots,\phi_K(x_i,z_{1i}),\tilde{\varphi}^L(z_i,v_i)')'$ and $\tilde{\varphi}^L(z_i,v_i) = (\tilde{\varphi}_1(z_i,v_i),\ldots,\tilde{\varphi}_L(z_i,v_i))'$.

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

To present the theorem, we need additional notation and assumptions. Let $a_L = (a_1,\ldots,a_L)'$ with an abuse of notation and for any differentiable function $c(w)$, let $|\mu| = \sum_{j=1}^{\dim(w)} \mu_j$ and define $\partial^\mu c(w) = \partial^{|\mu|}c(w)/\partial w_1 \cdots \partial w_{\dim(w)}$. Also define $|c(w)|_\delta = \max_{|\mu|\le\delta} \sup_{w\in\mathcal{W}} ||\partial^\mu c(w)||$ and others are defined similarly.

Assumption 4 (N1)**.** *(i) there exist δ, γ, and $\beta_{\mathbf{L}}$ such that $|g_0(z,v) - \beta_{\mathbf{L}}'\psi^{\mathbf{L}}(z,v)|_\delta \le C\mathbf{L}^{-\gamma}$*

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

(which also implies $|h_0(z,v)-a_L'\tilde{\varphi}^L(z,v)|_\delta \leq CL^{-\gamma}$); (ii) var$(Y_i|Z_i,V_i)$ is bounded away from zero, $E[\eta_i^4|Z_i,V_i]$ and $E[V_i^4|Z_i]$ are bounded and $E[\tilde{\varphi}_l(Z_i,V_i)^4|Z_i]$ is bounded for all l.

The assumption N1 (i) is satisfied for f_0 and h_0 that belong to the Hölder class. Then we can take (e.g.) $\gamma = s/d$. Next we impose the rate conditions that restrict the growth of k_n and $\mathbf{L}_n = K_n + L_n$ as n tends to infinity.

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

Theorem 3. *Suppose Assumptions C1, R1, and N1-N2 are satisfied. Then*

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

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

$$\begin{aligned}
\hat{\mathcal{T}} &= \sum\nolimits_{i=1}^n \hat{\psi}_i^{\mathbf{L}}\hat{\psi}_i^{\mathbf{L}\prime}/n, \hat{\Sigma} = \sum\nolimits_{i=1}^n (y_i-\hat{g}(z_i,\hat{v}_i))^2\hat{\psi}_i^{\mathbf{L}}\hat{\psi}_i^{\mathbf{L}\prime}/n \qquad (17)\\
\hat{\mathcal{T}}_1 &= P'P/n, \hat{\Sigma}_1 = \sum_{i=1}^n \hat{v}_i^2 p^k(z_i)p^k(z_i)'/n, \hat{\Sigma}_{2,l} = \sum_{i=1}^n \{\varphi_l(z_i,\hat{v}_i)-\hat{\bar{\varphi}}_l(z_i)\}^2 p^k(z_i)p^k(z_i)'/n\\
\hat{H}_{11} &= \sum_{i=1}^n\sum_{l=1}^L \hat{a}_l\frac{\partial\varphi_l(z_i,\hat{v}_i)}{\partial v_i}\hat{\psi}_i^{\mathbf{L}}p^k(z_i)'/n,\\
\hat{H}_{12} &= \sum_{i=1}^n p^k(z_i)'((P'P)^-\sum_{j=1}^n p^k(z_j)\frac{\partial\sum_{l=1}^L\hat{a}_l\varphi_l(z_j,\hat{v}_j)}{\partial v_j})\hat{\psi}_i^{\mathbf{L}}p^k(z_i)'/n,\\
\hat{H}_{2,l} &= \sum\nolimits_{i=1}^n \hat{a}_l\hat{\psi}_i^{\mathbf{L}}p^k(z_i)'/n, \hat{H}_1 = \hat{H}_{11}-\hat{H}_{12}.
\end{aligned}$$

Then, we can estimate Ω consistently by

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

Theorem 4. *Suppose Assumptions C1, R1, and N1-N2 are satisfied. Then $\hat{\Omega} \to_p \Omega$.*

This is the heteroskedasticity robust variance estimator that accounts for the first and second steps of estimation. The first variance term $\mathcal{A}\hat{\mathcal{T}}^{-1}\hat{\Sigma}\hat{\mathcal{T}}^{-1}\mathcal{A}'$ corresponds to the variance estimator without error from the first and second steps of estimation. The second variance term accounts for the estimation of v (and corresponds to the second term in (16)). The third variance term accounts for the estimation of $\bar{\varphi}_l(\cdot)$'s) (and corresponds to the third term in (16)). If we view our model as a parametric one with fixed k_n, K_n, and L_n, the same variance estimator $\hat{\Omega}$ can be used as the estimator of the variance for the parametric model (e.g, Newey (1984) and Murphy and Topel (1985)).

7.1 Discussion

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

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

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

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

where $\nu^*(z, v) = (E[q(Z, V)q(Z, V)'])^{-1}q(z, v)$ and $q(z, v)$ is the residual from the mean-square projection of x_1 on the space of functions that are additive in (x_{-1}, z_1) and any $h(z, v)$ such that $E[h(Z, V)|Z] = 0$.[10] Thus we can approximate $q(z, v)$ by the mean-square projection residual of x_1 on $\psi^{\mathbf{L}}_{-1}(z_i, v_i) \equiv (\phi_1(x_{-1i}, z_{1i}), \ldots, \phi_K(x_{-1i}, z_{1i}), \tilde{\varphi}^L(z_i, v_i)')'$, and then use these estimates to approximate $\nu^*(z, v)$.

Next consider a weighted average derivative of the form

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

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

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

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

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

8 Simulation Study

We conduct two types of monte carlos to evaluate the performance of the classic CF estimator, the NPVCF estimator, and our CMRCF estimator. The first set of monte carlos is based on the economic examples provided in Section 3.1 where the structural function $f(x)$ is parametric and the second set uses a non-parametric setup from Newey and Powell (2003) where the structural function is estimated nonparametrically.

8.1 Monte Carlos Based on Parametric Estimators

We consider six models motivated by the economic examples from Section 3.1. The outcome equations are parametric so $f(x)$ is known up to a finite set of parameters. The selection equations are treated as unknown to the practitioner and we use nonparametric estimators for them in the simulation.

The six designs are given as:

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

These designs can be obtained from the underlying decision problem of (8) by varying the structural function $f(x)$ and the cost function $c(x, z, \eta)$. For example we obtain design [1] by letting $c_2(z, \eta_2)$ be constant and $c_1(z, \eta_1)$ include the leading term z and the interaction term $\eta_1 \log(z)$, where $\eta_1 = 3\varepsilon + \varsigma$ is a noisy signal of ε. The selection equation (10) is then $x = \frac{\varphi_1 - c_1(z,\eta_1)}{c_2(z,\eta_2) - \varphi_2} = z + (3\varepsilon + \varsigma) \cdot \log(z)$. The other designs are derived in a similar way.

We generate simulation data based on the following distributions: $\varepsilon_i \sim U_\varepsilon$, $\varsigma_i \sim U_\varsigma$, $z_i = 2 + 2U_z$, where each U_ε, U_ς, and U_z independently follows the uniform distribution supported on $[-1/2, 1/2]$ so all three random variables ε_i, ς_i, and z_i are independent of one another. In all designs x_i is correlated with ε_i and the CMR condition, $E[\varepsilon_i|z_i] = 0$ holds. The **CF** restriction is violated in designs [1]-[5] and holds in design [6].[11] We set the true parameter values at $(\alpha_0, \beta_0, \gamma_0) = (1, 1, -1)$ and the data is generated with the sample size of $n = 1,000$.

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

All three estimators 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 estimates are robust to adding higher order terms.[12] The classic CF (CCF) estimates

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

using least squares where $f(x_i)$ is given by the designs [1]-[6]. The NPVCF estimator is obtained by estimating

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

where we approximate $h(\hat{v}_i)$ as $h(\hat{v}_i) = \sum_{l=1}^{5} a_l \hat{v}_i^l$.[13] Since the NPVCF 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 adding higher orders of polynomials to fit $h(\hat{v}_i)$.

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

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

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

We report the biases and the RMSE's based on 200 repetitions of the estimations. The simulation results in Tables I-VI show that CCF and NPVCF are biased in all designs except [5] and [6] for which the theory says they should be consistent. The CMRCF is robust regardless of the designs. In design [5] all three approaches produce correct estimates because the outcome equation is linear, which is consistent with our discussion in Section 2. In design [6] all three approaches are consistent because the **CF** restriction holds. We conclude that our CMRCF approach is consistent in these designs regardless of whether the model is linear or nonlinear or whether the **CF** restriction holds while the CCF and NPVCF approaches are not robust when the **CF** restriction does not hold.

[12] Root mean-squared errors were similar across all estimators whether we used two or more higher order terms. Thus if we followed Newey, Powell, and Vella (1999) and used cross validation (CV) to discriminate between alternative specifications we would be indifferent between this simplest specification and the ones with the higher order terms.

[13] We do not use the trimming device in Newey, Powell, and Vella (1999). Trimming is not necessary in these examples because the supports of variables are compact and tightly bounded.

8.2 Monte Carlos Based on Non-Parametric Estimators

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

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

$$\begin{aligned} y &= f(x) + \varepsilon = \ln(|x-1|+1)\mathrm{sgn}(x-1) + \varepsilon \\ [A]\; x &= z + \eta \end{aligned}$$

where the errors ε and η and instruments z are generated by

$$\begin{pmatrix} \varepsilon \\ \eta \\ z \end{pmatrix} \sim \text{i.i.d } N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho & 0 \\ \rho & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}\right)$$

with $\rho = 0.5$. This design satisfies the CF restriction with $v = x - E[x|z]$ because $v = \eta$.

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

$$[B]\; x = z + \eta / \exp(|z|)$$

and we use $\rho = 0.5$ and $\rho = 0.9$. The CF restriction is violated because $v = x - E[x|z] = \eta/\exp(|z|)$.

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

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

We estimate $f(x)$ using the classic CF approach, the NPVCF estimator and our CMRCF estimator. We fix $J = 5$ for design [A] and $J = 7$ for the design [B] and we use four different sample sizes (n=100, 400, 1000, and 2,000). In all of the designs we obtain the control using the first stage estimation residual $\hat{v}_i = x_i - (\hat{\pi}_0 + \hat{\pi}_1 z_i + \hat{\pi}_2 z_i^2)$. We experimented with adding several higher order terms in the first stage and found very similar simulation results across all three estimators. We also experimented with different choices of approximating functions of $h(v)$ and $h(z, v)$ for design [B].

The results are summarized in Tables A and B. We report the root mean-squared-error (RMSE) averaged across the 500 replications and the realized values of x. In both designs RMSE decreases as the sample size increases for all estimators. The RMSEs for non-parametric least squares (NPLS) (that does not include any control function in the estimation) are larger than RMSEs for the estimators that correct for endogeneity. In the design [A] both NPVCF and CMRCF perform similarly although the CMRCF estimator shows slightly larger RMSEs because it adds an irrelevant correction term (zv) in the control function. In the design [B] the CMRCF estimator dominates the NPVCF estimator in terms of RMSE.

Table A: Design [A], RMSE

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

Table B: Design [B], RMSE

	NPLS	NPVCF1	NPVCF2	CMRCF1	CMRCF2	CMRCF3
CF's	None	$\sum_{l=1}^{4} a_l\hat{v}^l$	$\sum_{l=1}^{5} a_l\hat{v}^l$	$a_1\hat{v} + a_2 z\hat{v}$	$a_1\hat{v} + a_2 z\hat{v} + a_3 z^2\hat{v}$	$a_1\hat{v} + a_2 z\hat{v} + a_3 z^2\hat{v} + a_4\tilde{v}_2$
	$\rho = 0.5$					
n=100	0.3896	0.3233	0.3241	0.3049	0.3104	0.3231
n=400	0.2775	0.1679	0.1671	0.1540	0.1422	0.1456
n=1000	0.2511	0.1364	0.1358	0.1190	0.0999	0.1014
n=2000	0.2440	0.1150	0.1156	0.0968	0.0737	0.0745
	$\rho = 0.9$					
n=100	0.5277	0.3059	0.3018	0.2833	0.2734	0.2889
n=400	0.4462	0.2042	0.2003	0.1680	0.1296	0.1322
n=1000	0.4375	0.1885	0.1865	0.1483	0.0941	0.0950
n=2000	0.4311	0.1762	0.1773	0.1356	0.0690	0.0696

We graph the average value of the function estimates $\hat{f}(x)$ from the three estimators against the true value of $f(x)$ (dashed line). The NPLS estimates are the light solid line, the CMRCF function estimates are the solid line and the NPVCF estimates are the dotted-and-dashed line. The upper and lower two standard deviation limits for the simulated distributions of $\hat{f}(x)$ for the CMRCF are given by the dotted lines. Both NPVCF estimators have almost identical RMSEs and we use NPVCF2 in Table B although NPVCF1 generated almost identical results.[14] We use the CMRCF2 from Table B because it has the smallest RMSE among the three CMRCF specifications.

In both designs the nonparametric estimator without the correction for endogeneity is substantially biased and often strays outside the simulated confidence interval of the CMRCF estimates. In design [A] where the CF restriction holds the CMRCF estimator and the NPVCF are almost identical. In design [B] where the CF restriction does not hold, the CMRCF estimator performs better than the NPVCF estimator which at some points approaches or strays outside the simulated confidence interval of the CMRCF estimator. The problem becomes worse as the sample size increases or when the endogeneity increases ($\rho = 0.5$ to $\rho = 0.9$). In these Monte Carlos motivated by the setup from Newey and Powell (2003) our proposed CMRCF estimator is robust to violations of the CF restriction while the NPVCF estimator is not.

9 Conclusion

We show that the classic CF estimator can be modified to allow the mean of the error to depend in a general way on the instruments and control. We do so by replacing the classic CF restriction with a generalized CF moment condition combined with the moment restrictions maintained by two-stage least squares. If the outcome equation is nonlinear or nonparametric in the endogenous regressor, then both the classical CF estimator and the NPVCF estimator of Newey, Powell, and Vella (1999) are inconsistent when the classic CF restriction does not hold. This restriction is often violated in economic settings including returns to education, production functions, and demand or supply with non-separable reduced forms for equilibrium prices. We use our results from the linear setting to develop an estimator robust to settings where the structural error depends on the instruments given the CF control. We augment the NPVCF setting with conditional moment restrictions and our estimator maintains the simplicity of the NPVCF estimator. In our simulation studies which are based on our economic examples we find that the classic CF estimator and the NPVCF estimator are biased when the CF restriction is violated while our estimator remains consistent.

[14] In this comparison we use the NPVCF estimator that possibly overfits $h(v)$ because we are more interested in the biases of estimators when the CF restriction does not hold.

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

Nonlinear & **CF** condition does not hold

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

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

Nonlinear & **CF** condition does not hold

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

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

Nonlinear & **CF** condition does not hold

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

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

Nonlinear & **CF** condition does not hold

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

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

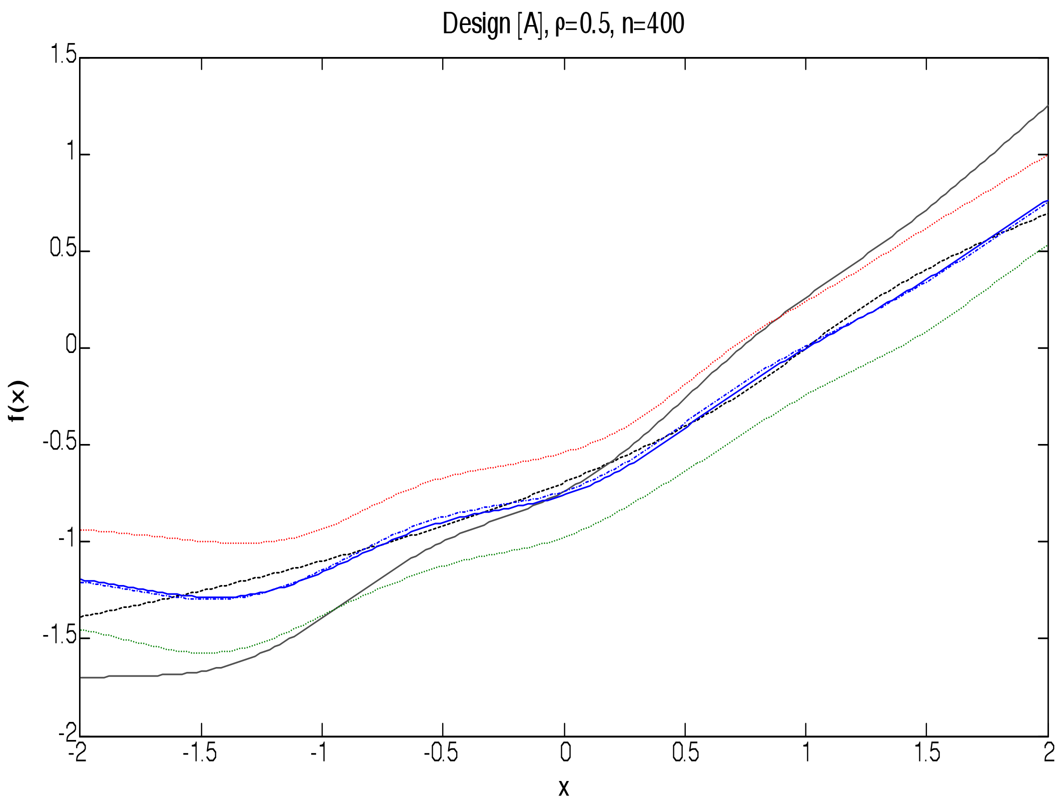
Linear & **CF** condition does not hold

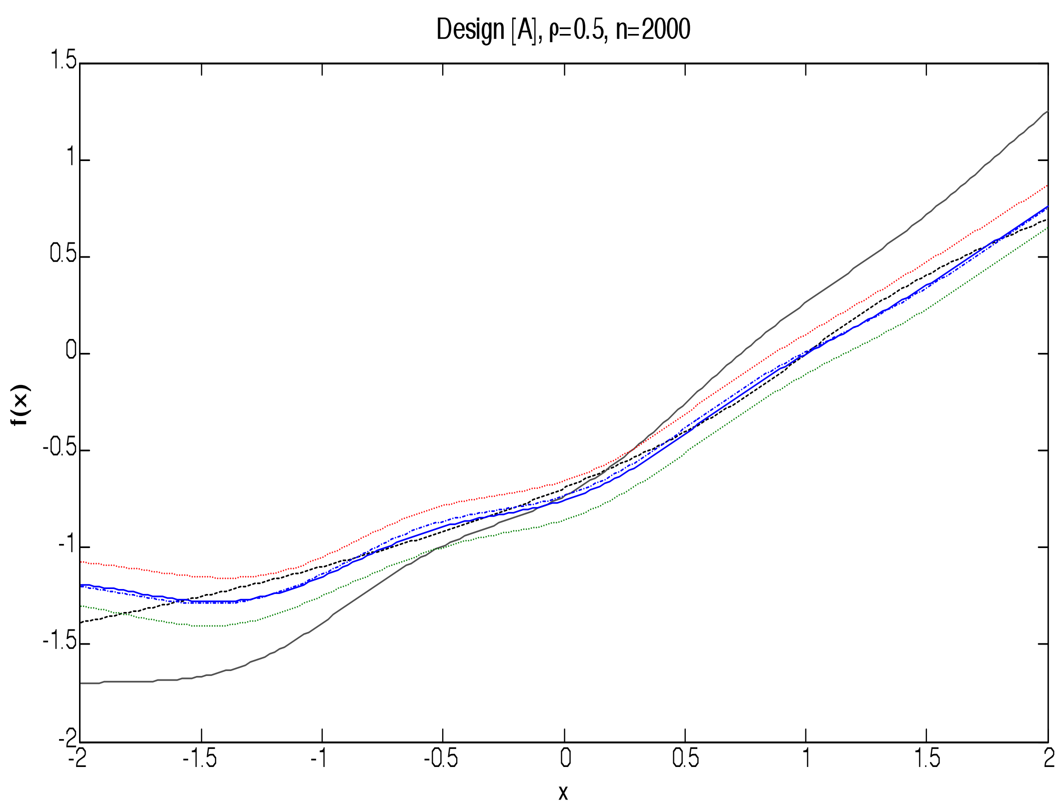
		mean	bias	RMSE
CCF	α	0.9993	-0.0007	0.0343
	β	1.0004	0.0004	0.0172
NPVCF	α	1.0010	0.0010	0.0417
	β	0.9997	-0.0003	0.0192
CMRCF	α	0.9991	-0.0009	0.0343
	β	1.0005	0.0005	0.0171

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

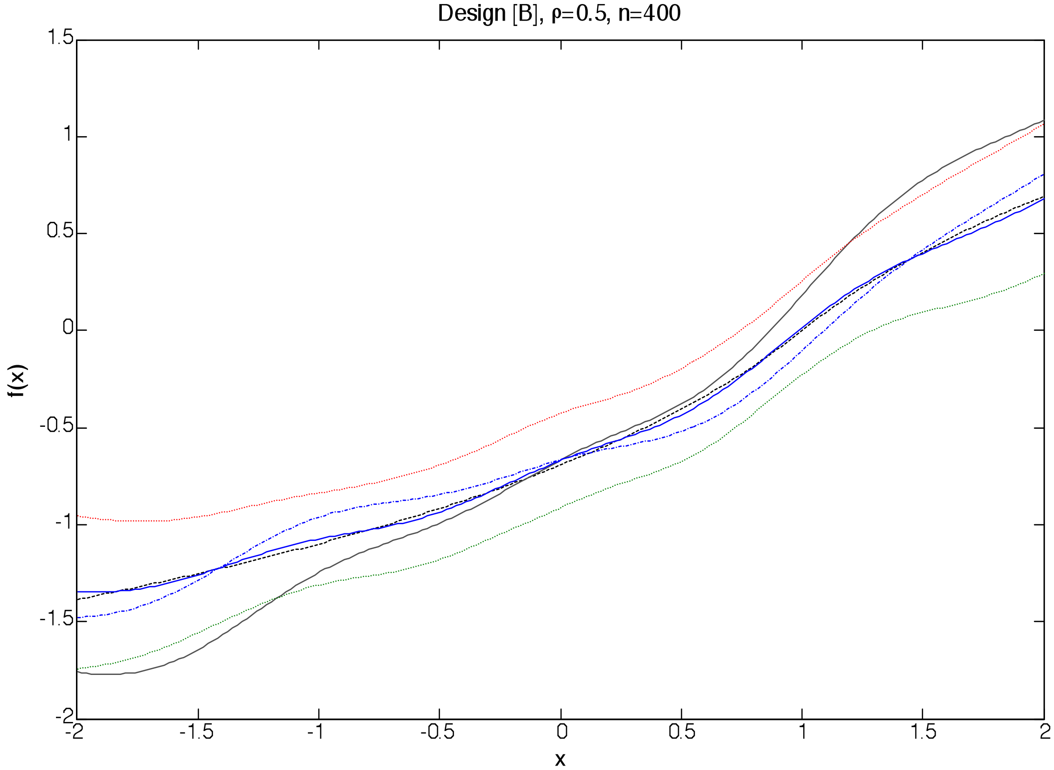
Nonlinear & **CF** condition holds

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

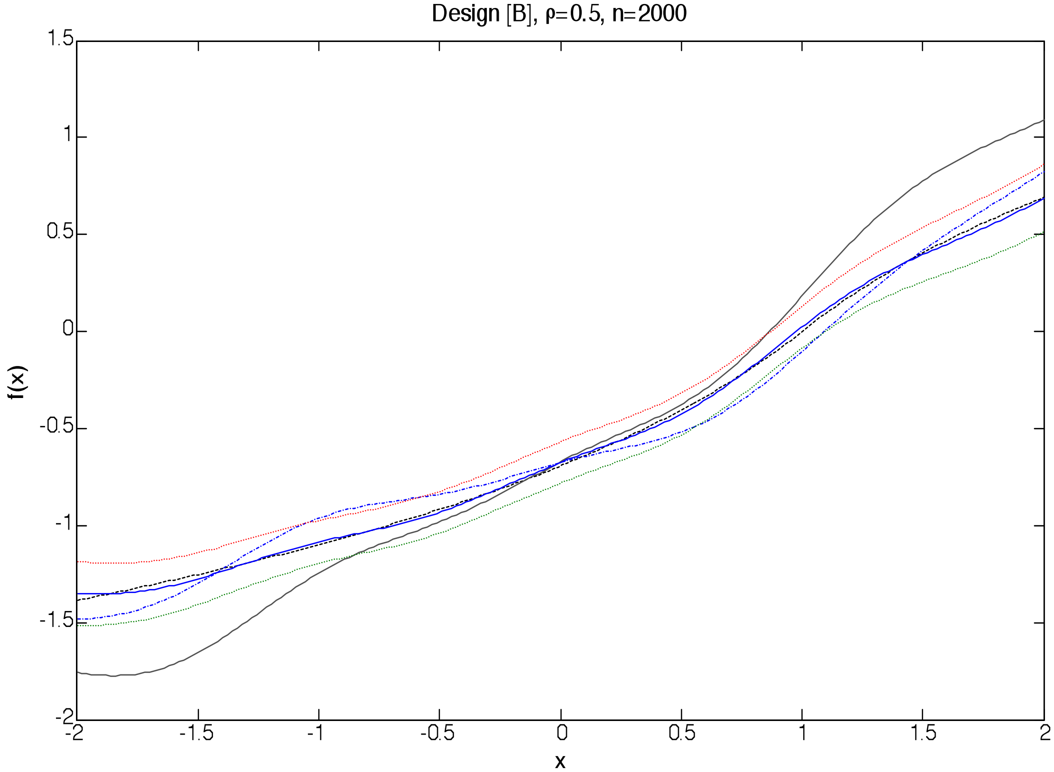

Design [A], ρ=0.5, n=400
f(x)
x
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2

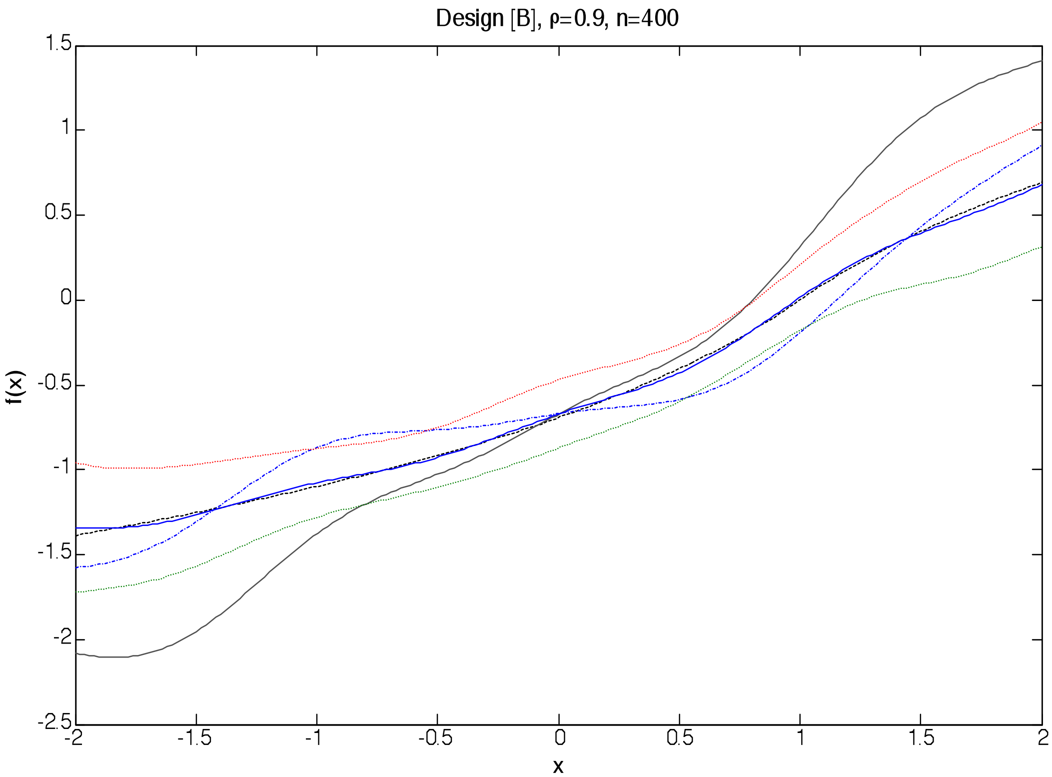

Design [A], ρ=0.5, n=2000
f(x)
x
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2

Design [B], ρ=0.5, n=400

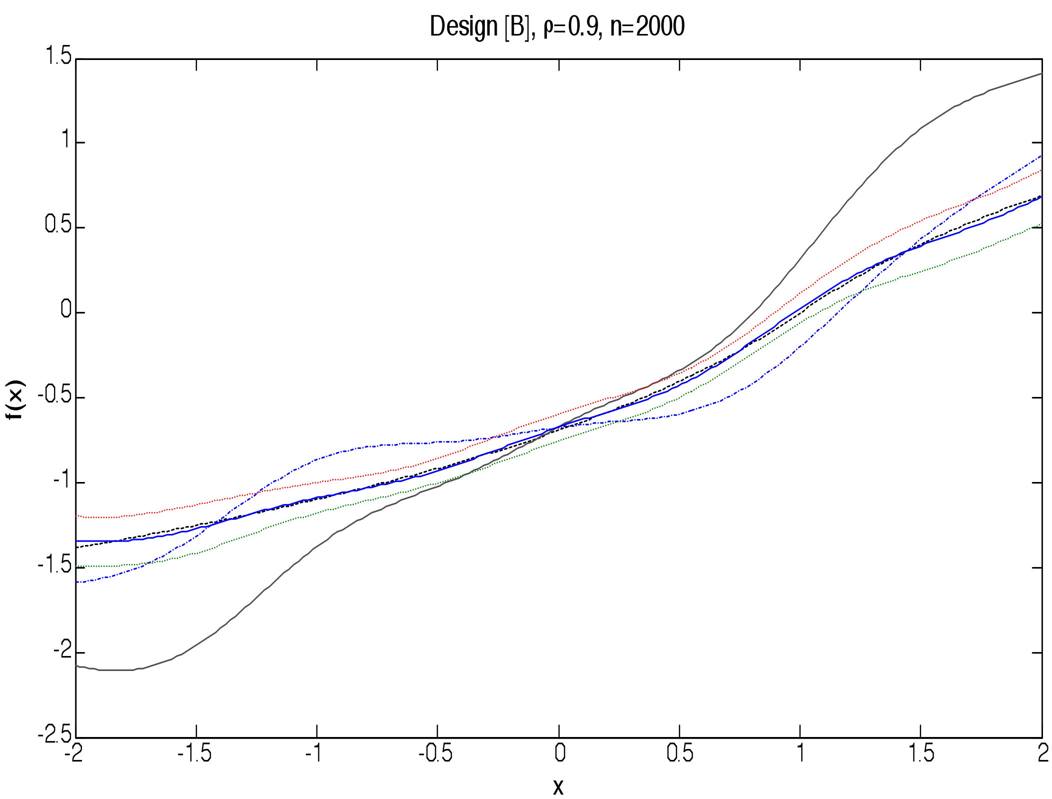


Design [B], ρ=0.5, n=2000





Design [B], ρ=0.9, n=400
f(x)
x
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2.5
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2



Design [B], ρ=0.9, n=2000
f(x)
x
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2.5
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2

Appendix

A Asymptotic irrelevance of the generalized control function in the classic CF approach

Theorem 5. *Assume (i) $E[\|z_i\| \cdot ||\tilde{h}(z_i, v_i)||] < \infty$, (ii) $\tilde{h}(z, v)$ is differentiable with respect to v, (iii) for $v_i(\pi) \equiv x_i - z_i'\pi$, assume $\sup_{\pi^* \in \Pi_0} E[\|z_i\|^2 \left\| \frac{\partial \tilde{h}(z_i, v_i(\pi^*))}{\partial v_i} \right\|] < \infty$ for Π_0 some neighborhood of π_0, (iv) assume $E[\|z_i\|^2 \left\| \frac{\partial \tilde{h}(z_i, v_i(\pi))}{\partial v_i} \right\|]$ is continuous at $\pi = \pi_0$, and (v) $\hat{\pi} \to_p \pi_0$. If (1) holds then $Z' M_{\hat{V}} \tilde{H}(Z, \hat{V})/n \to_p 0$ as $n \to \infty$.*

Proof. We can rewrite as

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

because $Z'\hat{V} = 0$. 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 $\sum_{i=1}^n z_i \tilde{h}(z_i, v_i)/n \to_p E[z_i \tilde{h}(z_i, v_i)]$ by the law of large numbers under (i). Obtain $|| \sum_{i=1}^n z_i(\tilde{h}(z_i, \hat{v}_i) - \tilde{h}(z_i, v_i))/n|| \le \|\hat{\pi}^* - \pi_0\| \sum_{i=1}^n \|z_i\|^2 \, || \frac{\partial \tilde{h}(z_i, v_i(\hat{\pi}^*))}{\partial v_i} ||/n$ by applying the mean-value expansion, where $\hat{\pi}^*$ lies between $\hat{\pi}$ and π_0 and $v_i(\pi) = x_i - z_i'\pi$. Then the term $\sum_{i=1}^n z_i(\tilde{h}(z_i, \hat{v}_i) - \tilde{h}(z_i, v_i))/n \to_p 0$ by the consistency of $\hat{\pi}$ and $\sum_{i=1}^n \|z_i\|^2 \left\| \frac{\partial \tilde{h}(z_i, v_i(\hat{\pi}^*))}{\partial v_i} \right\| /n \to_p E[\|z_i\|^2 \left\| \frac{\partial \tilde{h}(z_i, v_i(\pi_0))}{\partial v_i} \right\|] < \infty$ under (iii) and (iv). Therefore $\sum_{i=1}^n z_i \tilde{h}(z_i, \hat{v}_i)/n \to_p E[z_i \tilde{h}(z_i, v_i)] = 0$ by (1) and (4). $\square$

B Proof of convergence rates

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

Define $h_L(z, v) = a_L' \tilde{\varphi}^L(z, v)$ and $\hat{h}_L(z, v) = a_L' \hat{\tilde{\varphi}}^L(z, v)$ where a_L[15] satisfies Assumption L1 (iv). Define $\psi_i^{\mathbf{L}}(z_i, v_i) = (\phi_1(x_i, z_{1i}), \ldots, \phi_K(x_i, z_{1i}), \tilde{\varphi}^L(z_i, v_i)')'$ where $\tilde{\varphi}^L(z_i, v_i) = (\tilde{\varphi}_1(z_i, v_i), \ldots, \tilde{\varphi}_L(z_i, v_i))'$ and $\hat{\psi}_i^{\mathbf{L}}(z_i, v_i) = (\phi_1(x_i, z_{1i}), \ldots, \phi_K(x_i, z_{1i}), \hat{\tilde{\varphi}}^L(z_i, v_i)')'$ with $\hat{\tilde{\varphi}}^L(z_i, v_i) = (\hat{\tilde{\varphi}}_1(z_i, v_i), \ldots, \hat{\tilde{\varphi}}_L(z_i, v_i))'$. We further let $\hat{\hat{\psi}}_i^{\mathbf{L}} = \hat{\psi}^{\mathbf{L}}(z_i, \hat{v}_i)$, $\psi_i^{\mathbf{L}} = \psi^{\mathbf{L}}(z_i, v_i)$, and $\hat{\psi}_i^{\mathbf{L}} = \hat{\psi}^{\mathbf{L}}(z_i, v_i)$. We further let $\psi^{\mathbf{L},n} = (\psi_1^{\mathbf{L}}, \ldots, \psi_n^{\mathbf{L}})'$, $\hat{\psi}^{\mathbf{L},n} = (\hat{\psi}_1^{\mathbf{L}}, \ldots, \hat{\psi}_n^{\mathbf{L}})'$, and $\hat{\hat{\psi}}^{\mathbf{L},n} = (\hat{\hat{\psi}}_1^{\mathbf{L}}, \ldots, \hat{\hat{\psi}}_n^{\mathbf{L}})'$.

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

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

Assumption 6 (L1)**.** *(i) (X, Z, V) is continuously distributed with bounded density; (ii) for each k, L, and $\mathbf{L} = K + L$ there are nonsingular matrices B_1, B_2, and B such that for $p^k_{B_1}(z) = B_1 p^k(z)$, $\tilde{\varphi}^L_{B_2}(z, v) = B_2 \tilde{\varphi}^L(z, v)$, and $\psi^{\mathbf{L}}_B(z, v) = B\psi^{\mathbf{L}}(z, v)$, $E[p^k_{B_1}(Z_i)p^k_{B_1}(Z_i)']$, $E[\tilde{\varphi}^L_{B_2}(Z_i, V_i)\tilde{\varphi}^L_{B_2}(Z_i, V_i)']$, and $E[\psi^{\mathbf{L}}_B(Z_i, V_i)\psi^{\mathbf{L}}_B(Z_i, V_i)']$ have smallest eigenvalues that are bounded away from zero, uniformly in k, L, and $\mathbf{L}$; (iii) for each integer $\delta > 0$, there are $\zeta_\delta(\mathbf{L})$ and $\xi_\delta(k)$ such that $|\psi^{\mathbf{L}}(z, v)|_\delta \leq \zeta_\delta(\mathbf{L})$ (this also implies that $|\tilde{\varphi}^L(z, v)|_\delta \leq \zeta_\delta(L)$) and $|p^k(z)|_\delta \leq \xi_\delta(k)$; (iv) There exist γ, γ_1, $\gamma_2 > 0$, and $\beta_{\mathbf{L}}$, a_L, λ^1_k, and $\lambda^2_{l,k}$ such that $|\Pi_0(z) - \lambda^{1\prime}_k p^k(z)|_\delta \leq Ck^{-\gamma_1}$, $|\bar{\varphi}_{0l}(z) - \lambda^{2\prime}_{l,k} p^k(z)|_\delta \leq Ck^{-\gamma_2}$ for all l, $|h_0(z, v) - a'_L \tilde{\varphi}^L(z, v)|_\delta \leq CL^{-\gamma}$, and $|g_0(z, v) - \beta'_{\mathbf{L}}\psi^{\mathbf{L}}(z, v)|_\delta \leq C\mathbf{L}^{-\gamma}$; (v) both $\mathcal{Z}$ and $\mathcal{X}$ are compact.*

Let $\triangle_{n,1} = k_n^{1/2}/\sqrt{n} + k_n^{-\alpha_1}$ and $\triangle_{n,2} = k_n^{1/2}/\sqrt{n} + k_n^{-\alpha_2}$ and $\triangle_n = \max\{\triangle_{n,1}, \triangle_{n,2}\}$.

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

$$\left(\sum\nolimits_{i=1}^{n} (\hat{g}(z_i, v_i) - g_0(z_i, v_i))^2 /n\right)^{1/2} = O_p(\sqrt{\mathbf{L}/n} + L\xi_0(k)\triangle_{n,1}\sqrt{k/n} + L\triangle_{n,2} + \mathbf{L}^{-\gamma}).$$

B.1 Proof of Lemma L1

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

First note $(P'P)/n$ becomes nonsingular w.p.a.1 as $\xi_0(k)^2 k/n \to 0$ by Assumption L1 (ii) and 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

$$\sum\nolimits_{i=1}^{n} ||\hat{\Pi}_i - \Pi_i||^2/n = O_p(\triangle^2_{n,1}) \text{ and } \sum\nolimits_{i=1}^{n} ||\dot{\bar{\varphi}}_{l,i} - \bar{\varphi}_{l,i}||^2/n = O_p(\triangle^2_{n,2}) \text{ for all } l. \quad (18)$$

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

$$\begin{aligned} \max_{i \leq n} ||\hat{\Pi}_i - \Pi_i|| &= O_p(\xi_0(k)\triangle_{n,1}) \quad (19) \\ \max_{i \leq n} ||\dot{\bar{\varphi}}_{l,i} - \bar{\varphi}_{l,i}|| &= O_p(\xi_0(k)\triangle_{n,2}) \text{ for all } l. \end{aligned}$$

Define $\hat{\mathcal{T}} = (\hat{\hat{\psi}}^{\mathbf{L},n})'\hat{\hat{\psi}}^{\mathbf{L},n}/n$ and $\dot{\mathcal{T}} = (\psi^{\mathbf{L},n})'\psi^{\mathbf{L},n}/n$. Our goal is to show that $\hat{\mathcal{T}}$ is nonsingular w.p.a.1. We first show that $\dot{\mathcal{T}}$ is nonsingular w.p.a.1 and this is closely related with the identification result of Theorem 1. Recall that (x_i, z_{1i}) and $\kappa(z_i, v_i)$ has no additive functional relationship for any $\kappa(z_i, v_i)$ satisfying $E[\kappa(Z_i, V_i)|Z_i] = 0$ and $E[\tilde{\varphi}_i^L\tilde{\varphi}_i^{L\prime}]$ is nonsingular by Assumption L1 (ii). Therefore, $\dot{\mathcal{T}}$ is nonsingular w.p.a.1 by Assumption L1 (ii) as $\zeta_0(\mathbf{L})^2\mathbf{L}/n \to 0$ by the same proof in Lemma A1 of Newey, Powell, and Vella (1999). The same conclusion holds even when instead we take $\dot{\mathcal{T}} = \sum_{i=1}^n C(z_i, v_i)\psi_i^{\mathbf{L}}\psi_i^{\mathbf{L}\prime}/n$ for some positive bounded function $C(z_i, v_i)$ by the same proof in Lemma A1 of Newey, Powell, and Vella (1999) and this helps to derive the consistency of the heteroskedasticity robust variance estimator later.

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

Next note that

$$\begin{aligned} \left\|\hat{\tilde{\varphi}}_i^L - \tilde{\varphi}_i^L\right\| &\leq \left\|\varphi^L(z_i, \hat{v}_i) - \varphi^L(z_i, v_i)\right\| + \left\|\hat{\bar{\varphi}}^L(z_i) - \bar{\varphi}^L(z_i)\right\| \\ &\leq \left\|\varphi^L(z_i, \hat{v}_i) - \varphi^L(z_i, v_i)\right\| + \left\|\hat{\bar{\varphi}}^L(z_i) - \dot{\bar{\varphi}}^L(z_i)\right\| + \left\|\dot{\bar{\varphi}}^L(z_i) - \bar{\varphi}^L(z_i)\right\|. \end{aligned} \tag{20}$$

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 Π_i for all l (Assumption C1 (vi)). Combined with (18), it implies that

$$\sum\nolimits_{i=1}^n \left\|\varphi^L(z_i, \hat{v}_i) - \varphi^L(z_i, v_i)\right\|^2/n = O_p(\zeta_1(L)^2\triangle_{n,1}^2). \tag{21}$$

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

$$\begin{aligned} \sum\nolimits_{i=1}^n \left\|\hat{\bar{\varphi}}_l(z_i) - \dot{\bar{\varphi}}_l(z_i)\right\|^2/n &= \mathrm{tr}\left\{\sum\nolimits_{i=1}^n p^k(z_i)'(P'P)^-P'\hat{\omega}_l\hat{\omega}_l'P(P'P)^-p^k(z_i)\right\}/n \\ &= \mathrm{tr}\left\{(P'P)^-P'\hat{\omega}_l\hat{\omega}_l'P(P'P)^-\sum_{i=1}^n p^k(z_i)p^k(z_i)'\right\}/n \\ &= \mathrm{tr}\left\{(P'P)^-P'\hat{\omega}_l\hat{\omega}_l'P\right\}/n \\ &\leq C\max_{i\leq n}||\hat{\Pi}_i - \Pi_i||^2\mathrm{tr}\left\{(P'P)^-P'P\right\}/n \leq C\xi_0(k)^2\triangle_{n,1}^2k/n \end{aligned} \tag{22}$$

where the first inequality is obtained by (19) and applying a mean value expansion to $\varphi_l(z_i, v_i)$ which is Lipschitz in Π_i for all l (Assumption C1 (vi)). From (18), (20), (21), and (22), we conclude

$$\sum\nolimits_{i=1}^n ||\hat{\bar{\varphi}}^L(z_i) - \bar{\varphi}^L(z_i)||^2/n = O_p(L\xi_0(k)^2\triangle_{n,1}^2k/n) + O_p(L\triangle_{n,2}^2) = o_p(1) \tag{23}$$

and

$$\sum\nolimits_{i=1}^{n}\left\|\hat{\tilde{\varphi}}_i^L-\tilde{\varphi}_i^L\right\|^2/n=O_p(\zeta_1(L)^2\triangle_{n,1}^2)+O_p(L\xi_0(k)^2\triangle_{n,1}^2k/n)+O_p(L\triangle_{n,2}^2)=o_p(1).$$

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

$$\sum\nolimits_{i=1}^{n}||\hat{\tilde{\varphi}}_i^L||^2/n\leq 2\sum_{i=1}^{n}||\hat{\tilde{\varphi}}_i^L-\tilde{\varphi}_i^L||^2/n+2\sum_{i=1}^{n}||\tilde{\varphi}_i^L||^2/n=o_p(1)+O_p(L). \tag{24}$$

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

$$\sum\nolimits_{i=1}^{n}\left\|\hat{\tilde{\psi}}_i^{\mathbf{L}}-\psi_i^{\mathbf{L}}\right\|^2/n\leq\sum\nolimits_{i=1}^{n}\left\|\hat{\tilde{\varphi}}_i^L-\tilde{\varphi}_i^L\right\|^2/n=O_p((\triangle_n^{\varphi})^2)=o_p(1). \tag{25}$$

This also implies

$$\sum\nolimits_{i=1}^{n}\left\|\hat{\tilde{\psi}}_i^{\mathbf{L}}\right\|^2/n=O_p(\mathbf{L})$$

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

Then applying (25) and applying the triangle inequality and Cauchy-Schwarz inequality and by Assumption L1 (iii) , we obtain

$$\begin{aligned}||\hat{\mathcal{T}}-\dot{\mathcal{T}}|| &\leq \sum\nolimits_{i=1}^{n}\left\|\hat{\tilde{\psi}}_i^{\mathbf{L}}-\psi_i^{\mathbf{L}}\right\|^2/n+2\sum\nolimits_{i=1}^{n}\left\|\psi_i^{\mathbf{L}}\right\|\left\|\hat{\tilde{\psi}}_i^{\mathbf{L}}-\psi_i^{\mathbf{L}}\right\|/n \\ &\leq O_p((\triangle_n^{\varphi})^2)+2\left(\sum\nolimits_{i=1}^{n}\left\|\psi_i^{\mathbf{L}}\right\|^2/n\right)^{1/2}\left(\sum\nolimits_{i=1}^{n}\left\|\hat{\tilde{\psi}}_i^{\mathbf{L}}-\psi_i^{\mathbf{L}}\right\|^2/n\right)^{1/2} \\ &= O_p((\triangle_n^{\varphi})^2)+O_p(\mathbf{L}^{1/2}\triangle_n^{\varphi})=o_p(1).\end{aligned} \tag{26}$$

It follows that

$$\begin{aligned}||\hat{\mathcal{T}}-\mathcal{T}|| &\leq ||\hat{\mathcal{T}}-\dot{\mathcal{T}}||+||\dot{\mathcal{T}}-\mathcal{T}|| \\ &= O_p((\triangle_n^{\varphi})^2+\mathbf{L}^{1/2}\triangle_n^{\varphi}+\zeta_0(\mathbf{L})\sqrt{\mathbf{L}/n})\equiv O_p(\triangle_{\mathcal{T}})=o_p(1)\end{aligned} \tag{27}$$

where we obtain $||\dot{\mathcal{T}}-\mathcal{T}||=O_p(\zeta_0(\mathbf{L})\sqrt{\mathbf{L}/n})$ by the same proof in Lemma A1 of Newey, Powell, and Vella (1999).

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

Let $\eta_i=y_i-g_0(z_i,v_i)$ and let $\eta=(\eta_1,\ldots,\eta_n)'$. Let $(\mathbf{Z},\mathbf{V})=((Z_1,V_1),\ldots,(Z_n,V_n))$.

Then we have $E[\eta_i|\mathbf{Z},\mathbf{V}]=0$ and by the independence assumption of the observations, we have $E[\eta_i\eta_j|\mathbf{Z},\mathbf{V}]=0$ for $i\neq j$. We also have $E[\eta_i^2|\mathbf{Z},\mathbf{V}]<\infty$. Then by (25) and the triangle inequality, we bound

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

Then from the standard result (see Newey (1997) or Newey, Powell, and Vella (1999)) that the bound of a term in the conditional mean implies the bound of the term itself, we obtain $||(\hat{\hat{\psi}}^{\mathbf{L},n}-\psi^{\mathbf{L},n})'\eta/n||^2=o_p(n^{-1})$. Also note that $E[\left\|(\psi^{\mathbf{L},n})'\eta/n\right\|^2]=C\mathbf{L}/n$ (see proof of Lemma A1 in Newey, Powell, and Vella (1999)). Therefore, by the triangle inequality

$$\begin{aligned}
||(\hat{\hat{\psi}}^{\mathbf{L},n})'\eta/n||^2 &\leq 2||(\hat{\hat{\psi}}^{\mathbf{L},n}-\psi^{\mathbf{L},n})'\eta/n||^2+2||(\psi^{\mathbf{L},n})'\eta/n||^2. \\
&= o_p(1)+O_p(\mathbf{L}/n)=O_p(\mathbf{L}/n).
\end{aligned} \tag{28}$$

Define

$$\hat{g}_i = \hat{f}(x_i,z_{1i})+\hat{h}(z_i,\hat{v}_i),\hat{\hat{g}}_{\mathbf{L}i}=f_K(x_i,z_{1i})+\hat{h}_L(z_i,\hat{v}_i),\tilde{g}_{\mathbf{L}i}=f_K(x_i,z_{1i})+h_L(z_i,\hat{v}_i),$$

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

$$\begin{aligned}
0 &= \hat{\hat{\psi}}^{\mathbf{L},n\prime}(y-\hat{g})/n \\
&= \hat{\hat{\psi}}^{\mathbf{L},n\prime}(\eta-(\hat{g}-\hat{\hat{g}}_{\mathbf{L}})-(\hat{\hat{g}}_{\mathbf{L}}-\tilde{g}_{\mathbf{L}})-(\tilde{g}_{\mathbf{L}}-\tilde{g}_0))/n \\
&= \hat{\hat{\psi}}^{\mathbf{L},n\prime}(\eta-\hat{\hat{\psi}}^{\mathbf{L},n}(\hat{\beta}-\beta_{\mathbf{L}})-(\hat{\hat{g}}_{\mathbf{L}}-\tilde{g}_{\mathbf{L}})-(\tilde{g}_{\mathbf{L}}-\tilde{g}_0)-(\tilde{g}_0-g_0))/n.
\end{aligned} \tag{29}$$

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

$$\begin{aligned}
||\hat{\mathcal{T}}^{-1}\hat{\hat{\psi}}^{\mathbf{L},n\prime}(\tilde{g}_{\mathbf{L}}-\tilde{g}_0)/n|| &\leq O_p(1)\{(\tilde{g}_{\mathbf{L}}-\tilde{g}_0)'\hat{\hat{\psi}}^{\mathbf{L},n}(\hat{\hat{\psi}}^{\mathbf{L},n\prime}\hat{\hat{\psi}}^{\mathbf{L},n})^{-1}\hat{\hat{\psi}}^{\mathbf{L},n\prime}(\tilde{g}_{\mathbf{L}}-\tilde{g}_0)/n\}^{1/2} \\
&\leq O_p(1)\{(\tilde{g}_{\mathbf{L}}-\tilde{g}_0)'(\tilde{g}_{\mathbf{L}}-\tilde{g}_0)/n\}^{1/2}=O_p(\mathbf{L}^{-\gamma}).
\end{aligned} \tag{30}$$

Similarly we obtain by $\hat{\hat{\psi}}^{\mathbf{L},n}(\hat{\hat{\psi}}^{\mathbf{L},n\prime}\hat{\hat{\psi}}^{\mathbf{L},n})^{-1}\hat{\hat{\psi}}^{\mathbf{L},n\prime}$ idempotent, Assumption L1 (iv), and (23),

$$
\begin{aligned}
&||\hat{\mathcal{T}}^{-1}\hat{\hat{\psi}}^{\mathbf{L},n\prime}(\hat{\hat{g}}_{\mathbf{L}}-\tilde{g}_{\mathbf{L}})/n|| = O_p(1)\{(\hat{\hat{g}}_{\mathbf{L}}-\tilde{g}_{\mathbf{L}})'(\hat{\hat{g}}_{\mathbf{L}}-\tilde{g}_{\mathbf{L}})/n\}^{1/2} \qquad (31)\\
\leq\ & O_p(1)(\sum_{i=1}^{n}||\hat{h}_L(z_i,\hat{v}_i)-\tilde{h}_L(z_i,\hat{v}_i)||^2/n)^{1/2}\\
\leq\ & O_p(1)(\sum_{i=1}^{n}||a_L||^2||\hat{\bar{\varphi}}^L(z_i)-\bar{\varphi}^L(z_i)||^2/n)^{1/2} = O_p(L\xi_0(k)\triangle_{n,1}\sqrt{k/n}+L\triangle_{n,2}).
\end{aligned}
$$

Similarly also by $\hat{\hat{\psi}}^{\mathbf{L},n}(\hat{\hat{\psi}}^{\mathbf{L},n\prime}\hat{\hat{\psi}}^{\mathbf{L},n})^{-1}\hat{\hat{\psi}}^{\mathbf{L},n\prime}$ idempotent and (18) and applying the mean value expansion to $h_0(z_i,v_i)$, we have

$$
\begin{aligned}
&||\hat{\mathcal{T}}^{-1}\hat{\hat{\psi}}^{\mathbf{L},n\prime}(\tilde{g}_0-g_0)/n|| = O_p(1)(\sum_{i=1}^{n}||h_0(z_i,\hat{v}_i)-h_0(z_i,v_i)||^2/n)^{1/2} \qquad (32)\\
\leq\ & O_p(1)(\sum\nolimits_{i=1}^{n}||\hat{\Pi}_i-\Pi_i||^2/n)^{1/2} = O_p(\triangle_{n,1}) = o_p(1).
\end{aligned}
$$

Combining (28), (29), (30), (31), (32) and by $\hat{\mathcal{T}}$ is nonsingular w.p.a.1, we obtain

$$
\begin{aligned}
||\hat{\beta}-\beta_{\mathbf{L}}|| &\leq ||\hat{\mathcal{T}}^{-1}\hat{\hat{\psi}}^{\mathbf{L},n\prime}\eta/n||+||\hat{\mathcal{T}}^{-1}\hat{\hat{\psi}}^{\mathbf{L},n\prime}(\hat{\hat{g}}_{\mathbf{L}}-\tilde{g}_{\mathbf{L}})/n||+||\hat{\mathcal{T}}^{-1}\hat{\hat{\psi}}^{\mathbf{L},n\prime}(\tilde{g}_{\mathbf{L}}-\tilde{g}_0)/n||+o_p(1)\\
&= O_p(1)\{\sqrt{\mathbf{L}/n}+L\xi_0(k)\triangle_{n,1}\sqrt{k/n}+L\triangle_{n,2}+\mathbf{L}^{-\gamma}\}\equiv O_p(\triangle_{n,\beta}). \qquad (33)
\end{aligned}
$$

Define $g^*_{\mathbf{L}i}=f_K(x_i,z_{1i})+h^*_L(z_i,v_i)$ where $h^*_L(z_i,v_i)=a'_L(\varphi^L(z_i,v_i)-\hat{\bar{\varphi}}^L(z_i))$. Then applying the triangle inequality, by (23), (33), the Markov inequality, Assumption L1 (iv), and $\hat{\mathcal{T}}$ is nonsingular w.p.a.1 (by Assumption L1 (ii) and (27)), we conclude

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

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

$$\max_{i\leq n}|\hat{g}_i-g_{0i}|=O_p(\zeta_0(\mathbf{L})\triangle_{n,\beta}). \qquad (34)$$

B.2 Proof of Theorem 2

Under Assumptions C1, all the assumptions in Assumption L1 are satisfied. For the consistency, we require the following rate conditions: R(i) $\mathbf{L}^{1/2}\triangle_n^{\varphi} \to 0$ from (26), R(ii) $\zeta_0(\mathbf{L})^2\mathbf{L}/n \to 0$ (such that $\dot{\mathcal{T}}$ is nonsingular w.p.a.1), and R(iii) $\xi_0(k)^2 k/n \to 0$ (such that $P'P/n$ is nonsingular w.p.a.1). The other rate conditions are dominated by these three. From the definition of $\triangle_n^{\varphi} = (\zeta_1(L) + L^{1/2}\xi_0(k)\sqrt{k/n} + L^{1/2})\triangle_n$, we have R(i) : $\mathbf{L}^{1/2}(\zeta_1(L) + L^{1/2}\xi_0(k)\sqrt{k/n} + L^{1/2})\triangle_n$.

For the polynomial approximations, we have $\zeta_\delta(L) \leq CL^{1+2\delta}$ and $\xi_0(k) \leq Ck$ and for the spline approximations, we have $\zeta_\delta(L) \leq CL^{0.5+\delta}$ and $\xi_0(k) \leq Ck^{0.5}$. Therefore for the polynomial approximations, the rate condition becomes (i) $\mathbf{L}^{1/2}(L^3 + L^{1/2}k^{3/2}/\sqrt{n} + L^{1/2})\triangle_n \to 0$, (ii) $\mathbf{L}^3/n \to 0$, and (iii) $k^3/n \to 0$ and for the spline approximations, it becomes R(i) $\mathbf{L}^{1/2}(L^{3/2} + L^{1/2}k/\sqrt{n} + L^{1/2})\triangle_n \to 0$, (ii) $\mathbf{L}^2/n \to 0$, and (iii) $k^2/n \to 0$. Also note that

$$\begin{aligned}\triangle_{n,\beta} &\equiv \sqrt{\mathbf{L}/n} + L\xi_0(k)\triangle_{n,1}\sqrt{k/n} + L\triangle_{n,2} + \mathbf{L}^{-\gamma} \\ &= \sqrt{\mathbf{L}/n} + L\triangle_n + \mathbf{L}^{-\gamma}\end{aligned}$$

since $\xi_0(k)\sqrt{k/n} = o(1)$. We take $\gamma = s/d$ because f_0 and h_0 belong to the Hölder class and we can apply the approximation theorems (e.g., see Timan (1963), Schumaker (1981), Newey (1997), and Chen (2007)).

Therefore, the conclusion of Theorem C1 follows from Lemma L1 applying the dominated convergence theorem by $\hat{g}_i$ and g_{0i} are bounded.

C Proof of asymptotic normality

Along the proof, we will obtain a series of convergence rate conditions. We collect them here. First define

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

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

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

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

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

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

Let $p_i^k = p^k(Z_i)$. We start with introducing additional notation:

$$\begin{aligned}
\Sigma &= E[\psi_i^{\mathbf{L}}\psi_i^{\mathbf{L}\prime}\mathrm{var}(Y_i|Z_i,V_i)], \mathcal{T} = E[\psi_i^{\mathbf{L}}\psi_i^{\mathbf{L}\prime}], \mathcal{T}_1 = E[p_i^k p_i^{k\prime}], \qquad (35)\\
\Sigma_1 &= E[V_i^2 p_i^k p_i^{k\prime}], \Sigma_{2,l} = E[(\varphi_l(Z_i,V_i) - \bar{\varphi}_l(Z_i))^2 p_i^k p_i^{k\prime}],\\
H_{11} &= E[\frac{\partial h_{0i}}{\partial V_i}\psi_i^{\mathbf{L}}p_i^{k\prime}], \bar{H}_{11} = \sum_{i=1}^n \frac{\partial h_{0i}}{\partial V_i}\psi_i^{\mathbf{L}}p_i^{k\prime}/n\\
H_{12} &= E[E[\frac{\partial h_{0i}}{\partial V_i}|Z_i]\psi_i^{\mathbf{L}}p_i^{k\prime}], \bar{H}_{12} = \sum_{i=1}^n E[\frac{\partial h_{0i}}{\partial V_i}|Z_i]\psi_i^{\mathbf{L}}p_i^{k\prime}/n\\
H_{2,l} &= E[a_l\psi_i^{\mathbf{L}}p_i^{k\prime}], \bar{H}_{2,l} = \sum_{i=1}^n a_l\psi_i^{\mathbf{L}}p_i^{k\prime}/n, H_1 = H_{11} - H_{12}, \bar{H}_1 = \bar{H}_{11} - \bar{H}_{12}\\
\bar{\Omega} &= \mathcal{A}\mathcal{T}^{-1}[\Sigma + H_1\mathcal{T}_1^{-1}\Sigma_1\mathcal{T}_1^{-1}H_1' + \textstyle\sum_{l=1}^L H_{2,l}\mathcal{T}_1^{-1}\Sigma_{2,l}\mathcal{T}_1^{-1}H_{2,l}']\mathcal{T}^{-1}\mathcal{A}'.
\end{aligned}$$

We let $\mathcal{T}_1 = E[p_i^k p_i^{k\prime}] = I$ and $E[\tilde{\varphi}_i{}^L\tilde{\varphi}_i{}^{L\prime}] = I$ without loss of generality.

Then $\bar{\Omega} = \mathcal{A}\mathcal{T}^{-1}\left[\Sigma + H_1\Sigma_1H_1' + \sum_{l=1}^L H_{2,l}\Sigma_{2,l}H_{2,l}'\right]\mathcal{T}^{-1}\mathcal{A}'$. Let Γ be a symmetric square root of $\bar{\Omega}$. Because $\mathcal{T}$ is nonsingular and $\mathrm{var}(Y_i|Z_i,V_i)$ is bounded away from zero, $\Sigma - CI$ is positive semidefinite for some positive constant C. It follows that

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

Next we show $\bar{\Omega} \to \Omega$. Under Assumption R1, we have $\mathcal{A} = E[\nu^*(Z,V)\psi_i^{\mathbf{L}\prime}]$. Take $\nu_{\mathbf{L}}^*(Z,V) = \mathcal{A}\mathcal{T}^{-1}\psi_i^{\mathbf{L}}$. Then note $E[||\nu^*(Z,V) - \nu_{\mathbf{L}}^*(Z,V)||^2] \to 0$ because (i) $\nu_{\mathbf{L}}^*(Z,V) =$

$E[\nu^*(Z,V)\psi_i^{\mathbf{L}\prime}]\mathcal{T}^{-1}\psi_i^{\mathbf{L}}$ is a mean-squared projection of $\nu^*(z_i,v_i)$ on $\psi_i^{\mathbf{L}}$; (ii) $\nu^*(z_i,v_i)$ is smooth and the second moment of $\nu^*(z_i,v_i)$ is bounded, so it is well-approximated in the mean-squared error as assumed in Assumption R1. Let $\nu_i^* = \nu^*(Z_i,V_i)$ and $\nu_{\mathbf{L}i}^* = \nu_{\mathbf{L}}^*(Z_i,V_i)$. It follows that

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

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

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

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

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

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

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

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

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 (17) to the corresponding terms in (35). We show some of them first, which will be useful to derive the asymptotic normality. Note $||\hat{\mathcal{T}}-\mathcal{T}|| = O_p(\triangle_{\mathcal{T}}) = o_p(1)$ and $||\hat{\mathcal{T}}_1-\mathcal{T}_1|| = O_p(\triangle_{\mathcal{T}_1}) = o_p(1)$. We also have $||\Gamma\mathcal{A}(\hat{\mathcal{T}}^{-1}-\mathcal{T}^{-1})|| = o_p(1)$ and $||\Gamma\mathcal{A}\hat{\mathcal{T}}^{-1/2}||^2 = O_p(1)$ (see proof in Lemma A1 of Newey, Powell, and Vella (1999)). We next show $||\bar{H}_{11}-H_{11}||=o_p(1)$. Let $H_{11\mathbf{L}}=E[\sum_{l=1}^{L}a_l\frac{\partial\varphi_l(Z_i,V_i)}{\partial V_i}\psi_i^{\mathbf{L}}p_i^{k\prime}]$ and $\bar{H}_{11\mathbf{L}}=\sum_{i=1}^{n}\sum_{l=1}^{L}a_l\frac{\partial\varphi_l(Z_i,V_i)}{\partial V_i}\psi_i^{\mathbf{L}}p_i^{k\prime}/n$. Similarly define $H_{12\mathbf{L}}$ and $\bar{H}_{12\mathbf{L}}$ and let $H_{1\mathbf{L}}=H_{11\mathbf{L}}-H_{12\mathbf{L}}$. By Assumption N1 (i), Assumption L1 (iii), and the Cauchy-Schwarz inequality,

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

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

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

where the first equality holds because $\frac{\partial h_{Li}}{\partial V_i}=\sum_{l=1}^{L}a_l\frac{\partial\tilde{\varphi}_l(Z_i,V_i)}{\partial V_i}=\sum_{l=1}^{L}a_l\frac{\partial\varphi_l(Z_i,V_i)}{\partial V_i}$ and the last result holds because $h_{Li}\in\mathcal{H}_n$ (i.e. $|h_{Li}|_1$ is bounded). Similarly by (25), the Cauchy-Schwarz inequality, and the Markov inequality, we obtain

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

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

Now we derive the asymptotic expansion to obtain the influence functions. Further define $\hat{g}_{\mathbf{L}i}=f_K(x_i,z_{1i})+\tilde{h}_L(z_i,\hat{v}_i)$ where $\tilde{h}_L(z_i,\hat{v}_i)=a_L'(\varphi^L(z_i,\hat{v}_i)-E[\varphi^L(Z_i,\hat{V}_i)|z_i])$ and $g_{\mathbf{L}i}=f_K(x_i,z_{1i})+h_L(z_i,v_i)$. From the first order condition, we obtain the expansion similar to (29). Recall $\beta_{\mathbf{L}}=(\beta_1,\ldots,\beta_K,a_L')'$ and let this $\beta_{\mathbf{L}}$ satisfy Assumption N1 (i).

$$
\begin{aligned}
0 &= \hat{\hat{\psi}}^{\mathbf{L},n\prime}(y-\hat{g})/\sqrt{n} \qquad (36)\\
&= \hat{\hat{\psi}}^{\mathbf{L},n\prime}(\eta-(\hat{g}-\hat{\hat{g}}_{\mathbf{L}})-(\hat{\hat{g}}_{\mathbf{L}}-\hat{g}_{\mathbf{L}})-(\hat{g}_{\mathbf{L}}-g_{\mathbf{L}})-(g_{\mathbf{L}}-g_0))/\sqrt{n} \\
&= \hat{\hat{\psi}}^{\mathbf{L},n\prime}(\eta-\hat{\hat{\psi}}^{\mathbf{L},n}(\hat{\beta}-\beta_{\mathbf{L}})-(\hat{\hat{g}}_{\mathbf{L}}-\hat{g}_{\mathbf{L}})-(\hat{g}_{\mathbf{L}}-g_{\mathbf{L}})-(g_{\mathbf{L}}-g_0))/\sqrt{n}.
\end{aligned}
$$

Similar to (30), we obtain

$$||\hat{\mathcal{T}}^{-1}\hat{\hat{\psi}}^{\mathbf{L},n\prime}(g_{\mathbf{L}}-g_0)/\sqrt{n}|| = O_p(\sqrt{n}\mathbf{L}^{-\gamma}). \tag{37}$$

Also note that

$$\begin{aligned}\sqrt{n}||\Gamma(\alpha(g_{\mathbf{L}})-\alpha(g_0))|| &= \sqrt{n}||\Gamma||\cdot||\alpha(g_{\mathbf{L}}-g_0)|| \leq C\sqrt{n}\,\|\Gamma\|\cdot|\psi^{\mathbf{L}\prime}(\cdot)\beta_{\mathbf{L}}-g_0(\cdot)|_\delta \\ &= O_p(\sqrt{n}\mathbf{L}^{-\gamma}) = o_p(1)\end{aligned} \tag{38}$$

because $\alpha(\cdot)$ is a linear functional and by Assumption N1 (i).

From the linearity of $\alpha(\cdot)$, (36), (37), and (38) we have

$$\begin{aligned}\sqrt{n}\Gamma(\hat{\theta}-\theta_0) &= \sqrt{n}\Gamma(\alpha(\hat{g})-\alpha(g_0)) = \sqrt{n}\Gamma(\alpha(\hat{g})-\alpha(g_{\mathbf{L}}))+\sqrt{n}\Gamma(\alpha(g_{\mathbf{L}})-\alpha(g_0)) \\ &= \sqrt{n}\Gamma\mathcal{A}(\hat{\beta}-\beta_{\mathbf{L}})+\sqrt{n}\Gamma\{a(g_{\mathbf{L}})-a(g_0)\} \\ &= \Gamma\mathcal{A}\hat{\mathcal{T}}^{-1}\hat{\hat{\psi}}^{\mathbf{L},n\prime}(\eta-(\hat{\hat{g}}_{\mathbf{L}}-\hat{g}_{\mathbf{L}})-(\hat{g}_{\mathbf{L}}-g_{\mathbf{L}}))/\sqrt{n}+o_p(1).\end{aligned} \tag{39}$$

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

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

and the remainder term $||\hat{\varsigma}|| \leq C\sqrt{n}||\Gamma\mathcal{A}\hat{\mathcal{T}}^{-1/2}||\zeta_0(L)\sum_{i=1}^n C_i||\hat{\Pi}_i-\Pi_i||^2/n = O_p(\sqrt{n}\zeta_0(L)\triangle_{n,1}^2) = o_p(1)$. Then by the essentially same proofs ((A.18) to (A.23)) in Lemma A2 of Newey, Powell, and Vella (1999), under $\sqrt{n}k^{-s_1/d_z}\to 0$ and $k^{1/2}(\triangle_{\mathcal{T}_1}+\triangle_H)+\mathbf{L}^{1/2}\triangle_{\mathcal{T}}\to 0$ (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 obtain

$$\Gamma\mathcal{A}\hat{\mathcal{T}}^{-1}\hat{\hat{\psi}}^{\mathbf{L},n\prime}(\hat{g}_{\mathbf{L}}-g_{\mathbf{L}})/\sqrt{n} = \Gamma\mathcal{A}\mathcal{T}^{-1}H_1\sum\nolimits_{i=1}^n p_i^k v_i/\sqrt{n}+o_p(1). \tag{40}$$

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

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

$$\begin{aligned}
&\Gamma\mathcal{A}\hat{\mathcal{T}}^{-1}\sum_{i=1}^{n}\hat{\hat{\psi}}_i^{\mathbf{L}}(\hat{\hat{g}}_{\mathbf{L}i}-\hat{g}_{\mathbf{L}i})/\sqrt{n}=\Gamma\mathcal{A}\hat{\mathcal{T}}^{-1}\sum_{i=1}^{n}\hat{\hat{\psi}}_i^{\mathbf{L}}a_L'(\hat{\varphi}^L(z_i)-E[\varphi^L(Z_i,\hat{V}_i)|z_i])/\sqrt{n}\\
=\;&\Gamma\mathcal{A}\hat{\mathcal{T}}^{-1}\{\sum\nolimits_l\bar{H}_{2,l}\hat{\mathcal{T}}_1^{-1}\sum\nolimits_{i=1}^{n}p_i^k\tilde{\varphi}_{li}/\sqrt{n}+\sum\nolimits_l\bar{H}_{2,l}\hat{\mathcal{T}}_1^{-1}\sum\nolimits_{i=1}^{n}p_i^k(\bar{\varphi}_l(z_i)-p_i^{k\prime}\lambda_{l,k}^2)/\sqrt{n}\}\\
&+\Gamma\mathcal{A}\hat{\mathcal{T}}^{-1}\sum\nolimits_{i=1}^{n}\hat{\hat{\psi}}_i^{\mathbf{L}}\sum\nolimits_l a_l(p_i^{k\prime}\lambda_{l,k}^2-\bar{\varphi}_l(z_i))/\sqrt{n}+\Gamma\mathcal{A}\hat{\mathcal{T}}^{-1}\sum\nolimits_{i=1}^{n}\hat{\hat{\psi}}_i^{\mathbf{L}}\rho_i/\sqrt{n} \qquad (41)
\end{aligned}$$

where $\rho_i=p_i^{k\prime}\hat{\mathcal{T}}_1^{-1}\sum_{i=1}^n p_i^k\sum_l a_l\{(\varphi_l(z_i,\hat{v}_i)-\varphi_l(z_i,v_i))-(E[\varphi_l(Z_i,\hat{V}_i)|z_i]-\bar{\varphi}_l(z_i))\}$. We focus on the last term in (41). Note that $p_i^{k\prime}\hat{\mathcal{T}}_1^{-1}\sum_{i=1}^n p_i^k(\varphi_l(z_i,\hat{v}_i)-\varphi_l(z_i,v_i))$ is a projection of $\varphi_l(z_i,\hat{v}_i)-\varphi_l(z_i,v_i)$ on p_i^k and it converges to the conditional mean $E[\varphi_l(Z_i,\hat{V}_i)|z_i]-\bar{\varphi}_l(z_i)$. Note that $E[\rho_i|Z_1,\ldots,Z_n]=0$ and therefore $E[||\rho_i||^2|Z_1,\ldots,Z_n]\leq LO_p(\triangle_{n,2}^2)$ by a similar proof to (18). It follows that by Assumption L1 (iii) and the Cauchy-Schwarz inequality,

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

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

Then again by the essentially same proofs ((A.18) to (A.23)) in Lemma A2 of Newey, Powell, and Vella (1999), under $\sqrt{n}k^{-s_2/d_z}\to 0$, $\sqrt{n}k^{1/2}L^{-s/d}\to 0$, and $k^{1/2}(\triangle_{\mathcal{T}_1}+\triangle_H)+\mathbf{L}^{1/2}\triangle_{\mathcal{T}}+\triangle_{d\varphi}\to 0$ (so that we can replace $\hat{\mathcal{T}}_1$ with $\mathcal{T}_1$, $\bar{H}_{2,l}$ with $H_{2,l}$, and $\hat{\mathcal{T}}$ with $\mathcal{T}$ respectively and we can ignore the last term in (41)), we obtain

$$\Gamma\mathcal{A}\hat{\mathcal{T}}^{-1}\hat{\hat{\psi}}^{\mathbf{L},n\prime}(\hat{\hat{g}}_{\mathbf{L}}-\hat{g}_{\mathbf{L}})/\sqrt{n}=\Gamma\mathcal{A}\mathcal{T}^{-1}\sum\nolimits_l H_{2,l}\sum\nolimits_{i=1}^{n}p_i^k\tilde{\varphi}_{li}/\sqrt{n}+o_p(1). \qquad (42)$$

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

We can also show that replacing $\hat{\hat{\psi}}_i^{\mathbf{L}}$ with $\psi_i^{\mathbf{L}}$ does not influence the stochastic expansion by (25). Therefore by (39), (40), and (42), we obtain the stochastic expansion,

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

To apply the Lindeberg-Feller theorem, we check the Lindeberg condition. For any vector q with $||q||=1$, let $W_{in}=q'\Gamma\mathcal{A}\mathcal{T}^{-1}(\psi_i^{\mathbf{L}}\eta_i-H_1p_i^kv_i-\sum_l H_{2,l}p_i^k\tilde{\varphi}_{li})/\sqrt{n}$. Note that W_{in} is i.i.d, given n and by construction, $E[W_{in}]=0$ and $\mathrm{var}(W_{in})=1/n$. Also note that $||\Gamma\mathcal{A}\mathcal{T}^{-1}||\leq C$, $||\Gamma\mathcal{A}\mathcal{T}^{-1}H_j||\leq C||\Gamma\mathcal{A}\mathcal{T}^{-1}||\leq C$ by $CI-H_jH_j'$ being positive semidefinite for $j=1,(2,1),\ldots,(2,L)$. Also note that $(\sum_{l=1}^L\tilde{\varphi}_{li})^4\leq L^2(\sum_{l=1}^L\tilde{\varphi}_{li}^2)^2\leq L^3\sum_{l=1}^L\tilde{\varphi}_{li}^4$. It follows that for any $\varepsilon > 0$,

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

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

Now we show the convergence of the each term in (17) to the corresponding terms in (35). 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 $\max_{i\le n}|\hat{g}_i - g_{0i}| = O_p(\zeta_0(\mathbf{L})\triangle_{n,\beta}) = o_p(1)$ by (34). Let $\hat{D} = \Gamma\mathcal{A}\hat{\mathcal{T}}^{-1}\hat{\hat{\psi}}^{\mathbf{L},n\prime}\mathrm{diag}\{1+|\eta_i|,\ldots,1+|\eta_n|\}\hat{\hat{\psi}}^{\mathbf{L},n}\hat{\mathcal{T}}^{-1}\mathcal{A}'\Gamma'$ and note that $\hat{\hat{\psi}}^{\mathbf{L},n}$ and $\hat{\mathcal{T}}$ only depend on $(Z_1,V_1),\ldots,(Z_n,V_n)$ and thus $E[\hat{D}|(Z_1,V_1),\ldots,(Z_n,V_n)] \le C\Gamma\mathcal{A}\hat{\mathcal{T}}^{-1}\mathcal{A}'\Gamma' = O_p(1)$. Therefore, $||\hat{D}|| = O_p(1)$ as well. Next let $\tilde{\Sigma} = \sum_{i=1}^n \hat{\hat{\psi}}_i^{\mathbf{L}}\hat{\hat{\psi}}_i^{\mathbf{L}\prime}\eta_i^2/n$. Then,

$$
\begin{aligned}
||\Gamma\mathcal{A}\hat{\mathcal{T}}^{-1}(\hat{\Sigma}-\tilde{\Sigma})\hat{\mathcal{T}}^{-1}\mathcal{A}'\Gamma'|| &= ||\Gamma\mathcal{A}\hat{\mathcal{T}}^{-1}\hat{\hat{\psi}}^{\mathbf{L},n\prime}\mathrm{diag}\{\hat{\eta}_1^*,\ldots,\hat{\eta}_n^*\}\hat{\hat{\psi}}^{\mathbf{L},n}\hat{\mathcal{T}}^{-1}\mathcal{A}'\Gamma'|| \\
&\le C\mathrm{tr}(\hat{D})\max_{i\le n}|\hat{g}_i - g_{0i}| = O_p(1)o_p(1).
\end{aligned} \tag{43}
$$

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

$$
\begin{aligned}
||\tilde{\Sigma}-\Sigma|| &= O_p(\Delta_{\mathcal{T}} + \zeta_0(\mathbf{L})^2\mathbf{L}/n) \equiv O_p(\triangle_\Sigma) = o_p(1), \quad (44)\\
||\Gamma\mathcal{A}\hat{\mathcal{T}}^{-1}(\hat{\Sigma}-\Sigma)\hat{\mathcal{T}}^{-1}\mathcal{A}'\Gamma'|| &= o_p(1), \\
||\Gamma\mathcal{A}(\hat{\mathcal{T}}^{-1}\Sigma\hat{\mathcal{T}}^{-1} - \mathcal{T}^{-1}\Sigma\mathcal{T}^{-1})\mathcal{A}'\Gamma'|| &= o_p(1).
\end{aligned}
$$

Then, by (43), (44), and the triangle ineq., we conclude $||\Gamma\mathcal{A}\hat{\mathcal{T}}^{-1}\hat{\Sigma}\hat{\mathcal{T}}^{-1}\mathcal{A}'\Gamma' - \Gamma\mathcal{A}\mathcal{T}^{-1}\Sigma\mathcal{T}^{-1}\mathcal{A}'\Gamma'|| = o_p(1)$. It remains to show that for $j = 1,(2,1),\ldots,(2,L)$,

$$
\Gamma\mathcal{A}(\hat{\mathcal{T}}^{-1}\hat{H}_j\hat{\mathcal{T}}_1^{-1}\hat{\Sigma}_j\hat{\mathcal{T}}_1^{-1}\hat{H}_j'\hat{\mathcal{T}}^{-1} - \mathcal{T}^{-1}H_j\Sigma_jH_j'\mathcal{T}^{-1})\mathcal{A}'\Gamma' = o_p(1). \tag{45}
$$

As we have shown $||\hat{\Sigma}-\Sigma|| = o_p(1)$, similarly we can show $||\hat{\Sigma}_j - \Sigma_j|| = o_p(1)$, $j = 1,(2,1),\ldots,(2,L)$.

We focus on showing $||\hat{H}_j - \bar{H}_j|| = o_p(1)$ for $j = 1,(2,1),\ldots,(2,L)$. First note that

$$
||\hat{H}_{11} - \bar{H}_{11}|| = ||\sum\nolimits_{i=1}^n(\sum\nolimits_{l=1}^L \hat{a}_l\frac{\partial\varphi_l(z_i,\hat{v}_i)}{\partial v_i} - a_l\frac{\partial\varphi_l(z_i,v_i)}{\partial v_i})\hat{\hat{\psi}}_i^{\mathbf{L}}p^k(z_i)'/n||
$$

By the Cauchy-Schwarz inequality, (24), and Assumption L1 (iii), we have $\sum_{i=1}^n ||\hat{\hat{\psi}}_i^{\mathbf{L}} p_i^{k\prime}||^2/n \leq \sum_{i=1}^n ||\hat{\hat{\psi}}_i^{\mathbf{L}}||^2 ||p_i^k||^2/n = O_p(\mathbf{L}\xi_0(k)^2)$. Also note that by the triangle inequality, the Cauchy-Schwarz inequality, and by Assumption C1 (vi) and (19), applying a mean value expansion to $\frac{\partial \varphi_l(z_i,v_i)}{\partial v_i}$ w.r.t v_i,

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

where $\tilde{v}_i$ lies between $\hat{v}_i$ and v_i, which may depend on l. We therefore conclude by the triangle inequality and the Cauchy-Schwarz inequality, $||\hat{H}_{11} - \bar{H}_{11}|| \leq O_p((\zeta_1(L)\triangle_{n,\beta} + \xi_0(k)\triangle_{n,1})\mathbf{L}^{1/2}\xi_0(k)) = O_p(\triangle_{\hat{H}}) = o_p(1)$. Similarly we can show that $||\hat{H}_{12} - \bar{H}_{12}|| = o_p(1)$ and $||\hat{H}_{2,l} - \bar{H}_{2,l}|| = o_p(1)$ $l = 1, \ldots, L$. We have shown that $||\bar{H}_j - H_j|| = o_p(1)$ for $j = 1, (2,1), \ldots, (2,L)$ previously. Therefore, $||\hat{H}_j - H_j|| = o_p(1)$ for $j = 1, (2,1), \ldots, (2,L)$. Then by the similar proof like (43) and (44), the conclusion (45) follows. From (45) 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

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

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

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

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

Gagliardini, P., and O. Scaillet (2009): "Tikhonov Regularisation for Functional Minimum Distance Estimation," Working paper, HEC Universite de Geneve.

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

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

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

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

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

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

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

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

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

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

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

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

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