Alternative Asymptotics and the Partially Linear Model with Many Regressors[1]

Matias D. Cattaneo, Department of Economics, University of Michigan, cattaneo@umich.edu.

Michael Jansson, Department of Economics, UC Berkeley, mjansson@econ.berkeley.edu.

Whitney K. Newey, Department of Economics, MIT, wnewey@mit.edu.

July, 2010

Revised November, 2010

JEL classification: C13, C31.

Keywords: partially linear model, many terms, adjusted variance.

[1]The authors thank comments from Alfonso Flores-Lagunes, seminar participants at Indiana, Michigan and MSU, and conference participants at the 2010 Joint Statistical Meetings and the 2010 LACEA Impact Evaluation Network Conference.

Proposed Running Head: Semiparametric Asymptotics

Corresponding Author:
Whitney K. Newey
Department of Economics
MIT, E52-262D
Cambridge, MA 02142-1347

Abstract

1 Introduction

Many instrument asymptotics, where the number of instruments grows as fast as the sample size, has proven useful for instrumental variable estimators. Kunitomo (1980) and Morimune (1983) derived asymptotic variances that are larger than the usual formulae when the number of instruments and sample size grow at the same rate, and Bekker (1994) and others provided consistent estimators of these larger variances. Hansen, Hausman, and Newey (2008) showed that using many instrument standard errors provides a theoretical improvement for a range of number of instruments and a practical improvement for estimating the returns to schooling. Thus, many instrument asymptotics and the associated standard errors have been demonstrated to be a useful alternative to the usual asymptotics for instrumental variables.

Instrumental variable estimators implicitly depend on a nonparametric series estimator. Many instrument asymptotics has the number of series terms growing so fast that the series estimator is not consistent. Analogous asymptotics for kernel density weighted average derivative estimators has been considered by Cattaneo, Crump, and Jansson (2010). They find that when the bandwidth shrinks faster than needed for consistency of the kernel estimator the variance of the estimator is larger than the usual formula. They also find that correcting the variance provides an improvement over standard asymptotics for a range of bandwidths.

The purpose of this paper is to show that these results share a common structure and that this structure can be used to derive new results. The common structure is that the object determining the limiting distribution is a V-statistic, with a remainder that is a asymptotically normal degenerate U-statistic. Asymptotic normality of the remainder distinguishes this setting from other ones with V-statistics. Here the asymptotically normal remainder comes from the number of series terms going to infinity (or bandwidth shrinking to zero), while the behavior of a degenerate U-statistic is more complicated in other settings. When the number of terms grows as fast as the sample size the remainder has the same magnitude as the leading term, resulting in an asymptotic variance larger than just the variance of the leading term. The many instrument and small bandwidth results share this structure. In keeping with this common structure, we will henceforth refer to such results under the general heading of alternative asymptotics.

Applying the common structure to a series estimator of the partially linear model leads to new results. These results allow the number of terms in the series approximation to grow as fast as the sample size. The asymptotic distribution of the estimator is derived and it is shown that it has a larger asymptotic variance than the usual formula. When the disturbance is homoskedastic this larger variance is consistently estimated by using by the usual homoskedasticity consistent estimator with the proper degrees of freedom. This provides a large sample justification for the use of a degrees of freedom correction without normality of disturbances. It is also found that the White (1980) variance estimator is inconsistent with many regressors, being too small when the disturbance is homoskedastic. We give a variance estimator that is heteroskedasticity consistent when the number of series terms grows as fast as the sample size. We also show that the new variance estimator provides an improvement when the number of terms grows slower than the sample size. These results suggest that the new standard errors should be useful for inference for a partially linear model with many regressors.

In Section 2 we describe the common structure of many instrument and small bandwidth asymptotics. In Section 3 we show how the structure leads to new results for the partially linear model. Section 4 describes previous results for the partially linear model and Section 5 gives the new results. Section 6 provides some Monte Carlo evidence and Section 7 concludes.

2 A Common Structure

To describe the common structure of many instrument and small bandwidth asymptotics, let $W_1, ..., W_n$ denote independent data observations. Also, let $\hat{\mu}$ denote an estimator and μ_0 a corresponding true value. We consider estimators that satisfy

$$\sqrt{n}\,(\hat{\mu} - \mu_0) = \hat{\Gamma}^{-1} S_n, S_n = \sum_{i,j=1}^{n} u_{ij}^n(W_i, W_j), \tag{1}$$

where $u_{ij}^n(w_i, w_j)$ is a function of a pair of observations that can depend on i, j, and n. We allow u to depend on n to account for number of terms or bandwidths that change with the sample size. Also, we allow u to vary with i and j to account for dependence on variables that are being conditioned on in the asymptotics, and so treated as nonrandom. We will illustrate in examples to follow.

We will assume throughout that $\hat{\Gamma} \xrightarrow{p} \Gamma$ non-singular and focus on the V-statistic S_n. That V-statistic has a well known decomposition that we describe here because it is an essential feature of the common structure. For notational implicitly we will drop the W_i and W_j arguments and let $u_{ij}^n = u_{ij}^n(W_i, W_j)$ and $\tilde{u}_{ij}^n = u_{ij}^n + u_{ji}^n - E[u_{ij}^n + u_{ji}^n]$. We have the following result.

PROPOSITION 1: *If* $\mathbb{E}[\|u_{ij}^n\|^2] < \infty$ *for all* i, j, n *then*

$$S_n = \Psi_n + U_n + B_n, \tag{2}$$

where

$$\begin{aligned}
\Psi_n &= \sum_{i=1}^n \psi_i^n(W_i), \quad \psi_i^n(W_i) = u_{ii}^n - \mathbb{E}[u_{ii}^n] + \sum_{j \neq i} \mathbb{E}[\tilde{u}_{ij}^n | W_i], \\
U_n &= \sum_{i=2}^n D_i^n(W_i, W_{i-1}, ..., W_1), \quad D_i^n(W_i, ..., W_1) = \sum_{j<i} \left(\tilde{u}_{ij}^n - \mathbb{E}[\tilde{u}_{ij}^n | W_i] - \mathbb{E}[\tilde{u}_{ij}^n | W_j] \right), \\
B_n &= \mathbb{E}[S_n],
\end{aligned}$$

and

$$\mathbb{E}[\psi_i^n(W_i)] = 0, \quad \mathbb{E}[D_i^n(W_i, ..., W_1) | W_{i-1}, ..., W_1] = 0, \quad \mathbb{E}[\Psi_n U_n] = 0.$$

This result shows that S_n can be decomposed into a sum of independent terms Ψ_n, a U-statistic remainder U_n that is a martingale difference sum and uncorrelated with Ψ_n, and a pure bias term B_n. This decomposition of a V-statistic is well known, being referred to by van der Vaart (1998, Chapter 11), and is included here for exposition. It is important in many of the proofs of asymptotic normality of semiparametric estimators, including Powell, Stock, and Stoker (1989), with the limiting distribution being determined by Ψ_n, and U_n being treated as a remainder that is of smaller order when the bandwidth shrinks slowly enough.

An interesting feature of this decomposition in semiparametric settings is that U_n is asymptotically normal at some rate when the number of series terms grow or the bandwidth shrinks to zero. In other settings the asymptotic behavior of U_n is more complicated. It is a degenerate U-statistic, that in general converges to a weighted sum of chi-squares, e.g. see van der Vaart (1998, Chapter 12). Apparently what occurs in semiparametric settings as the number of instruments grows or the bandwidth shrinks is that the individual contributions $D_i^n(W_i, ..., W_1)$ to U_n are small enough to satisfy a Lindeberg-Feller condition. Combined with the martingale property of U_n, as in Proposition 1 and is well known, this leads to asymptotic normality of U_n. This asymptotic normality property of U_n has been shown for both series and kernel estimators, as further explained below.

Alternative asymptotics occurs when the number of series terms grows or the bandwidth shrinks fast enough that Ψ_n and U_n have the same magnitude in the limit. Because of uncorrelatedness of Ψ_n and U_n the asymptotic variance will be larger than the usual formula which is $\lim_{n\longrightarrow\infty}\mathbb{V}[\Psi_n]$. Thus, consistent variance estimation under alternative asymptotics requires accounting for the presence of U_n.

Accounting for the presence of U_n should also yield improvements when numbers of series terms and bandwidths do not satisfy the knife-edge conditions of alternative asymptotics. Intuitively, if the number of series terms grows just slightly slower than the sample size or the bandwidth shrinks slightly slower than $1/n$. then accounting for the presence of U_n should still give a better large sample approximation. Hansen, Hausman, and Newey (2008) show such an improvement for many instrument asymptotics. It would be good to consider such improved approximations more generally, though it is beyond the scope of this paper to do so.

We can show that many instrument asymptotics has the structure we have outlined. To keep the exposition simple we focus on the jackknife instrumental variables estimator JIVE2 of Angrist, Imbens, and Krueger (1999). The limited information maximum likelihood estimator could also be considered, for which many instrument asymptotics was carried out by Kunitomo (1980) and Morimune (1983) some time ago, but the analysis is more complicated. Consider a linear structural equation

$$y_i = x_i'\beta_0 + \varepsilon_i, \quad \mathbb{E}[\varepsilon_i] = 0, \quad i = 1, ..., n,$$

where x_i is a vector and y_i and ε_i are scalar dependent variable and disturbance respectively. Let Z_i be a $K \times 1$ vector of instrumental variables that we treat as constants. It is also equivalent to allow instrumental variables to be random but condition on the matrix of observations $Z = [Z_1, ..., Z_n]'$ and replace unconditional moment conditions with conditional ones.

To describe the estimator let $Q = Z(Z'Z)^{-}Z'$ denote the projection matrix on the column space of Z. The JIVE2 estimator takes the form

$$\hat{\beta} = \left(\sum_{i\neq j} Q_{ij} x_i x_j'\right)^{-1} \sum_{i\neq j} Q_{ij} x_i y_j.$$

Substituting for y_j and collecting terms gives

$$\sqrt{n}(\hat{\beta} - \beta_0) = \left(\sum_{i \neq j} Q_{ij} x_i x_j' / n \right)^{-1} \sum_{i \neq j} Q_{ij} x_i \varepsilon_j / \sqrt{n}.$$

Here we can see that that JIVE2 is a special case of equation (1) with

$$\mu_n = \beta_0, \quad \hat{\Gamma} = \sum_{i \neq j} Q_{ij} x_i x_j' / n, \quad u_{ii}^n(W_i, W_i) = 0, \quad u_{ij}^n(W_i, W_j) = Q_{ij} x_i \varepsilon_j / \sqrt{n}, \quad i \neq j.$$

Note that $\mathbb{E}[u_{ij}^n(W_i, W_j)] = 0$ and for $\Upsilon_i = \mathbb{E}[x_i]$,

$$\mathbb{E}[u_{ij}^n(W_i, W_j)|W_i] = Q_{ij} x_i \mathbb{E}[\varepsilon_i] = 0, \quad \mathbb{E}[u_{ij}^n(W_i, W_j)|W_j] = Q_{ij} \Upsilon_i \varepsilon_j / \sqrt{n}.$$

Here Υ_i can be interpreted as the reduced form for observation i. Let $v_i = x_i - \Upsilon_i$. Applying Proposition 1, and using symmetry of the matrix Q, we find that equation (2) is satisfied with

$$\psi_i^n(W_i) = \left(\sum_{j \neq i} Q_{ij} \Upsilon_j \right) \varepsilon_i = \Upsilon_i (1 - Q_{ii}) \varepsilon_i / \sqrt{n} - \left(\Upsilon_i - \sum_i Q_{ij} \Upsilon_j \right) \varepsilon_i / \sqrt{n},$$

$$D_i^n(W_i, ..., W_1) = \sum_{j < i} Q_{ij} \left(v_i \varepsilon_j + v_j \varepsilon_i \right) / \sqrt{n}, \quad B_n = 0.$$

Note that $\Upsilon_i - \sum_i Q_{ij} \Upsilon_j$ is the i-th residual from regressing the reduced form observations on Z, so that by appropriate definition of the reduced form this can generally be assumed to go to zero as the sample size grows. In that case

$$\Psi_n = \sum_i \Upsilon_i (1 - Q_{ii}) \varepsilon_i / \sqrt{n} + o_p(1).$$

Furthermore, under standard asymptotics Q_{ii} will go to zero, so the variance of this does indeed correspond to the usual asymptotic variance for IV.

The degenerate U-statistic term is

$$U_n = \sum_{i=1}^{n} \sum_{j < i} Q_{ij} \left(v_i \varepsilon_j + v_j \varepsilon_i \right) / \sqrt{n}.$$

Chao, Swanson, Hausman, Newey, and Woutersen (2010) apply the martingale central limit theorem to show that this U_n will be asymptotically normal when x_i and ε_i have uniformly bounded fourth moments, $\text{rank}(Q) = \dim(Z) = K \longrightarrow \infty$, and Q_{ii} is bounded away from 1 uniformly in i and n. The conditions of the martingale central limit theorem are verified by showing that certain linear combinations with coefficients depending on the elements of Q go to zero as $K \longrightarrow \infty$. In the proof, this makes individual terms asymptotically negligible, with a Lindeberg-Feller condition being satisfied. Alternative asymptotics occurs when K grows as fast as n, resulting in Ψ_n and U_n having the same magnitude in the limit.

As mentioned above, small bandwidth asymptotics for kernel estimators also has the structure outlined above. To illustrate this we consider kernel estimation of the integrated squared density. We use this example to keep the exposition relatively simple and because it shares the common structure with the more interesting density weighted average derivative estimator of Powell, Stock, and Stoker (1989) treated in Cattaneo, Crump, and Jansson (2010). In this example the parameter of interest is

$$\mu_0 = \int f_0(w)^2 \mathrm{d}w = \mathbb{E}[f_0(W_i)],$$

where W_i denotes a continuously distributed random variable with pdf f_0. A "leave one out estimator," analogous to that of Powell, Stock, and Stoker (1989), is

$$\hat{\mu} = \sum_{i \neq j} K_h(W_i - W_j)/n(n-1),$$

where $K(u)$ be a symmetric kernel with $K(-u) = K(u)$ and $K_h(u) = h^{-d}K(u/h)$. This estimator has the V-statistic form of equation (1) with

$$\mu_0 = \int f_0(w)^2 \mathrm{d}w, \quad \hat{\Gamma} = 1, \quad u^n_{ii}(W_i, W_i) = 0, \quad u^n_{ij}(W_i, W_j) = K_h(W_i - W_j)/\sqrt{n}(n-1), \quad i \neq j.$$

Define

$$f_h(w) = \int K(u) f(w + hu) \mathrm{d}u.$$

Note that by symmetry of $K(u)$,

$$\mathbb{E}[u^n_{ij}(W_i, W_j)|W_i] = f_h(W_i)/\sqrt{n}(n-1), \quad \mathbb{E}[u^n_{ij}(W_i, W_j)|W_j] = f_h(W_j)/\sqrt{n}(n-1),$$

$$\mathbb{E}[u^n_{ij}(W_i, W_j)] = \mathbb{E}[f_h(X_i)]/\sqrt{n}(n-1).$$

Applying Proposition 1, and using symmetry of $K(u)$, we find that equation (2) is satisfied with

$$\psi_i^n(W_i) = 2\{f_h(W_i) - E[f_h(W_i)]\}/\sqrt{n},$$

$$D_i^n(W_i, ..., W_1) = 2\sum_{j<i}\{K_h(W_i - W_j) - f_h(W_i) - f_h(W_j) + E[f_h(W_i)]\}/\sqrt{n}(n-1),$$

$$B_n = \sqrt{n}\{E[f_h(W_i)] - \mu_0\}.$$

Note that $2\{f_h(W_i) - E[f_h(W_i)]\}$ is an approximation to the well known influence function $2[f_0(W_i) - \mu_0]$ for estimators of the integrated squared density. As $h \to 0$ we will have $f_h(W_i)$ converging to $f_0(W_i)$ in mean square, so that

$$\Psi_n = \sum_i 2[f_0(W_i) - \mu_0]/\sqrt{n} + o_p(1).$$

The asymptotic variance of this leading term does correspond to the usual asymptotic variance for estimators of the integrated square density.

The degenerate U-statistic term is

$$U_n = 2\sum_{i=1}^{n}\sum_{j<i}\{K_h(W_i - W_j) - f_h(W_i) - f_h(W_j) + E[f_h(W_i)]\}/\sqrt{n}(n-1).$$

The martingale central limit theorem can be applied as in Cattaneo, Crump, and Jansson (2010) to show that this U_n will be asymptotically normal as $h \to 0$ and $n \to \infty$ growing. Intuitively, the bandwidth shrinking leads to $K_h(W_i - W_j)$ being small except when W_i and W_j are close, so that tail of the distribution of $\sqrt{n}D_i^n(W_i, ..., W_1)$ becomes thin and a Lindeberg-Feller condition is satisfied. Alternative asymptotics occurs when h shrinks as fast as $1/n$, resulting in Ψ_n and U_n having the same magnitude in the limit.

This pair of examples shows how several estimators share the common structure outlined above. We now show how that structure can be applied to derive new results.

3 Series Estimation of the Partially Linear Model

To describe the partially linear model let $(y_i, x_i', z_i')'$, $i = 1, \ldots, n$, be a random sample of the random vector $(y, x', z')'$, where $y \in \mathbb{R}$ is a dependent variable, and $x \in \mathbb{R}^{d_x \times 1}$ and $z \in \mathbb{R}^{d_z \times 1}$

are explanatory variables. The model is given by

$$y_i = x_i'\beta + g(z_i) + \varepsilon_i, \quad \mathbb{E}\left[\varepsilon_i|x_i, z_i\right] = 0, \quad \sigma^2(x_i, z_i) = \mathbb{E}\left[\varepsilon_i^2|x_i, z_i\right],$$

where $v_i = x_i - h(z_i)$, with $h(z_i) = \mathbb{E}\left[x_i|z_i\right]$.[2] A series estimator of β is obtained by regressing y_i on x_i and approximating functions of z_i. To describe the estimator let $p^K(z) = (p_{K1}(z), \ldots, p_{KK}(z))'$ be an approximating functions, such as polynomials or splines, where K denotes the number of terms in the regression. Here the unknown function $g(z)$ will be approximated by a linear combination of $p_{kK}(z)$. Also, let $Y = [y_1, \cdots, y_n]' \in \mathbb{R}^{n\times 1}$, $X = [x_1, \cdots, x_n]' \in \mathbb{R}^{n\times d_x}$, $P = [p^K(z_1), \ldots, p^K(z_n)]'$. A series estimator of β is given by

$$\hat{\beta} = (X'MX)^{-1}X'MY, \quad M = I - Q, \quad Q = P(P'P)^{-}P',$$

where A^- denotes a generalized inverse of a matrix A (satisfying $AA^-A = A$) and $X'MX$ will be non-singular with probability approaching one under conditions given below. Donald and Newey (1994) gave conditions for the asymptotic normality of this estimator.

Conditional on $Z = [z_1, ..., z_n]'$ this estimator has the structure outlined earlier. For this reason we will condition on Z throughout the following discussion so that expectations are implicitly conditional on Z. To explain how $\hat{\beta}$ fits within the common structure, let $h_i = \mathbb{E}[x_i]$ and $v_i = x_i - h_i$. Also let $g_i = g(z_i)$,

$$G = [g_1, \cdots, g_n]', \quad \varepsilon = [\varepsilon_1, \cdots, \varepsilon_n]', \quad H = [h_1, \cdots, h_n]', \quad V = [v_1, \cdots, v_n]'.$$

By $Y = X\beta + G + \varepsilon$ we have

$$\sqrt{n}(\hat{\beta} - \beta) = (X'MX/n)^{-1}X'M(\varepsilon + G)/\sqrt{n} = \hat{\Gamma}^{-1}S_n,$$

$$\hat{\Gamma} = X'MX/n, \quad S_n = \sum_{i,j} x_i M_{ij}(g_j + \varepsilon_j)/\sqrt{n}.$$

This estimator has the V-statistic form of equation (1) with $W_i = (y_i, x_i)$ and

$$\mu_0 = \beta, \quad \hat{\Gamma} = X'MX/n, \quad u_{ij}^n(W_i, W_j) = x_i M_{ij}(g_j + \varepsilon_j)/\sqrt{n}.$$

By $\mathbb{E}[\varepsilon_i|x_i] = 0$ we have $\mathbb{E}[x_i\varepsilon_i] = 0$. Therefore, letting $u_{ij}^n = u_{ij}^n(W_i, W_j)$ as we have done

[2] See Robinson (1988) for the analysis of this model when using kernel regression, and Linton (1995) for the corresponding higher-order properties.

previously, we have

$$\mathbb{E}[u_{ij}^n] = H_i M_{ij} g_j/\sqrt{n}, \quad u_{ij}^n - \mathbb{E}[u_{ij}^n] = M_{ij}\left(v_i g_j + x_i \varepsilon_j\right)/\sqrt{n},$$

$$\tilde{u}_{ij} = M_{ij}\left(v_j g_i + v_i g_j + x_j \varepsilon_i + x_i \varepsilon_j\right)/\sqrt{n}, \quad \mathbb{E}[\tilde{u}_{ij}|W_i] = M_{ij}\left(v_i g_j + h_j \varepsilon_i\right).$$

Plugging these expressions into the formulas on Proposition 1 we obtain

$$\begin{aligned}\psi_i^n(W_i) &= M_{ii}\left(v_i g_i + h_i \varepsilon_i + v_i \varepsilon_i\right)/\sqrt{n} + \sum_{j \neq i} M_{ij}\left(v_i g_j + \varepsilon_i h_j\right)/\sqrt{n} \\ &= \left[v_i M_{ii} \varepsilon_i + v_i \sum_j M_{ij} g_j + \varepsilon_i \sum_j h_j M_{ij}\right]/\sqrt{n},\end{aligned}$$

$$D_i^n(W_i, ..., W_1) = \sum_{j<i} M_{ij}\left(v_i \varepsilon_j + v_j \varepsilon_i\right)/\sqrt{n}, B_n = H'MG/\sqrt{n}.$$

Summing up $\psi_i^n(W_i)$ it follows that

$$\Psi_n = \sum_i M_{ii} v_i \varepsilon_i/\sqrt{n} + \left(V'MG + H'M\varepsilon\right)/\sqrt{n}.$$

The last term in this expression has mean zero and will converge to zero in mean square as K grows, as further discussed below. Therefore, by $M_{ii} = 1 - Q_{ii}$,

$$\Psi_n = \sum_i (1 - Q_{ii}) v_i \varepsilon_i/\sqrt{n} + o_p(1). \tag{3}$$

Under standard asymptotics Q_{ii} will go to zero and hence the variance of this corresponds to the usual asymptotic variance.

Noting that for $j < i$, $M_{ij} = -Q_{ij}$, plugging in we find that the degenerate U-statistic term is

$$U_n = -\sum_{i=1}^n \sum_{j<i} Q_{ij}\left(v_i \varepsilon_j + v_j \varepsilon_i\right)/\sqrt{n}.$$

Remarkably this term is essentially the same as the degenerate U-statistic term for JIVE2 that was shown above. Consequently, the central limit theorem of Chao, Swanson, Hausman, Newey, and Woutersen (2010) applies immediately to show that U_n is asymptotically normal as $K \longrightarrow \infty$. This leads to many regressor asymptotics for series estimators of the partially linear model. It also illustrates the usefulness of the common structure, showing how that structure can be used to derive new results.

4 Standard Asymptotics for the Series Estimator

The estimator $\hat{\beta}$ may be intuitively interpreted as a two-step semiparametric estimator with smoothing parameter $p_K(\cdot)$ and tuning parameter K, because the unknown (regression) functions $g(\cdot)$ and $h(\cdot)$ are non-parametrically estimated in a preliminary step by the series estimator. In particular, the following assumption characterizes the rate at which the approximation error of series estimator should vanish.

Assumption B. (i) For some $\alpha_h > 0$, there exists η_h so that

$$\frac{1}{n}H'MH = \frac{1}{n}\min_{\eta}\|H - P\eta\|^2 = \frac{1}{n}\sum_{i=1}^{n}\left\|h(z_i) - p_K(z_i)'\eta_h\right\|^2 = O_{as}(K^{-2\alpha_h}).$$

(ii) For some $\alpha_g > 0$, there exists η_g so that

$$\frac{1}{n}G'MG = \frac{1}{n}\min_{\eta}\|G - P\eta\|^2 = \frac{1}{n}\sum_{i=1}^{n}\left[g(z_i) - p_K(z_i)'\eta_g\right]^2 = O_{as}(K^{-2\alpha_g}).$$

The conditions required in Assumption B are implied by conventional assumptions from the series-based nonparametric literature (e.g., Newey (1997, Assumption 3)). Thus, under appropriate assumptions, commonly used basis of approximation such as polynomials or splines will satisfy Assumption B with $\alpha_h = d_z/s_h$ and $\alpha_g = d_z/s_g$, where s_h and s_g denotes the number of continuous derivatives of h and g, respectively.

Under regularity conditions (given below) and Assumption B, Donald and Newey (1994) obtained the following (infeasible) classical asymptotic approximation for $\hat{\beta}$ corresponding to equation (3): If

$$nK^{-2(\alpha_h+\alpha_g)} \to 0 \qquad \text{and} \qquad \frac{K}{n} \to 0 \tag{4}$$

then

$$\sqrt{n}(\hat{\beta} - \beta) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_i + o_p(1) \to_d \mathcal{N}(0, \Omega), \qquad \Omega = \Gamma^{-1}\Sigma\Gamma^{-1}, \tag{5}$$

where

$$\psi_i = \Gamma^{-1}v_i\varepsilon_i, \qquad \Gamma = \mathbb{E}[v_iv_i'], \qquad \Sigma = \mathbb{E}[v_i\mathbb{V}[\varepsilon|X,Z]v_i'].$$

The classical asymptotic linear representation of $\hat{\beta}$, given in (5), is established by analyzing the first-order stochastic properties of the "numerator" $X'M(Y-X\beta)/\sqrt{n}$ and "denominator" $\hat{\Gamma}$ of the estimator. Specifically, for the "denominator", it can be shown (see Lemma 1 below) that if Condition (4) holds, then the (Hessian) matrix satisfies (recall that $M = I - Q$ and $QP = P$)

$$\begin{aligned}\hat{\Gamma}_n &= \frac{1}{n}X'MX = \frac{1}{n}V'MV + o_p(1) = \frac{1}{n}V'V - \frac{1}{n}V'QV + o_p(1) \\ &= \frac{1}{n}\sum_{i=1}^{n} v_i v_i' + o_p(1) \to_p \Gamma,\end{aligned}$$

where the second equality captures the bias introduced by the series estimator (Assumption B(i)), and the fourth equality requires the contribution of $V'QV$ to vanish asymptotically. Similarly, it follows from equation (3) that the "numerator" of $\hat{\beta}$ satisfies

$$\frac{1}{\sqrt{n}}X'M\left(Y - X\beta\right) = \sum_i v_i(1-Q_{ii})\varepsilon_i/\sqrt{n} + o_p(1) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} v_i\varepsilon_i + o_p(1) \to_d \mathcal{N}\left(0, \Sigma\right),$$

where the first equality is again related to the bias introduced by the nonparametric estimator and the second equality will follow by $Q_{ii} \longrightarrow 0$.

In both cases, the approximation error associated with the bias is controlled by the condition $nK^{-2(\alpha_h+\alpha_g)} \to 0$, which requires K to be "large". On the other hand, condition $K/n \to 0$ guarantees that both $V'QV$ and $V'Q\varepsilon$ are asymptotically negligible, as required for the classical, asymptotically linear, approximation to be valid. The latter condition controls the variance of the estimator, and it is directly related to the behavior of the nonparametric estimator, which in this case is described by Q.

The classical approach to form a confidence interval for β_0 is to use the asymptotic distributional result coupled with a consistent standard errors estimator. A plug-in approach employs the (asymptotically) pivotal test statistic $T_{0,n}(K) = \Omega_0^{-1/2}\sqrt{n}(\hat{\beta}-\beta)$, together with a plug-in consistent estimator for Ω_0. Under heteroskedasticity, a feasible test statistic is given by

$$\hat{T}_{0,n}(K_n) = \hat{\Omega}_n^{-1/2}\sqrt{n}(\hat{\beta}-\beta), \qquad \hat{\Omega}_n = \hat{\Gamma}_n^{-1}\hat{\Sigma}_n\hat{\Gamma}_n^{-1},$$

where

$$\hat{\Gamma}_n = \frac{X'MX}{n}, \qquad \hat{\Sigma}_n = \frac{1}{n-K-d_x}\sum_{i=1}^{n}\tilde{x}_i\tilde{x}_i'\hat{\varepsilon}_i^2, \qquad \hat{\varepsilon} = \tilde{Y} - \tilde{X}'\hat{\beta} = [\hat{\varepsilon}_1, \cdots, \hat{\varepsilon}_n]'.$$

In this case, the standard error estimator is the classical Heteroskedasticity-Robust (HR) standard errors estimator commonly used in regression analysis. Under Condition (4) it is not difficult to show that $\hat{\Omega}_n^{-1}\Omega_0 \to_p I_{d_x}$.

5 Many Regressor Asymptotics

This section derives a generalized asymptotic distribution for $\sqrt{n}(\hat{\beta}-\beta)$, which relaxes the condition $K_n/n \to 0$. This non-standard asymptotic theory encompasses the classical result discussed in the previous section, and also captures the effect of the degenerate U-statistic term of the expansion, which is negligible when $K/n \longrightarrow 0$. Intuitively, this asymptotic experiment captures the effect of K "large" (relative to n) by breaking down the asymptotic linearity of the estimator.

To characterize the generalized central limit theory, it is natural to study the stochastic behavior of the estimator $\sqrt{n}(\hat{\beta}-\beta)$ as a "ratio" of two bilinear forms:

$$\sqrt{n}(\hat{\beta}-\beta) = \left(\frac{1}{n}X'MX\right)^{-}\frac{1}{\sqrt{n}}X'M\left(Y-X\beta\right) = \hat{\Gamma}_n^{-}\frac{1}{\sqrt{n}}X'M\left(Y-X\beta\right).$$

The following lemma characterizes the behavior of the Hessian matrix $\hat{\Gamma}_n$ under quite weak conditions.

Lemma 1. Suppose that $\mathbb{E}[\|v_i\|^4] \leq C_v < \infty$ (a.s.) and Assumption B(i) holds. Then,

$$\hat{\Gamma}_n = \frac{1}{n}X'MX = \Gamma_n + o_p(K/n), \qquad \Gamma_n = \frac{1}{n}\sum_{i=1}^{n} M_{ii}\mathbb{E}\left[v_i v_i'|z_i\right] = O_{as}(1+K/n).$$

This lemma characterizes the stochastic behavior of the Hessian matrix under conditions that are weaker than those entertained by the classical, asymptotically linear, distribution theory. Specifically, because $M = I - Q$ is an idempotent symmetric matrix, $M_{ii} \in (0,1)$ and $\sum_{i=1}^{n} M_{ii} \leq n-K$, Lemma 1 implies that $\hat{\Gamma}_n$ remains asymptotically bounded even when $K/n \not\to 0$. In particular, $\Gamma_n = \mathbb{E}\left[v_i v_i'\right] + o_p(1)$ when $K/n \to 0$. Moreover, in the case of homoskedasticity of v_i, that is, if $\mathbb{E}\left[v_i v_i'|z_i\right] = \mathbb{E}\left[v_i v_i'\right]$ (and $\operatorname{rank}(Q) = K$), then $\Gamma_n = (1-K/n)\,\mathbb{E}\left[v_i v_i'\right]$. Finally, if $\lambda_{\min}\left(\mathbb{E}\left[v_i v_i'|z_i\right]\right) \geq F_V > 0$ and $M_{ii} = 1 - Q_{ii} \geq F_Q > 0$ (a.s.) then $\lambda_{\min}(\Gamma_n) \geq F_Q F_V > 0$. ($\lambda_{\min}(A)$ denotes the minimum eigenvalue of a matrix A.)

To fully characterize the asymptotic behavior of the "numerator" of $\sqrt{n}(\hat{\beta}-\beta)$ it is convenient to proceed in two steps. First, under appropriate bias assumptions, it is possible to show that the numerator is asymptotically equivalent to a quadratic form based on mean-zero random variables.

Lemma 2. Suppose the assumptions of Lemma 1 hold, and $\mathbb{E}[\varepsilon_i^4] \leq C_\varepsilon < \infty$ (a.s.) and Assumption B(ii) holds. Then,

$$\left(V'MG + H'M\varepsilon + H'MG\right)/\sqrt{n} = O_p(\sqrt{n}K^{-(\alpha_h+\alpha_g)} + K^{-\alpha_h} + K^{-\alpha_g}).$$

As in Lemma 1, this result only requires bounded moments and a bias condition. In this case, the bias arises from the approximation of both unknown functions h and g. As mentioned above, the high-level Assumption B is implied by the standard assumption of best approximation from the sieve literature. Interestingly, in this model there is a trade-off in terms of curse of dimensionality: provided that $\min\{\alpha_h, \alpha_g\} > 0$, the bias condition is given by $\sqrt{n}K^{-(\alpha_h+\alpha_g)} \to 0$, which implies a trade-off between smoothness and dimensionality between h and g.

Lemma 2 justifies focusing on the bilinear form

$$\theta_n = \frac{1}{\sqrt{n}}V'M\varepsilon = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\sum_{j=1}^{n} M_{ij}v_i\varepsilon_j,$$

where $\mathbb{E}\left[\theta_n | Z, X\right] = 0$. Moreover, under the assumptions imposed in Lemma 2, a simple variance calculation yields

$$\Sigma_n = \mathbb{V}\left[\theta_n|Z\right] = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} M_{ij}^2\mathbb{E}[v_iv_i'\varepsilon_j^2] = O_{as}\left(1 + K/n\right).$$

In particular, if $K/n \to 0$ then

$$\Sigma_n = \frac{1}{n}\sum_{i=1}^{n} M_{ii}^2\mathbb{E}\left[v_iv_i'\varepsilon_i^2|z_i\right] + o_p\left(1\right) = \mathbb{E}\left[v_iv_i'\varepsilon_i^2\right] + o_p\left(1\right) = \Sigma + o_p\left(1\right),$$

as given in Section 2. Moreover, under homoskedasticity, that is, if $\mathbb{E}\left[\varepsilon_i^2|x_i, z_i\right] = \sigma^2$, then

$$\mathbb{V}\left[\left.\frac{1}{\sqrt{n}}V'M\varepsilon\right| X, Z\right] = \frac{\sigma^2}{n}V'MV,$$

and

$$\Sigma_n = \frac{\sigma^2}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} M_{ij}^2 \mathbb{E}\left[v_i v_i' | z_i\right] = \sigma^2 \Gamma_n,$$

because $\sum_{j=1}^{n} M_{ij}^2 = M_{ii}$. Furthermore, if $\mathbb{E}\left[\varepsilon_i^2 | x_i, z_i\right] = \sigma^2$ and $\mathbb{E}\left[v_i v_i' | z_i\right] = \mathbb{E}\left[v_i v_i'\right]$ (and $\operatorname{rank}(Q) = K$ (a.s.)), then $\Sigma_n = \sigma^2 \left(1 - K/n\right) \mathbb{E}\left[v_i v_i'\right]$. Finally, if $\lambda_{\min}\left(\mathbb{E}\left[v_i v_i' \varepsilon^2 | z_i\right]\right) \geq F_{V\varepsilon} > 0$ and $M_{ii} = 1 - Q_{ii} \geq F_Q > 0$ (a.s.), then $\lambda_{\min}\left(\Sigma_n\right) \geq F_Q^2 F_{V\varepsilon} > 0$.

The following Lemma characterizes the asymptotic distribution of the bilinear form θ_n.

Lemma 3. Suppose the assumptions of Lemma 2 hold, and $M_{ii} > F_M > 0$ (a.s.) and $\lambda_{\min}\left(\Sigma_n\right) > F_\Sigma > 0$ (a.s.). Then,

$$\Sigma_n^{-1/2} \frac{1}{\sqrt{n}} V' M \varepsilon \to_d \mathcal{N}\left(0, I_{d_x}\right).$$

The following theorem is a direct consequence of the previous lemmas and Slutsky Theorem, and constitutes the main result of this section.

Theorem 1. Suppose the assumptions of Lemma 3 hold and suppose $\lambda_{\min}\left(\Gamma_n\right) > F_H > 0$. Then, if

$$n K^{-2(\alpha_x + \alpha_g)} \to 0 \qquad \text{and} \qquad \frac{K}{n} \to \alpha \in [0, 1) \tag{6}$$

then

$$\hat{\Omega}_n^{-1/2}(\hat{\beta} - \beta) \to_d \mathcal{N}\left(0, I_{d_x}\right). \tag{7}$$

If, moreover, $\left(\Gamma_n, \Sigma_n\right) = \left(\Gamma_\infty, \Sigma_\infty\right) + o_p(1)$ for some $\left(\Gamma_\infty, \Sigma_\infty\right)$, then

$$\sqrt{n}(\hat{\beta} - \beta) \to_d \mathcal{N}\left(0, \Omega_\infty\right), \qquad \Omega_\infty = \Gamma_\infty^{-1} \Sigma_\infty \Gamma_\infty^{-1}.$$

If, moreover, $\mathbb{E}\left[\varepsilon_i^2 | x_i, z_i\right] = \sigma^2$ for some σ^2, then

$$\sqrt{n}(\hat{\beta} - \beta) \to_d \mathcal{N}\left(0, \sigma^2 \Gamma_\infty^{-1}\right).$$

Theorem 1 shows that the central limit theorem for $\hat{\beta}$ holds under the weaker condition (6). (Compare to Condition (4).) This result does not rely on asymptotic linearity, nor on the actual convergence of the matrices Γ_n and Σ_n. However, if $K/n \to 0$, then $\left(\Gamma_n, \Sigma_n\right) = (\Gamma, \Sigma) + o_p(1)$ with $\Gamma = \mathbb{E}\left[v_i v_i'\right]$ and $\Sigma = \mathbb{E}\left[v_i v_i' \varepsilon_i^2\right]$, and the resulting large sample distribution theory does rely on the asymptotically linear representation of $\hat{\beta}$, as given in (5).

Importantly, if $K/n \nrightarrow 0$ and $(\Gamma_n, \Sigma_n) = (\Gamma, \Sigma) + o_p(1)$, then $\Gamma \neq \mathbb{E}[v_i v_i']$ and $\Sigma \neq \mathbb{E}[v_i v_i' \varepsilon_i^2]$, in general. For instance, if (v_i, ε_i) is independent of z_i, then $\Gamma_n = (1 - K/n)\,\mathbb{E}[v_i v_i']$ and

$$\Sigma_n = \left(1 - \frac{K}{n}\right) \mathbb{E}\left[v_i v_i' \varepsilon_i^2\right] + \left(\frac{1}{n}\sum_{i=1}^{n} Q_{ii}^2 - \frac{K}{n}\right)\left(\mathbb{E}\left[v_i v_i' \varepsilon_i^2\right] - \mathbb{E}\left[v_i v_i'\right]\mathbb{E}\left[\varepsilon_i^2\right]\right).$$

6 Standard Errors

This section discusses different homoskedastic- and heteroskedastic- standard errors estimators, and their properties under the generalized asymptotics studied in this paper.

6.1 Homoskedasticity

If $\mathbb{E}[\varepsilon_i^2 | x_i, z_i] = \sigma^2$ for all $i = 1, 2, \cdots, n$, then $\Sigma_n = \sigma^2 \Gamma_n^-$. Thus, a natural plug-in estimator is given by $\hat{V}_{HOM} = \hat{\sigma}^2 \hat{\Gamma}^-$, where $\hat{\sigma}^2$ is chosen so that $\hat{\sigma}^2 = \sigma^2 + o_p(1)$. The usual OLS estimator is

$$\hat{\sigma}_n^2 = \frac{1}{n - K - d_x} \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \frac{1}{n - K - d_x} \hat{\varepsilon}' \hat{\varepsilon}.$$

As shown in Lemma 1, $\hat{\Gamma}_n = \Gamma_n + o_p(1)$ under the many terms asymptotics discussed in this paper (Condition (6)), and therefore it remains to verify that $\hat{\sigma}_n^2$ is also consistent under this generalized asymptotic experiment. This estimator is consistent because

$$\hat{\varepsilon}'\hat{\varepsilon} = (Y - X\hat{\beta})' M (Y - X\hat{\beta}) = \varepsilon' M \varepsilon + o_p(1) = (n - K)\,\sigma^2 + o_p(1),$$

where the first approximation is based on the $\sqrt{n}$-consistency of $\hat{\beta}$ and the approximation bias of the series estimator, while the second approximation is based on the fact that the bilinear form $\varepsilon' M \varepsilon = \varepsilon'(I - Q)\varepsilon$ is dominated by its diagonal.

Theorem 2. Suppose the assumptions of Theorem 1 hold. Then, $\hat{\sigma}_n^2 = \sigma^2 + o_p(1)$.

It follows by Lemma 1, Theorem 2 and Slutsky Theorem that

$$\hat{V}_{HOM} = \sigma^2 \Gamma_n^- + o_p(1) = \Sigma_n + o_p(1),$$

and therefore

$$\hat{V}_{HOM}^{-1/2}(\hat{\beta} - \beta_0) \to_d \mathcal{N}(0, I_d),$$

under Condition (6). Thus, under known homoskedasticity, the usual finite sample standard errors estimator from least squares theory turns out to be valid even when K is large. However, the "consistent" but biased standard errors estimator $(\hat{\varepsilon}'\hat{\varepsilon}/n)\Gamma_n^-$ will not be consistent unless $K/n \to 0$, which implies that using the finite sample, unbiased standard errors estimator under homoskedasticity is important even in large samples when the degrees of freedom is small (i.e., when K is large).

6.2 Heteroskedasticity

Under heteroskedasticity of unknown form, a natural candidate for standard errors estimator is the (family of) Eicker-Huber-White estimators given by

$$\hat{V}_{HR} = (\tilde{X}'\tilde{X})^- \tilde{X}'\Upsilon\tilde{X}(\tilde{X}'\tilde{X})^- = \frac{1}{n}\hat{\Gamma}_n^- \hat{\Sigma}_n \hat{\Gamma}_n^-,$$

$$\hat{\Sigma}_n = \frac{1}{n}X'M\Upsilon MX, \qquad \Upsilon = \operatorname{diag}\left(\omega_1\hat{\varepsilon}_1^2, \cdots, \omega_n\hat{\varepsilon}_n^2\right) = \begin{bmatrix} \omega_1\hat{\varepsilon}_1^2 & & \\ & \ddots & \\ & & \omega_n\hat{\varepsilon}_n^2 \end{bmatrix}$$

where $\{\omega_i : i = 1, \cdots, n\}$ are appropriate weights. Classical choices of weights include $\omega_i = 1$, $\omega_i = n/(n - K - d_x)$, $\omega_i = M_{ii}^{-1}$, etc. Since $\hat{\Gamma}_n = \Gamma_n + o_p(1)$ according to Lemma 1, it only remains to characterized the middle matrix of this classical sandwich formula. The asymptotic properties of $\hat{\Sigma}_n$ are given by

$$\hat{\Sigma}_n = \frac{1}{n}\sum_{i=1}^n \omega_i\tilde{x}_i\tilde{x}_i'\hat{\varepsilon}_i^2 = \frac{1}{n}\sum_{i=1}^n \omega_i\tilde{v}_i\tilde{v}_i'\tilde{\varepsilon}_i^2 + o_p(1) = \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\omega_i\tilde{v}_i\tilde{v}_i'\tilde{\varepsilon}_i^2\middle| Z\right] + o_p(1),$$

where, as before, the first approximation captures the bias of the series estimator and removes the estimation error of β, and the second approximation shows that a (conditional) law of large numbers holds in this case (i.e., a variance condition). This results is summarized in the following theorem.

Theorem 3. Suppose the assumptions of Theorem 1 hold, $nK^{-2(\alpha_g+\alpha_h)+1} \to 0$ with $\min\{\alpha_g, \alpha_h\} > 1/2$, and $\omega_i \in \sigma(Z)$ for all i with $|\omega_i| \le C_\omega$. Then, $\hat{\Sigma}_n = \tilde{\Sigma}_n + o_p(1)$, where

$$\tilde{\Sigma}_n = \frac{1}{n}\sum_{i=1}^n\sum_{j=1}^n\left(\sum_{\ell=1}^n \omega_i M_{i\ell}^2 M_{j\ell}^2\right)\mathbb{E}\left[v_i v_i'\varepsilon_j^2\middle| z_i, z_j\right].$$

Recall that the population asymptotic middle matrix Σ_n is given by

$$\Sigma_n = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} M_{ij}^2 \mathbb{E}[v_i v_i' \varepsilon_j^2 | z_i, z_j],$$

which implies that the classical HR standard errors will not be consistent in general when Condition (6). For example, the bias may be characterized under homoskedasticity: assuming $\mathbb{E}\left[\varepsilon_i^2 | x_i, z_i\right] = \sigma^2$ and $\omega_i = 1$, a simple calculation yields

$$\Sigma_n = \frac{\sigma^2}{n}\sum_{i=1}^{n} M_{ii}\mathbb{E}\left[v_i v_i' | z_i\right],$$

while, using basic properties of idempotent matrices, it is easy to verify that

$$\tilde{\Sigma}_n = \Sigma_n - \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} M_{ij}^2 \left(1 - M_{jj}\right)\mathbb{E}\left[v_i v_i' | z_i\right] < \Sigma_n.$$

As a consequence, the classical Eicker-Huber-White estimator is downward biased when K is "large". It is important to note that this result continue to hold in a simple linear model where the number of regressors is "large" when compared to the sample size. Therefore, there is an important sense in which the classical HR standard errors estimator is not robust, even in a simple linear model.

On the other hand, if $K/n \to 0$, then the asymptotic results presented above imply that $\tilde{\Sigma}_n = \Sigma_n + o_p(1)$, which verifies that the classical HR standard errors estimator is indeed consistent under heteroskedasticity of unknown form.

6.3 HR and Many Terms Robust Standard Errors

Intuitively, the failure of the classical HR standard errors estimator is due to the fact that both $\tilde{x}_i$ and $\hat{\varepsilon}_i$ are estimated with too much error when $K/n \not\to 0$. Thus, it is possible to fix this problem by considering alternative (consistent) estimators for either $\tilde{x}_i$ or $\hat{\varepsilon}_i$. To describe the new estimators, let K_g and K_h be two choices of truncation for an approximation series, and let

$$\tilde{X} = (I - Q_h)X, \qquad Q_h = P_{K_h}(P_{K_h}' P_{K_h})^- P_{K_h}', \qquad M_h = I - Q_h,$$

$$\hat{\varepsilon} = (I - Q_g)\varepsilon, \qquad Q_g = P_{K_g}(P_{K_g}' P_{K_g})^- P_{K_g}', \qquad M_g = I - Q_g.$$

Using this notation, Theorem 3 may be extended to the following result.

Theorem 4. Suppose the assumptions of Theorem 1 hold, $nK_g^{1/2}K_g^{-2\alpha_g}K_h^{1/2}K_h^{-2\alpha_h} \to 0$ with $\min\{\alpha_g, \alpha_h\} > 1/2$, and $\omega_i \in \sigma(Z)$ for all i with $|\omega_i| \leq C_\omega$. Then

$$\check{\Sigma}_n = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\sum_{\ell=1}^{n}\omega_i M_{h,i\ell}^2 M_{g,j\ell}^2\right)\mathbb{E}\left[v_i v_i' \varepsilon_j^2 | z_i, z_j\right].$$

This theorem leads naturally to two alternative recipes for heteroskedasticity and many terms robust estimators. Specifically, for $\omega_i = 1$, if $\min\{K_h, K_g\} = o(n)$ and $\max\{K_h, K_g\} = K$, then $\check{\Sigma}_n = \Sigma_n + o_p(1)$. This result implies that

$$\hat{V}_{HET}^{-1/2}(\hat{\beta} - \beta_0) \to_d \mathcal{N}(0, I_d), \qquad \hat{V}_{HET} = \frac{1}{n}\hat{\Gamma}_n^{-}\check{\Sigma}_n\hat{\Gamma}_n^{-}.$$

7 Simulations

To explore the consequences of using many terms in the partially linear model, or alternatively using many covariates in a linear model, this section reports preliminary results from a Monte Carlo experiment. Specifically, the simulations consider the following model:

$$\begin{aligned} y_i &= x_i'\beta + g(z_i) + \varepsilon_i, & \varepsilon_i &= \sigma_\varepsilon(x_i, z_i)\, u_{1i}, \\ x_i &= h(z_i) + v_i, & v_i &= \sigma_v(z_i)\, u_{2i}, \end{aligned}$$

with $d_x = 1$, $d_z = 10$, $g(z) = 1$, $h(z) = 0$, and $u = (u_{1i}, u_{1i})' \sim \mathcal{N}(0, I_2)$ and $z_i \sim \mathcal{U}(-1, 1)$ independently of u. Note that this data generating process does not have smoothing bias. Four models of heteroskedasticity are considered, as given in Table 1 (with $\iota = (1, 1, \cdots, 1)' \in \mathbb{R}^{d_z}$).

Table 1: Simulation Models ()

	$\sigma_v^2(z_i) = 1$	$\sigma_v^2(z_i) = (z_i'\iota)^2$
$\sigma_\varepsilon^2(x_i, z_i) = 1$	Model 1	Model 3
$\sigma_\varepsilon^2(x_i, z_i) = (z_i'\iota + x_i)^2$	Model 2	Model 4

For simplicity, the simulations consider additive-separable power series, that is, the unknown function $g(z_i)$ is assumed to satisfy $g(z_i) = 1 + g_1(z_{1i}) + \cdots + g_{d_z}(z_{d_z i})$ and each component is estimated by $g_j(z_{ji}) \approx p_K(z_{ji})'\gamma_j$, $j = 1, 2, \cdots, d_z$, with $p_K(z_{ji}) = (0, z_{ji}, z_{ji}^2, \cdots, z_{ji}^{K-1})'$.

We consider the classical least squares homoskedasticity-consistent standard errors estimators

$$V_{HO1} = \frac{\hat{\varepsilon}'\hat{\varepsilon}}{n}\hat{\Gamma}^{-} \qquad \text{and} \qquad V_{HO2} = \frac{\hat{\varepsilon}'\hat{\varepsilon}}{n-K-d_x}\hat{\Gamma}^{-},$$

and the classical heteroskedasticity-consistent standard errors estimators

$$V_{HR1} = \hat{\Gamma}^{-}\tilde{X}'\Upsilon\tilde{X}\hat{\Gamma}^{-}, \qquad \Upsilon = \operatorname{diag}\left(\hat{\varepsilon}_1, \cdots, \hat{\varepsilon}_n\right)/n,$$

$$V_{HR2} = \hat{\Gamma}^{-}\tilde{X}'\Upsilon\tilde{X}\hat{\Gamma}^{-}, \qquad \Upsilon = \operatorname{diag}\left(\hat{\varepsilon}_1, \cdots, \hat{\varepsilon}_n\right)/(n-K-d_x).$$

Also, we report the two alternative heteroskedasticity- and many terms- robust standard errors estimators proposed in Theorem 4. These estimators are given by

$$V_{CJN1} = \hat{\Gamma}^{-}\tilde{X}'\Upsilon\tilde{X}\hat{\Gamma}^{-}, \qquad \Upsilon = \operatorname{diag}\left(\hat{\varepsilon}_1, \cdots, \hat{\varepsilon}_n\right)/n,\ K_h = K_{CV},\ K_g = K,$$

$$V_{CJN2} = \hat{\Gamma}^{-}\tilde{X}'\Upsilon\tilde{X}\hat{\Gamma}^{-}, \qquad \Upsilon = \operatorname{diag}\left(\hat{\varepsilon}_1, \cdots, \hat{\varepsilon}_n\right)/n,\ K_h = K,\ K_g = K_{CV},$$

where K_{CV} represents a cross-validation estimate of the optimal K.

The results are presented in Figure 1, for a grid of $K = (0, 1, 2, \cdots, 20)$. The effective degrees of freedom is determined by the choice of K and d_z.

8 Technical Appendix

All statements involving conditional expectations are understood to hold almost surely. Recall that $M = I - Q$ is symmetric and idempotent, and therefore $|M_{ii}| \leq 1$, $n - K = \sum_{i=1}^{n} M_{ii}$ and $M_{ij} = \sum_{\ell=1}^{n} M_{i\ell} M_{\ell j}$.

Proof of Lemma 1. It follows from $H'MH/n = o_p(1)$ and the Cauchy-Schwarz inequality that $X'MX/n = (V+H)' M (V+H)/n = V'MV/n + H'MH/n + 2H'MV/n = V'MV/n + o_p(1)$, provided that

$$V'MV/n = \sum_{i=1}^{n} M_{ii} v_i v_i'/n + \sum_{i=1}^{n} \sum_{j \neq i} M_{ij} v_i v_j'/n = O_p(1).$$

Using $|M_{ii}| \leq 1$ and the Markov inequality,

$$\sum_{i=1}^{n} M_{ii} v_i v_i'/n = \sum_{i=1}^{n} M_{ii} \mathbb{E}\left[v_i v_i' | z_i\right]/n + o_p(1),$$

because

$$\mathbb{V}\left[\sum_{i=1}^{n} M_{ii} \left\|v_i\right\|^2 /n \middle| Z\right] = \sum_{i=1}^{n} M_{ii}^2 \mathbb{V}[\left\|v_i\right\|^2 |z_i]/n^2 = O_{as}\left(n^{-1}\right),$$

while

$$\sum_{i=1}^{n} \sum_{j \neq i} M_{ij} v_i v_j'/n = O_p(n^{-1}K^{1/2}) = o_p(1)$$

by Lemma A1 of Chao, Swanson, Hausman, Newey, and Woutersen (2010). ■

Proof of Lemma 2. First note that

$$X'M(Y - X\beta_0)/\sqrt{n} = V'M\varepsilon/\sqrt{n} + H'M\varepsilon/\sqrt{n} + X'MG/\sqrt{n},$$

where $H'M\varepsilon/\sqrt{n} = O_p\left(K^{-\alpha_h}\right)$ because

$$\mathbb{E}\left[\left\|H'M\varepsilon/\sqrt{n}\right\|^2 \middle| Z\right] = \operatorname{tr}\left(H'M\mathbb{E}[\varepsilon\varepsilon'|Z]MH\right)/n \leq C_\varepsilon \operatorname{tr}\left(H'MH\right)/n = O\left(K_n^{-2\alpha_h}\right).$$

Next,

$$X'Mg/\sqrt{n} = V'MG/\sqrt{n} + \bar{X}'MG/\sqrt{n} = O_p\left(K^{-\alpha_g} + \sqrt{n}K^{-(\alpha_X + \alpha_g)}\right),$$

because

$$\mathbb{E}\left[\left\|V'MG/\sqrt{n}\right\|^2 \middle| Z\right] = G'M\mathbb{E}\left[VV'|Z\right]MG/n \leq C_v G'MG/n = O_{as}\left(K^{-2\alpha_g}\right),$$

and

$$H'MG/\sqrt{n} \leq \sqrt{n}\sqrt{H'MH/n}\sqrt{G'MG/n} = O_{as}\left(\sqrt{n}K^{-(\alpha_X + \alpha_g)}\right),$$

which completes the proof. ■

Proof of Lemma 3. Follows from Lemma A2 in Chao, Swanson, Hausman, Newey, and Woutersen (2010). ■

Proof of Theorem 1. Follows by standard arguments from Lemmas 1–3. ■

Proof of Theorem 2. First, it follows from $G'MG/n = o_p(1)$ and the Cauchy-Schwarz inequality that

$$\begin{aligned}\hat{\varepsilon}'\hat{\varepsilon}/n &= (Y - X\hat{\beta})'M(Y - X\hat{\beta})/n \\ &= (Y - X\hat{\beta} - G)'M(Y - X\hat{\beta} - G)/n + G'MG/n - 2(Y - X\hat{\beta} - G)'MG/n \\ &= (Y - X\hat{\beta} - G)'M(Y - X\hat{\beta} - G)/n + o_p(1),\end{aligned}$$

provided $(Y - X\hat{\beta} - G)'M(Y - X\hat{\beta} - G)/n = O_p(1)$. Next, note that Lemma 1 and $\hat{\beta} - \beta = o_p(1)$ imply $(\hat{\beta} - \beta)'X'MX(\hat{\beta} - \beta)/n = o_p(1)$, which together with the Cauchy-Schwarz inequality gives

$$\begin{aligned}&(Y - X\hat{\beta} - G)'M(Y - X\hat{\beta} - G)/n \\ &= \varepsilon'M\varepsilon/n + (\hat{\beta} - \beta)'X'MX(\hat{\beta} - \beta)/n - 2(Y - X(\hat{\beta} - \beta) - G)'M(\hat{\beta} - \beta)/n \\ &= \varepsilon'M\varepsilon/n + o_p(1) = \tilde{\varepsilon}'\tilde{\varepsilon}/n + o_p(1).\end{aligned}$$

Finally, consider the bilinear form

$$\tilde{\varepsilon}'\tilde{\varepsilon}/n = \varepsilon'M\varepsilon/n = \sum_i M_{ii}\varepsilon_i^2/n + \sum_i \sum_{j \neq i} \varepsilon_i M_{ij} \varepsilon_j/n.$$

Using $|M_{ii}| \leq 1$ and the Markov inequality,

$$\sum_i M_{ii}\varepsilon_i^2/n = \sum_i M_{ii}\mathbb{E}\left[\varepsilon_i^2|z_i\right]/n + o_p(1) = \frac{n-K}{n}\sigma^2 + o_p(1)$$

because

$$\mathbb{E}\left[\mathbb{V}\left[\sum_i M_{ii}\varepsilon_i^2/n \middle| Z\right]\right] = \mathbb{E}\left[\sum_i M_{ii}^2\mathbb{V}[\varepsilon_i^2|z_i]/n^2\right] \leq \sum_i \mathbb{E}\left[\mathbb{V}[\varepsilon_i^2|z_i]\right]/n^2 \leq C_v/n = o(1),$$

while

$$\sum_i \sum_{j \neq i} M_{ij}\varepsilon_i\varepsilon_j/n = O_p\left(n^{-1}K^{1/2}\right) = o_p(1),$$

by Lemma A1 of Chao, Swanson, Hausman, Newey, and Woutersen (2010). Therefore, since $(n-K)/(n-K-d_x) \to 1$, $\hat{\sigma}^2 = \hat{\varepsilon}'\hat{\varepsilon}/(n-K-d) = \sigma^2 + o_p(1)$, which completes the proof. ■

Proof of Theorem 3. Apply Theorem 4 with $K_g = K_h = K$. ■

Proof of Theorem 4. First note that $\hat{\varepsilon} = \tilde{Y} - \tilde{X}'\hat{\beta} = \tilde{\varepsilon} - \tilde{X}'(\hat{\beta} - \beta) + \tilde{g}_0$, for $\tilde{X} = (I - Q_g)X$ and $\tilde{g} = (I - Q_g)g$. Thus, $\hat{\Sigma}_n = T_{1,n} + T_{2,n} + T_{3,n} + T_{4,n} + T_{5,n} + T_{6,n}$, where

$$T_{1,n} = \frac{1}{n}\sum_i w_i\tilde{\varepsilon}_i^2\tilde{v}_i\tilde{v}_i', \qquad T_{2,n} = \frac{1}{n}\sum_i w_i\tilde{\varepsilon}_i^2\tilde{v}_i\left(\tilde{x}_i - \tilde{v}_i\right)', \qquad T_{3,n} = \frac{1}{n}\sum_i w_i\tilde{\varepsilon}_i^2\left(\tilde{x}_i - \tilde{v}_i\right)\tilde{v}_i',$$

$$T_{4,n}=\frac{1}{n}\sum_i w_i\tilde{\varepsilon}_i^2\left(\tilde{x}_i-\tilde{v}_i\right)\left(\tilde{x}_i-\tilde{v}_i\right)', \qquad T_{5,n}=\frac{1}{n}\sum_i w_i\left(\tilde{x}_i'(\hat{\beta}-\beta)+\tilde{g}_i\right)^2\tilde{x}_i\tilde{x}_i',$$

$$T_{6,n}=\frac{1}{n}\sum_i w_i\tilde{\varepsilon}_i\left(\tilde{x}_i'(\hat{\beta}-\beta)+\tilde{g}_i\right)\tilde{x}_i\tilde{x}_i',$$

with $\tilde{X}=(I-Q_h)X=[\tilde{x}_1,\cdots,\tilde{x}_n]'$ and $\tilde{V}=(I-Q_h)V=[\tilde{v}_1,\cdots,\tilde{v}_n]'$.

First consider the leading term $T_{1,n}$. Note that $T_{1,n}=T_{11,n}+T_{12,n}+T_{13,n}$, where

$$T_{1,n}=\frac{1}{n}\sum_i w_i\left(\sum_j M_{g,ij}^2\varepsilon_j^2\right)\left(\sum_k M_{h,ik}^2 v_k v_k'\right),$$

$$T_{2,n}=\frac{1}{n}\sum_i w_i\left(\sum_j M_{g,ij}^2\varepsilon_j^2\right)\left(\sum_{k_1}\sum_{k_2\neq k_1} M_{h,ik_1}M_{h,ik_2}v_{k_1}v_{k_2}'\right),$$

$$T_{3,n}=\frac{1}{n}\sum_i w_i\left(\sum_{j_1}\sum_{j_2\neq j_1} M_{g,ij_1}M_{g,ij_2}\varepsilon_{j_1}\varepsilon_{j_2}\right)\left(\sum_{k_1,k_2} M_{h,ik_1}M_{h,ik_2}v_{k_1}v_{k_2}'\right),$$

with $\mathbb{E}\left[T_{12,n}|Z\right]=0=\mathbb{E}\left[T_{13,n}|Z\right]$. Moreover,

$$T_{11,n}=\mathbb{E}\left[T_{11,n}|\,Z\right]+o_p\left(1\right), \qquad \mathbb{E}\left[T_{11,n}|\,Z\right]=\frac{1}{n}\sum_{i,j}\left(\sum_\ell w_\ell M_{g,\ell i}^2 M_{h,\ell j}^2\right)\mathbb{E}[\varepsilon_i^2 v_j v_j'|z_i,z_j],$$

because $\mathbb{V}[T_{11,n}|Z]=o_{as}(1)$. To see this last result, let $a_{1,ij}=\sum_\ell w_\ell M_{g,\ell i}^2 M_{h,\ell j}^2$ and $A_{1,ij}=\varepsilon_i^2 v_j v_j'-\mathbb{E}[\varepsilon_i^2 v_j v_j'|Z]$, and note after expanding the sums and collecting terms,

$$\mathbb{V}[T_{11,n}|Z]=\frac{1}{n^2}\mathbb{E}\left[\left\|\sum_{i,j}a_{ij}A_{ij}\right\|^2\middle| Z\right]\leq\frac{C}{n^2}\sum_{i,j,k}\left[a_{ij}a_{ik}+a_{ik}a_{ji}+a_{ij}a_{jk}+a_{ik}a_{jk}\right]\leq Cn^{-1},$$

because

$$\begin{aligned}\left|\sum_{i,j,k}a_{ij}a_{ik}\right| &= \left|\sum_{i,j,k}\left(\sum_{\ell_1}w_{\ell_1}M_{g,\ell_1 i}^2 M_{h,\ell_1 j}^2\right)\left(\sum_{\ell_2}w_{\ell_2}M_{g,\ell_2 j}^2 M_{h,\ell_2 k}^2\right)\right| \\ &\leq C\sum_i\sum_{\ell_1}M_{g,\ell_1 i}^2\sum_j M_{h,\ell_1 j}^2\sum_{\ell_2}M_{g,\ell_2 j}^2\sum_k M_{h,\ell_2 k}^2\leq Cn,\end{aligned}$$

and similarly for the other terms.

Next, $T_{12,n}=o_p\left(1\right)$ because $\mathbb{E}\left[T_{2,n}|Z\right]=0$ and $\mathbb{E}[\|T_{2,n}\|^2|Z]\leq Cn^{-1/2}$. To see the last conclusion, let $a_{2,ijk}=\mathbf{1}\left(j\neq k\right)\sum_\ell w_\ell M_{g,\ell i}^2 M_{h,\ell j}M_{h,\ell k}$ and $A_{2,ijk}=\varepsilon_i^2 v_j v_k'$, and observe that $a_{ijk}=a_{ikj}$. Expanding the sums, using the mean-zero property of $A_{2,ijk}$, and collecting terms gives

$$\mathbb{E}[\|T_{2,n}\|^2 |Z] \leq \frac{C}{n^2}\mathbb{E}\left[\left\|\sum_i \sum_{j\neq i} a_{2,iji}\varepsilon_i^2 v_i v_j'\right\|^2 \middle| Z\right] + \frac{C}{n^2}\mathbb{E}\left[\left\|\sum_i \sum_{j\neq i} \sum_{k\neq i,k\neq j} a_{2,ijk}A_{2,ijk}\right\|^2 \middle| Z\right].$$

Now, for the first term, define $a_{2,ij} = \mathbf{1}(i \neq j)\, a_{2,iji}$ and $A_{2,ij} = \varepsilon_i^2 v_i v_j'$, and expanding the sums and collecting non-mean-zero terms

$$\begin{aligned}\frac{1}{n^2}\mathbb{E}\left[\left\|\sum_i \sum_{j\neq i} a_{2,ij}\varepsilon_i^2 v_i v_j'\right\|^2 \middle| Z\right] &\leq \frac{C}{n^2}\sum_i \sum_{j\neq i}[a_{2,ij}^2 + a_{2,ij}a_{2,ji}] + \frac{C}{n^2}\sum_i \sum_{j\neq i} \sum_{k\neq i,k\neq j} a_{2,ik}a_{2,jk} \\ &\leq Cn^{-1} + Cn^{-1/2},\end{aligned}$$

while for the second term

$$\begin{aligned}\frac{1}{n^2}\mathbb{E}\left[\left\|\sum_i \sum_{j\neq i} \sum_{k\neq i,k\neq j} a_{2,ijk}A_{2,ijk}\right\|^2 \middle| Z\right] &\leq \frac{C}{n^2}\sum_i \sum_{j\neq i} \sum_{k\neq i,k\neq j}[a_{ijk}^2 + a_{ijk}a_{jik}] + \frac{C}{n^2}\sum_i \sum_{j\neq i} \sum_{k\neq i,k\neq j} \sum_{l\neq i,l\neq j,l\neq k} a_{ijk}a_{ljk} \\ &\leq Cn^{-1/2} + Cn^{-1},\end{aligned}$$

where the final bounds as a function of n are obtained by properties of the idempotent matrix M as above.

Finally, $T_{3,n} = o_p(1)$ because $\mathbb{E}[T_{3,n}|Z] = 0$ and $\mathbb{E}[\|T_{3,n}\|^2 |Z] \leq Cn^{-1}$. To see the last conclusion, first let $a_{3,ij} = \varepsilon_i\varepsilon_j$ and $A_{3,ij} = \mathbf{1}(i \neq j)\sum_\ell w_\ell M_{g,\ell i}M_{g,\ell j}\left(\sum_{k_1}\sum_{k_2} M_{h,\ell k_1}M_{h,\ell k_2}v_{k_1}v_{k_2}'\right)$, and observe that $a_{3,ij} = a_{3,ji}$ and $A_{3,ij} = A_{3,ji} = A_{3,ij}' = A_{3,ji}'$. Then, expanding the sums and collecting non-mean-zero terms,

$$\begin{aligned}\mathbb{E}[\|T_{3,n}\|^2 |Z] &= \mathbb{E}\left[\left\|\frac{1}{n}\sum_i \sum_j a_{3,ij}A_{3,ij}\right\|^2 \middle| Z\right] = \frac{2}{n^2}\sum_i \sum_j \mathbb{E}[a_{3,ij}^2 \|A_{3,ij}\|^2 |Z] \\ &\leq \frac{C}{n^2}\sum_i \sum_{j\neq i}\mathbb{E}\left[\left\|\sum_\ell w_\ell M_{g,\ell i}M_{g,\ell j}\left(\sum_{k_1}\sum_{k_2} M_{h,\ell k_1}M_{h,\ell k_2}v_{k_1}v_{k_2}'\right)\right\|^2 \middle| Z\right].\end{aligned}$$

Next, for each (i,j) let $a_{3,ijkl} = \sum_\ell w_\ell M_{g,\ell i}M_{g,\ell j}M_{h,\ell k}M_{h,\ell l}$ and $A_{3,kl} = v_k v_l'$, and note that $a_{3,ijkl} =$

$a_{3,ijlk}$ and $A_{3,kl} = A'_{3,lk}$. Thus, expanding the sums and collecting non-mean-zero terms,

$$
\begin{aligned}
\mathbb{E}[\|T_{3,n}\|^2 |Z] &\leq \frac{C}{n^2}\sum_i \sum_{j\neq i} \mathbb{E}\left[\left\| \sum_{k=1}^n \sum_{l=1}^n a_{3,ijkl} A_{3,kl} \right\|^2 \middle| Z \right] \\
&\leq \frac{C}{n^2}\sum_i \sum_{j\neq i} \mathbb{E}\left[\left\| \sum_{k} a_{3,ijkk} A_{3,kk} \right\|^2 \middle| Z \right] + \frac{C}{n^2}\sum_i \sum_{j\neq i} \mathbb{E}\left[\left\| \sum_{k} \sum_{l\neq k} a_{3,ijkl} A_{3,kl} \right\|^2 \middle| Z \right] \\
&\leq \frac{C}{n^2}\sum_i \sum_{j\neq i} \sum_k a_{ijkk}^2 + \frac{C}{n^2}\sum_i \sum_{j\neq i} \sum_k \sum_{l\neq k} a_{ijkl}^2 \leq Cn^{-1},
\end{aligned}
$$

where the last inequality follows by properties of the idempotent matrix M as above.

Next, consider the smaller order terms $T_{k,n}$, $k = 2, \cdots, 6$. By $\tilde{\varepsilon}_i = M_{g,ii}\varepsilon_i + \sum_{j=1,j\neq i}^n M_{g,ij}\varepsilon_j$ with $\mathbb{E}\left[\varepsilon_i | z_i\right] = 0$,

$$
\mathbb{E}\left[\tilde{\varepsilon}_i^4 \middle| Z\right] \leq C \sum_j \sum_k M_{g,ij}^2 M_{g,ik}^2 \mathbb{E}[\varepsilon_j^2 \varepsilon_k^2 | z_j, z_k] \leq C,
$$

by properties of the idempotent matrix M. Similarly, $\mathbb{E}[\|\tilde{v}_i\|^4 |Z] \leq C$. For terms $T_{2,n}$ and $T_{3,n}$, using $\tilde{x}_i - \tilde{v}_i = M'_{h,i}H = M'_{h,i}(H - P'_{K_h}\eta_h)$, it follows that for $k = 2, 3$,

$$
\begin{aligned}
\mathbb{E}\left[\|T_{k,n}\| | Z\right] &\leq \frac{1}{n}\sum_i |w_i| \, \mathbb{E}[\tilde{\varepsilon}_i^2 \|\tilde{V}_i\| |Z] \left\| M'_{h,i}(H - P'_{K_h}\eta_h) \right\| \\
&\leq \frac{C}{n}\sum_i \sum_k |M_{h,ik}| \, \|h(z_k) - p_{K_h}(z_k)'\eta_h\| \leq C K_h^{1/2} K_h^{-\alpha_h},
\end{aligned}
$$

by Cauchy-Schwarz inequality, and therefore $T_{2,n} = o_p(1)$ and $T_{3,n} = o_p(1)$. Also, by the same arguments,

$$
\begin{aligned}
\mathbb{E}\left[\|T_{4,n}\| | Z\right] &\leq \frac{C}{n}\sum_i \mathbb{E}\left[\tilde{\varepsilon}_i^2 \middle| Z\right] \left\| M'_{h,i}(H - P'_{K_h}\eta_h) \right\|^2 \\
&\leq \frac{C}{n}\sum_i M_{h,ii} \sum_k \|h(z_k) - p_{K_h}(z_k)'\eta_h\|^2 \leq C K_h K_h^{-2\alpha_h}
\end{aligned}
$$

which implies that $T_{4,n} = o_p(1)$.

Therefore, it remains to show that $T_{5,n} = o_p(1)$, which then implies by the Cauchy-Schwarz inequality that $T_{6,n} = o_p(1)$. To this end, note that

$$
T_{5,n} = \frac{1}{n}\sum_i w_i \left(\tilde{x}'_i(\hat{\beta} - \beta) + \tilde{g}_i\right)^2 \tilde{x}_i \tilde{x}'_i = O_p(n^{-1} + K_h K_h^{-4\alpha_h} + K_g^{1/2} K_g^{-2\alpha_g} + n K_g^{1/2} K_g^{-2\alpha_g} K_h^{1/2} K_h^{-2\alpha_h}),
$$

because, using $\hat{\beta} - \beta_0 = O_p\left(n^{-1/2}\right)$, $\mathbb{E}[\|\tilde{v}_i\|^4|Z] \leq C$,

$$\begin{aligned}
\frac{1}{n}\sum_i |w_i| \, \|\tilde{x}_i'(\hat{\beta} - \beta_0)\|^2 \|\tilde{x}_i\|^2 &\leq C\|\hat{\beta} - \beta_0\|^2 \frac{1}{n}\sum_i \|\tilde{v}_i\|^4 + C\|\hat{\beta} - \beta_0\|^2 \frac{1}{n}\sum_i \|M_{h,i}'(H - P_{K_h}'\eta_h)\|^4 \\
&\leq O_p(n^{-1}) + O_p(1)\frac{1}{n^2}\sum_i M_{h,ii}^2 \left(\sum_k \|h(z_k) - p_{K_h}(z_k)'\eta_h\|^2\right)^2 \\
&= O_p(n^{-1} + K_h K_h^{-4\alpha_h}),
\end{aligned}$$

and

$$\begin{aligned}
\frac{1}{n}\sum_i |w_i| \, |\tilde{g}_i|^2 \|\tilde{x}_i\|^2 &\leq C\sqrt{\sum_i M_{g,ii}^2 \left(\sum_k \|g(z_k) - p_{K_g}(z_k)'\eta_g\|^2\right)^2} \sqrt{\frac{1}{n^2}\sum_i \|\tilde{x}_i\|^4} \\
&\leq C\sqrt{K_g}\left(\sum_k \|g(z_k) - p_{K_g}(z_k)'\eta_g\|^2\right) O_p(n^{-1} + K_h^{1/2} K_h^{-2\alpha_h}) \\
&= O_p(K_g^{1/2} K_g^{-2\alpha_g} + n K_g^{1/2} K_g^{-2\alpha_g} K_h^{1/2} K_h^{-2\alpha_h}).
\end{aligned}$$

This concludes the proof. ■

References

Angrist, J., G. W. Imbens, and A. Krueger (1999): "Jackknife Instrumental Variables Estimation," *Journal of Applied Econometrics*, 14(1), 57–67.

Bekker, P. A. (1994): "Alternative Approximations to the Distributions of Instrumental Variables Estimators," *Econometrica*, 62, 657–681.

Cattaneo, M. D., R. K. Crump, and M. Jansson (2010): "Small Bandwidth Asymptotics for Density-Weighted Average Derivatives," working paper.

Chao, J. C., N. R. Swanson, J. A. Hausman, W. K. Newey, and T. Woutersen (2010): "Asymptotic Distribution of JIVE in a Heteroskedastic IV Regression with Many Instruments," forthcoming in *Econometric Theory*.

Donald, S. G., and W. K. Newey (1994): "Series Estimation of Semilinear Models," *Journal of Multivariate Analysis*, 50(1), 30–40.

Hansen, C., J. Hausman, and W. K. Newey (2008): "Estimation with Many Instrumental Variables," *Journal of Business and Economic Statistics*, 26(4), 398–422.

Kunitomo, N. (1980): "Asymptotic Expansions of the Distributions of Estimators in a Linear Functional Relationship and Simultaneous Equations," *Journal of the American Statistical Association*, 75(371), 693–7000.

Linton, O. (1995): "Second Order Approximation in the Partialy Linear Regression Model," *Econometrica*, 63(5), 1079–1112.

Morimune, K. (1983): "Approximate Distributions of k-Class Estimators when the Degree of Overidentifiability is Large Compared with the Sample Size," *Econometrica*, 51(3), 821–841.

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

Powell, J. L., J. H. Stock, and T. M. Stoker (1989): "Semiparametric Estimation of Index Coefficients," *Econometrica*, 57(6), 1403–1430.

Robinson, P. M. (1988): "Root-N-Consistent Semiparametric Regression," *Econometrica*, 56(4), 931–954.

van der Vaart, A. W. (1998): *Asymptotic Statistics*. Cambridge University Press, New York.

White, H. (1980): "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," *Econometrica*, 48(4), 817–838.

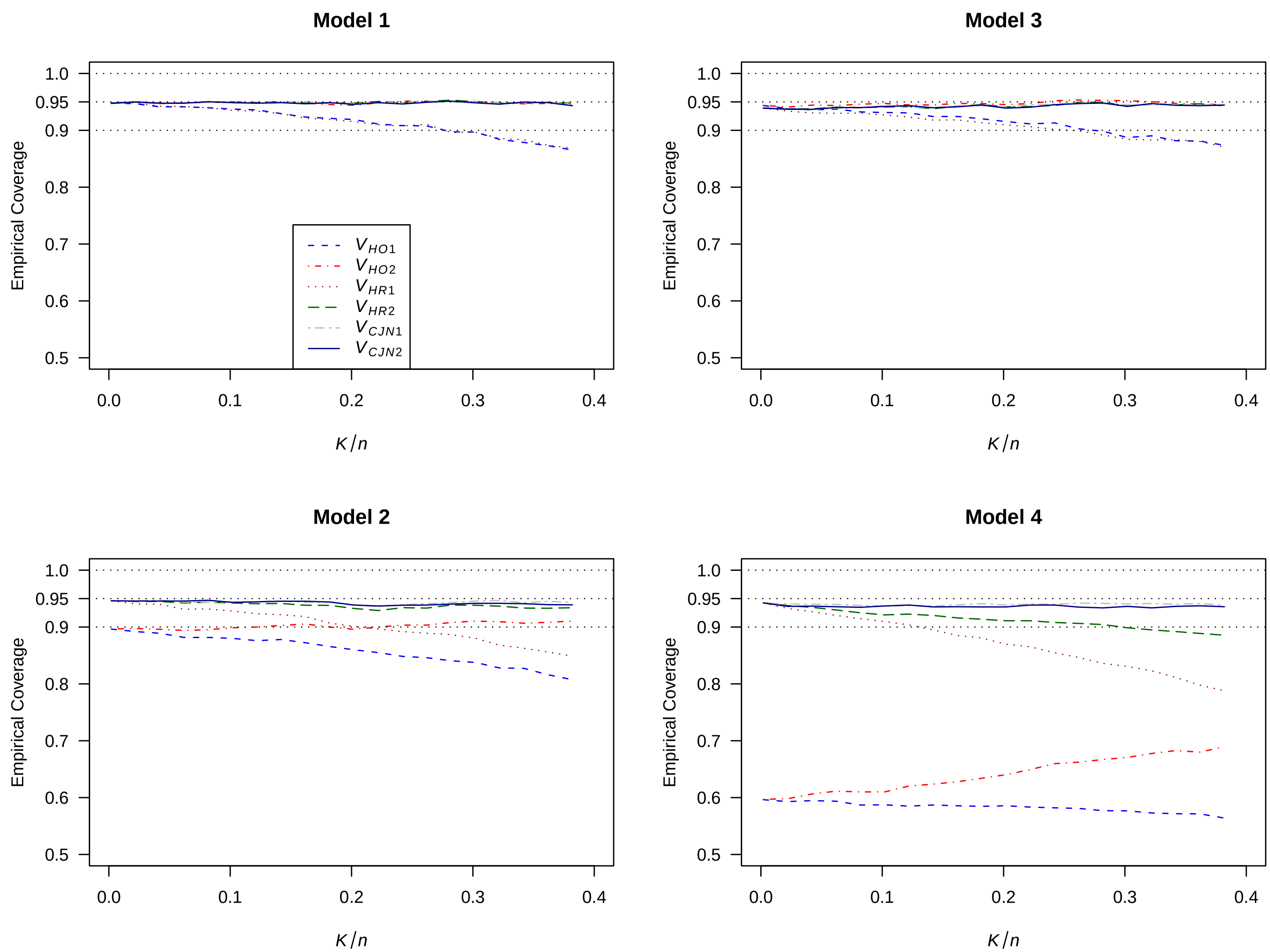

Figure 1: Empirical Coverage Rates for 95% Confidence Intervals: $n = 500$, $S = 3,000$