CrossMark

Econometric Theory, **34**, 2018, 277–301.
doi:10.1017/S026646661600013X

ALTERNATIVE ASYMPTOTICS AND THE PARTIALLY LINEAR MODEL WITH MANY REGRESSORS

MATIAS D. CATTANEO
University of Michigan

MICHAEL JANSSON
University of California Berkeley and CREATES

WHITNEY K. NEWEY
Massachusetts Institute of Technology

Many empirical studies estimate the structural effect of some variable on an outcome of interest while allowing for many covariates. We present inference methods that account for many covariates. The methods are based on asymptotics where the number of covariates grows as fast as the sample size. We find a limiting normal distribution with variance that is larger than the standard one. We also find that with homoskedasticity this larger variance can be accounted for by using degrees-of-freedom-adjusted standard errors. We link this asymptotic theory to previous results for many instruments and for small bandwidth(s) distributional approximations.

1. INTRODUCTION

Many empirical studies estimate the structural, causal, or treatment effect of some variable on an outcome of interest. For example, we might be interested in estimating the effect of some government policy on an outcome such as income. Since policies and many other variables are not exogenous, researchers rely on a variety of approaches based on observational data when trying to estimate such effects. One important method is based on assuming that the variable of interest can be taken as exogenous after controlling for a sufficient set of other factors or covariates. See, for example, Heckman and Vytlacil (2007) and Imbens and Wooldridge (2009) for recent reviews and further references.

The authors are grateful for comments from Xiaohong Chen, Victor Chernozhukov, Alfonso Flores–Lagunes, Lutz Kilian, seminar participants at Bristol, Brown, Cambridge, Exeter, Indiana, LSE, Michigan, MSU, NYU, Princeton, Rutgers, Stanford, UCL, UCLA, UCSD, UC-Irvine, USC, Warwick and Yale, and conference participants at the 2010 Joint Statistical Meetings and the 2010 LACEA Impact Evaluation Network Conference. We also thank the editor, Peter Phillips, the co-editor and two reviewers for their comments. The first author gratefully acknowledges financial support from the National Science Foundation (SES 1122994 and SES 1459931). The second author gratefully acknowledges financial support from the National Science Foundation (SES 1124174 and SES 1459967) and the research support of CREATES (funded by the Danish National Research Foundation under Grant No. DNRF78). The third author gratefully acknowledges financial support from the National Science Foundation (SES 1132399). Address correspondence to Whitney K. Newey, Department of Economics, MIT, E52-424, Cambridge, MA 02139, USA; e-mail: wnewey@mit.edu.

© Cambridge University Press 2016

https://doi.org/10.1017/S026646661600013X Published online by Cambridge University Press

A problem empirical researchers face when relying on covariates to estimate a structural effect is the availability of many potential controls. Typically, intuition will suggest a set of variables that might be important but will not identify exactly which variables are important or the functional form with which variables should enter the model. This lack of clear guidance about what variables to use leaves researchers with a potentially vast set of potential covariates including raw regressors available in the data as well as interactions and other nonlinear transformations thereof. Many economic studies include many of these variables in order to control for as broad array of covariates as possible. For example, it is common to include dummy variables for many potentially overlapping groups based on age, cohort, geographic location, etc. Even when some controls are dropped after valid covariate selection, as was developed by Belloni, Chernozhukov, and Hansen (2014), many controls may remain in the final regression specification.

We present inference methods that account for the presence of many controls in regression models. We do this using a large sample approximation where the number of covariates grows as fast as the sample size. We find a limiting normal distribution with variance that is larger than the standard asymptotic variance. We show that with homoskedasticity this larger variance is fully accounted for by using standard errors with a degrees-of-freedom adjustment for inclusion of many covariates. This asymptotics and the associated standard errors provides an important justification for the practice of adjusting for degrees of freedom even when disturbances are not normally distributed. As always the asymptotics are meant as an approximation that provides useful inference methods for applications. In this way the asymptotic approximation given here should prove useful in practice.

This paper also adds to the literature on regression where the number of regressors grow with the sample size. Huber (1973) showed that fitted regression values are not asymptotically normal when the number of regressors grows as fast as sample size. The problem is circumvented here by focusing on the coefficients of some regressors when the number of covariates gets large. Recently, El Karoui, Bean, Bickel, Lim, and Yu (2013) showed that, with a Gaussian distributional assumption on the regressors, certain coefficients and contrasts are asymptotically normal when the number of regressors grows as fast as sample size, but do not give inference results. We do give inference results in showing that the degrees-of-freedom adjustment to standard errors accounts correctly for many covariates and do not impose distributional assumptions on the regressors. We also use a different and simpler approach to the asymptotic theory. We note that our results were presented at the 2010 Joint Statistical Meetings and are independent of El Karoui et al. (2013).

The asymptotics here are based on asymptotic normality results for degenerate U-statistics. To help explain and motivate this theory we note that asymptotic normality for degenerate U-statistics has already been used in other settings. Such results are the basis for the many instrument asymptotics where the number of instruments grows as fast as the sample size. Kunitomo (1980) and Morimune (1983) derived asymptotic variances that are larger than the usual formulae

https://doi.org/10.1017/S026646661600013X Published online by Cambridge University Press

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 the use of many instrument standard errors provides an improvement for a range of number of instruments. Such asymptotics have also proven useful for small bandwidth approximations for kernel-based density-weighted average derivative estimators in Cattaneo, Crump, and Jansson (2010, 2014b). They show 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.

We use a common framework for these results to motivate the asymptotic theory. The common framework is that the object determining the limiting distribution is a V-statistic, which can be decomposed into a bias term, a sample average, and a "remainder" that is an asymptotically normal degenerate U-statistic. Asymptotic normality of the remainder distinguishes this setting from others with degenerate U-statistic. Here asymptotic normality occurs because the number of covariates goes to infinity, while the behavior of a degenerate U-statistic is different in other settings. When the number of covariates 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 covariate, many instrument, and small bandwidth results share this structure. In keeping with this common structure, we refer here to such results under the general heading of "alternative asymptotics." While not all semiparametric estimation problems share this structure, we show by example that its scope may indeed be useful for econometrics. In the conclusions section below we also discuss its limitations and its relation to other types of alternative asymptotic approximations in semiparametrics problems and other loosely related contexts.

An important generalization to the results presented herein is to asymptotics and inference with many covariates under heteroskedasticity. Constructing consistent standard error estimators under heteroskedasticity of unknown form in this setting turns out to be quite challenging. In Cattaneo, Jansson, and Newey (2015), we present a detailed discussion of heteroskedasticity-robust standard errors for linear models where the number of covariates increases at the same rate as the sample size, which covers the partially linear model when the number terms grows at the same rate as the sample size.

The rest of the paper is organized as follows. Section 2 describes the common structure of many instrument and small bandwidth asymptotics, and also shows how the structure leads to new results for the partially linear model. Section 3 formalizes the new distributional approximation for many covariates. Section 4 reports results from a small simulation study aimed to illustrate our results in small samples. Section 5 concludes. Appendix A collects the proofs of our results, while Appendix B discusses heuristically how our results can be extended to the case of generated regressors and related problems.

https://doi.org/10.1017/S026646661600013X Published online by Cambridge University Press

2. A COMMON STRUCTURE

We consider inference on structural effects in an environment where variables of interest may be taken as exogenous conditional on covariates. We pose the problem in the framework of a partially linear model. Let $(y_i, x_i', z_i')', i = 1, \ldots, n$, be a random sample satisfying

$$y_i = x_i'\beta_0 + g(z_i) + \varepsilon_i, \qquad \mathbb{E}[\varepsilon_i | x_i, z_i] = 0, \tag{1}$$

where y_i is a scalar dependent variable, $x_i \in \mathbb{R}^d$ are the treatment/policy variables of interest, z_i are explanatory variables, $g(z)$ is an unknown function, and $\mathbb{E}[\mathbb{V}[x_i|z_i]]$ is of full rank. The goal of the analysis is to conduct inference about the structural effect β_0.

A series estimator of β_0 is obtained by regressing y_i on x_i and functions of z_i. To describe the estimator, let $p^1(z)$, $p^2(z)$, ... be approximating functions, such as polynomials or splines, and let $p_K(z) = (p^1(z), \ldots, p^K(z))'$ be a K-dimensional vector of such functions. We consider a regression that includes a $K \times 1$ vector of covariates $p_K(z_i)$ that may consist of z_i and transformations of z_i to adequately approximate $g(z_i)$. The conditional mean restriction $\mathbb{E}[\varepsilon_i | x_i, z_i] = 0$ means that x_i may be considered exogenous after controlling linearly for variables that can approximate $g(z_i)$. We will assume that linear combinations of these variables provide approximations to $g(z_i)$ and to $\mathbb{E}[x_i|z_i]$ with relatively small approximation errors for each object. To describe the estimator let M_{ij} denote the (i, j)-th element of $M = I_n - P_K(P_K'P_K)^{-1}P_K'$, where $P_K = [p_K(z_1), \ldots, p_K(z_n)]'$. A series estimator of β_0 in (1) is given by

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

Donald and Newey (1994) gave conditions for asymptotic normality of this estimator using standard asymptotics. See also Linton (1995) and references therein for related asymptotic results when using kernel estimators.

Conditional on $Z = [z_1, \ldots, z_n]'$, $\hat{\beta}$ depends on a V-statistic. Plugging in for y_i for each i and solving gives

$$\sqrt{n}\left(\hat{\beta} - \beta_0\right) = \hat{\Gamma}_n^{-1} S_n, \tag{2}$$

with

$$\hat{\Gamma}_n = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} M_{ij} x_i x_j', \qquad S_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \sum_{j=1}^{n} x_i M_{ij} \left(g_j + \varepsilon_j\right),$$

where $g_i = g(z_i)$. Conditional on Z, the term S_n is a V-statistic

$$S_n = \sum_{i=1}^{n} \sum_{j=1}^{n} u_{ij}^n \left(W_i, W_j\right),$$

https://doi.org/10.1017/S026646661600013X Published online by Cambridge University Press

where $W_i = (x_i', \varepsilon_i)'$ and $u_{ij}^n(W_i, W_j) = x_i M_{ij}(g_j + \varepsilon_j)/\sqrt{n}$. We assume throughout this section that there exists a sequence of nonrandom matrices Γ_n satisfying $\Gamma_n^{-1}\hat{\Gamma}_n \to_p I_d$ for I_d the $d \times d$ identity matrix, and hence we focus on the V-statistic S_n. (All limits are taken as $n \to \infty$ unless explicitly stated otherwise.)

To explain the many covariate asymptotics, and to provide a link to previous work on many instruments and small bandwidths, it is helpful to provide a general analysis of the V-statistic S_n. This V-statistic has a well known (Hoeffding-type) decomposition that we describe here because it is an essential feature of the common structure. For notational simplicity we will drop the W_i and W_j arguments and set $u_{ij}^n = u_{ij}^n(W_i, W_j)$ and $\tilde{u}_{ij}^n = u_{ij}^n + u_{ji}^n - \mathbb{E}[u_{ij}^n + u_{ji}^n]$. Let $\|\cdot\|$ denote the Euclidean norm. If $\mathbb{E}[\|u_{ij}^n\|] < \infty$ for all i, j, n, then

$$S_n = B_n + \Psi_n + U_n, \tag{3}$$

where

$$B_n = \mathbb{E}[S_n], \qquad \Psi_n = \sum_{i=1}^n \psi_i^n(W_i), \qquad U_n = \sum_{i=2}^n D_i^n(W_i, \ldots, W_1),$$

$$\psi_i^n(W_i) = u_{ii}^n - \mathbb{E}[u_{ii}^n] + \sum_{j=1, j\neq i}^n \mathbb{E}\left[\tilde{u}_{ij}^n \middle| W_i\right],$$

$$D_i^n(W_i, \ldots, W_1) = \sum_{j=1, j<i}^n \left(\tilde{u}_{ij}^n - \mathbb{E}\left[\tilde{u}_{ij}^n \middle| W_i\right] - \mathbb{E}\left[\tilde{u}_{ij}^n \middle| W_j\right]\right).$$

It is straightforward to see that $\mathbb{E}[\psi_i^n(W_i)] = 0$, $\mathbb{E}[D_i^n(W_i, \ldots, W_1)| W_{i-1}, \ldots, W_1] = 0$, and $\mathbb{E}[\Psi_n U_n] = 0$. This decomposition of a V-statistic is well known (e.g., van der Vaart (1998, Chapter 11)), and shows that S_n can be decomposed into a sum Ψ_n of independent terms, a U-statistic remainder U_n that is a martingale difference sum and uncorrelated with Ψ_n, and a pure bias term B_n.[1] The decomposition 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 under a particular restriction on the tuning parameter sequence (e.g., when the number of covariates increases slowly enough).

An interesting property of U_n is that it is asymptotically normal at some rate when the number of covariates grows. To be specific, if regularity conditions specified below hold and $K \to \infty$ with the sample size, it turns out that

$$\begin{bmatrix} \mathbb{V}[\Psi_n]^{-1/2}\Psi_n \\ \mathbb{V}[U_n]^{-1/2}U_n \end{bmatrix} \to_d \mathcal{N}(0, I_{2d}).$$

In other settings, where the underlying kernel of the U-statistic does not vary with the sample size, the asymptotic behavior of U_n can be different. Many degenerate U-statistics will converge to a weighted sum of independent chi-squared random

https://doi.org/10.1017/S026646661600013X Published online by Cambridge University Press

variables (e.g., van der Vaart (1998, Chapter 12)). However, as the number of covariates grows, the kernel of the underlying U-statistic forming U_n changes with the sample in such a way that the individual contributions $D_i^n(W_i, \ldots, W_1)$ to U_n are small enough to satisfy a Lindeberg–Feller condition leading to a Gaussian limiting distribution (usually established using the martingale property of U_n). For an interesting discussion of this phenomenon, see de Jong (1987). This type of asymptotic normality result for degenerate U-statistics has previously been shown in other settings, as further explained below.

When the number of covariates grows as fast as the sample size $\mathbb{V}[\Psi_n]$ and $\mathbb{V}[U_n]$ 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 $\lim_{n\to\infty} \mathbb{V}[\Psi_n]$ (assuming the limit exists). As a consequence, consistent variance estimation under many covariate asymptotics requires accounting for the contribution of U_n to the (asymptotic) sampling variability of the statistic.

To apply this calculation to many covariates, note that by $\mathbb{E}[\varepsilon_i|x_i, z_i] = 0$ we have $\mathbb{E}[x_i\varepsilon_i|Z] = 0$. Therefore, for $u_{ij}^n = u_{ij}^n(W_i, W_j)$ as introduced previously, we have

$$\mathbb{E}\left[u_{ij}^n \middle| Z\right] = h_i M_{ij} g_j/\sqrt{n}, \qquad u_{ij}^n - \mathbb{E}\left[u_{ij}^n \middle| Z\right] = M_{ij}\left(v_i g_j + x_i\varepsilon_j\right)/\sqrt{n},$$

$$\tilde{u}_{ij}^n = M_{ij}\left(v_j g_i + v_i g_j + x_j\varepsilon_i + x_i\varepsilon_j\right)/\sqrt{n},$$

$$\mathbb{E}\left[\tilde{u}_{ij}^n \middle| W_i, Z\right] = M_{ij}\left(v_i g_j + h_j\varepsilon_i\right)/\sqrt{n},$$

for $i \neq j$, where $h_i = h(z_i) = \mathbb{E}[x_i|z_i]$ and $v_i = x_i - h_i$. In this case, the bias term in (3) is

$$B_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \sum_{j=1}^{n} M_{ij} h_i g_j,$$

which will be negligible under regularity conditions, as shown in the next section. Moreover,

$$\Psi_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} M_{ii} v_i \varepsilon_i + R_n, \qquad R_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \sum_{j=1}^{n} M_{ij}\left(v_i g_j + h_i \varepsilon_j\right),$$

where R_n has mean zero and converges to zero in mean square as K grows, as further discussed below. Under standard asymptotics M_{ii} will go to one and hence the limiting variance of the leading term in Ψ_n corresponds to the usual asymptotic variance. Finally, we find that the degenerate U-statistic term is

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

where Q_{ij} is the (i, j)-th component of $P_K(P_K'P_K)^{-1}P_K'$. Remarkably, as discussed below, this term is essentially the same as the degenerate U-statistic term for certain instrumental variables estimators. Consequently, a central limit

https://doi.org/10.1017/S026646661600013X Published online by Cambridge University Press

theorem of Chao, Swanson, Hausman, Newey, and Woutersen (2012) that was applied to many instrument asympotics is applicable to regression with many covariates. We will employ it to show that U_n is asymptotically normal as $K \to \infty$.

Distribution theory with many covariates may be seen as a generalization of the conventional asymptotics in the sense that under conventional asymptotics the asymptotic variances emerging from both approaches coincide. But, the alternative asymptotic approximation also allows for the covariates to grow at the same rate as the sample size, where the limiting asymptotic variance is larger. Thus, in general, there is no reason to expect that the usual standard error formulas derived under conventional asymptotics will remain valid more generally. From this perspective, our many covariate asymptotics provides a theoretical justification for new standard error formulas that are consistent under both conventional and alternative asymptotics. We refer to the latter standard error formulas as being more robust than the usual standard errors available in the literature. For instance, using these ideas, more robust standard errors were derived previously for many instrument asymptotics in IV models (Hansen, Hausman, and Newey (2008)) and small bandwidth asymptotics in kernel-based semiparametrics (Cattaneo, Crump, and Jansson (2014b)).

Accounting for the presence of U_n should also yield improvements when numbers of covariates do not satisfy the knife-edge condition of growing at the same rate as the sample size. For instance, if the number of covariates grows just slightly slower than the sample size 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.

To motivate and provide background for this approach we show next that both many instrument asymptotics and small bandwidth asymptotics have the structure described above.

2.1. Connection with Many Instrument Asymptotics

To link many covariate asymptotics with many instrument asymptotics we focus on the JIVE2 estimator of Angrist, Imbens, and Krueger (1999), but the idea applies to other IV estimators such as the limited information maximum likelihood estimator. See Chao et al. (2012) for more details, including regularity conditions under which the following discussion can be made rigorous. See also Alvarez and Arellano (2003) for a discussion of many instrument IV asymptotics for panels.

Let $(y_i, x_i', z_i')'$, $i = 1, \ldots, n$, be a random sample generated by the model

$$y_i = x_i'\beta_0 + \varepsilon_i, \qquad \mathbb{E}[\varepsilon_i | z_i] = 0, \tag{4}$$

where y_i is a scalar dependent variable, $x_i \in \mathbb{R}^d$ is a vector of endogenous variables, ε_i is a disturbance, and $z_i \in \mathbb{R}^K$ is a vector of instrumental variables.

To describe the JIVE2 estimator of β_0 in (4), now let Q_{ij} denote the (i, j)-th element of $Q = Z(Z'Z)^{-1}Z'$, where $Z = [z_1, \cdots, z_n]'$. After centering and

https://doi.org/10.1017/S026646661600013X Published online by Cambridge University Press

scaling, the JIVE2 estimator $\hat{\beta}$ satisfies

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

Conditional on Z, $\hat{\beta}$ has the structure in (2) with $W_i=(x_i',\varepsilon_i)'$ and

$$\hat{\Gamma}_n=\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}Q_{ij}x_ix_j',\qquad u_{ij}^n\left(W_i,W_j\right)=\mathbf{1}(i\neq j)Q_{ij}x_i\varepsilon_j/\sqrt{n},$$

where $\mathbf{1}(\cdot)$ is the indicator function.

For $i\neq j$, $\mathbb{E}\left[u_{ij}^n(W_i,W_j)\middle|Z\right]=0$ and

$$\mathbb{E}\left[u_{ij}^n(W_i,W_j)\middle|W_i,Z\right]=Q_{ij}x_i\mathbb{E}\left[\varepsilon_j|Z\right]=0,$$
$$\mathbb{E}\left[u_{ji}^n(W_j,W_i)\middle|W_i,Z\right]=Q_{ij}\Upsilon_j\varepsilon_i/\sqrt{n},$$

where $\Upsilon_i=\mathbb{E}[x_i|z_i]$ can be interpreted as the reduced form for observation i. As a consequence, (3) is satisfied with $B_n=0$,

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

$$D_i^n(W_i,\ldots,W_1)=\sum_{j=1,j<i}^{n}Q_{ij}\left(\upsilon_i\varepsilon_j+\upsilon_j\varepsilon_i\right)/\sqrt{n},\qquad \upsilon_i=x_i-\Upsilon_i.$$

Because $\Upsilon_i-\sum_{j=1}^{n}Q_{ij}\Upsilon_j$ is the i-th residual from regressing the reduced form observations on Z, by appropriate definition of the reduced form this can generally be assumed to vanish as the sample size grows. In that case,

$$\Psi_n=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Upsilon_i(1-Q_{ii})\varepsilon_i+o_p(1).$$

Furthermore, under standard asymptotics Q_{ii} will go to zero, so the limiting variance of the leading term in Ψ_n corresponds to the usual asymptotic variance for IV. The degenerate U-statistic term is

$$U_n=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\sum_{j=1,j<i}^{n}Q_{ij}\left(\upsilon_i\varepsilon_j+\upsilon_j\varepsilon_i\right).$$

Chao et al. (2012) apply a martingale central limit theorem to show that this U_n will be asymptotically normal when $K\to\infty$ and certain regularity conditions hold. Here we see that the U_n term for JIVE2 has the same form as for many covariates. Thus, many covariate asymptotics can be obtained by using previous results for many instruments.

https://doi.org/10.1017/S026646661600013X Published online by Cambridge University Press

2.2. Connection with Small Bandwidth Asymptotics

We can also show that small bandwidth asymptotics for certain kernel-based semiparametric estimators are based on a degenerate U-statistic like that considered above. To keep the exposition simple we focus on an estimator of the integrated squared density, but the structure of this estimator is shared by the density-weighted average derivative estimator of Powell, Stock, and Stoker (1989) treated in Cattaneo, Crump, and Jansson (2014b) and more generally by estimators of density-weighted averages and ratios thereof (see, e.g., Newey, Hsieh, and Robins (2004, Section 2) and references therein). Furthermore, these ideas are also applicable to other semiparametric problems such as those involving (*i*) certain functionals of U-processes arising in latent models as in Aradillas–Lopéz, Honoré, and Powell (2007) and references therein, (*ii*) U-statistics used for specification testing as in Li and Racine (2007, Chapter 12) and references therein, and (*iii*) U-statistics obtained from convolution estimators as in Schick and Wefelmeyer (2013) and references therein. Since the main purpose here is to highlight the connections between many covariate asymptotics and other alternative asymptotics in the literature, rather than to extend the scope of alternative asymptotics, we do not discuss those other potential applications here.

Suppose x_i, $i = 1, \ldots, n$, are i.i.d. continuously distributed p-dimensional random vectors with smooth p.d.f. f_0 and consider estimation of the integrated squared density

$$\beta_0 = \int_{\mathbb{R}^p} f_0(x)^2 \mathrm{d}x = \mathbb{E}[f_0(x_i)].$$

A leave-one-out kernel-based estimator is

$$\hat{\beta} = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} \mathcal{K}_h(x_i - x_j),$$

where $\mathcal{K}(u)$ is a symmetric kernel and $\mathcal{K}_h(u) = h^{-p}\mathcal{K}(u/h)$. As shown by Giné and Nickl (2008), this estimator is optimal, attaining root-n consistency under weak conditions. This estimator has the V-statistic form of (2) with $W_i = x_i$ and

$$\hat{\Gamma}_n = 1, \qquad u_{ij}^n(W_i, W_j) = \frac{1}{\sqrt{n}(n-1)} \mathbb{1}(i \neq j) \left\{ \mathcal{K}_h\left(x_i - x_j\right) - \beta_0 \right\}.$$

Let $f_h(x) = \int_{\mathbb{R}^p} \mathcal{K}(u) f_0(x + hu) \mathrm{d}u$ and $\beta_h = \int_{\mathbb{R}^p} f_h(x) f_0(x) \mathrm{d}x$. By symmetry of $\mathcal{K}(u)$,

$$\mathbb{E}\left[u_{ij}^n\left(W_i, W_j\right) \middle| W_i \right] = \mathbb{E}\left[u_{ji}^n\left(W_j, W_i\right) \middle| W_i \right] = \frac{1}{\sqrt{n}(n-1)} \{ f_h(x_i) - \beta_0 \},$$

$$\mathbb{E}\left[u_{ij}^n\left(W_i, W_j\right) \right] = \frac{1}{\sqrt{n}(n-1)} \{ \beta_h - \beta_0 \},$$

https://doi.org/10.1017/S026646661600013X Published online by Cambridge University Press

so the terms in the decomposition (3) are of the form

$$B_n = \sqrt{n}\{\beta_h - \beta_0\}, \qquad \Psi_n = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} 2\{f_h(x_i) - \beta_h\},$$

$$U_n = \frac{2}{\sqrt{n}(n-1)}\sum_{i=1}^{n}\sum_{j=1, j<i}^{n}\left\{\mathcal{K}_h(x_i - x_j) - f_h(x_i) - f_h(x_j) + \beta_h\right\}.$$

Here, $2\{f_h(x_i) - \beta_h\}$ is an approximation to the well known influence function $2\{f_0(x_i) - \beta_0\}$ for estimators of the integrated squared density. Under regularity conditions, $f_h(x_i)$ converges to $f_0(x_i)$ in mean square as $h \to 0$, so that

$$\Psi_n = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} 2\{f_0(x_i) - \beta_0\} + o_p(1).$$

A martingale central limit theorem can be applied as in Cattaneo, Crump, and Jansson (2014b) to show that the degenerate U-statistic term U_n will be asymptotically normal as $h \to 0$ and $n \to \infty$, provided that $n^2 h^p \to \infty$. It is easy to show that $n^2 h^p \mathbb{V}[U_n] \to \Delta = \beta_0 \int_{\mathbb{R}^p} \mathcal{K}(u)^2 \mathrm{d}u$, under mild regularity conditions. Alternative asymptotics occurs when h^p shrinks as fast as $1/n$, resulting in $\mathbb{V}[\Psi_n]$ and $\mathbb{V}[U_n]$ having the same magnitude in the limit.

3. MANY COVARIATE ASYMPTOTICS

In this section we make precise the previous discussion for many covariate asymptotics and also consider inference under homoskedasticity. The estimator $\hat{\beta}$ described above for many covariates can be interpreted as a two-step semiparametric estimator with tuning parameter K, the first step involving series estimation of the unknown (regression) functions $g(z)$ and $h(z)$. Donald and Newey (1994) gave conditions for asymptotic normality of this estimator when $K/n \to 0$. Here we generalize their findings by obtaining an asymptotic distributional result that is valid even when K/n is bounded away from zero.

The analysis proceeds under the following assumption.

Assumption PLM (Partially Linear Model)

(a) $(y_i, x_i', z_i')'$, $i = 1, \ldots, n$, is a random sample.

(b) There is a $C < \infty$ such that $\mathbb{E}\left[\varepsilon_i^4 \middle| x_i, z_i\right] \leq C$ and $\mathbb{E}\left[\|v_i\|^4 \middle| z_i\right] \leq C$.

(c) There is a $C > 0$ such that $\mathbb{E}\left[\varepsilon_i^2 \middle| x_i, z_i\right] \geq C$ and $\lambda_{\min}(\mathbb{E}[v_i v_i' | z_i]) \geq C$.

(d) $\operatorname{rank}(P_K) = K$ (a.s.) and there is a $C > 0$ such that $M_{ii} \geq C$.

(e) For some $\alpha_g, \alpha_h > 0$, there is a $C < \infty$ such that

$$\min_{\eta_g \in \mathbb{R}^K} \mathbb{E}\left[|g(z_i) - \eta_g' p_K(z_i)|^2\right] \leq C K^{-2\alpha_g},$$

$$\min_{\eta_h \in \mathbb{R}^{K \times d}} \mathbb{E}\left[\|h(z_i) - \eta_h' p_K(z_i)\|^2\right] \leq C K^{-2\alpha_h}.$$

https://doi.org/10.1017/S026646661600013X Published online by Cambridge University Press

Because $\sum_{i=1}^{n} M_{ii} = n - K$, an implication of part (d) is that $K/n \leq 1 - C < 1$, but crucially Assumption PLM does not imply that $K/n \to 0$. Part (e) is implied by conventional assumptions from approximation theory. For instance, when the support of z_i is compact, commonly used bases for approximation, such as polynomials or splines, will satisfy this assumption with $\alpha_g = s_g/d_z$ and $\alpha_h = s_h/d_z$, where s_g and s_h denote the number of continuous derivatives of $g(z)$ and $h(z)$, respectively. Further discussion and related references for several bases of approximation may be found in Chen (2007).

3.1. Asymptotic Distribution

From the discussion in the previous section, we see that the asymptotic distribution of $\hat{\beta}$ will be determined by the behavior of $\hat{\Gamma}_n$ and S_n. The following lemma approximates $\hat{\Gamma}_n$ without requiring that $K/n \to 0$.

LEMMA 1. *If Assumption PLM is satisfied and if $K \to \infty$, then*

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

Because $\sum_{i=1}^{n} M_{ii} = n - K$, it follows from this result that in the homoskedastic v_i case (i.e., when $\mathbb{E}[v_i v_i' | z_i] = \mathbb{E}[v_i v_i'])$ $\hat{\Gamma}_n$ is close to

$$\Gamma_n = (1 - K/n)\Gamma, \qquad \Gamma = \mathbb{E}[v_i v_i'],$$

in probability. More generally, with heteroskedasticity, $\hat{\Gamma}_n$ will be close to the weighted average Γ_n. Importantly, this result includes standard asymptotics as a special case when $K/n \to 0$, where $\sum_{i=1}^{n}(1 - M_{ii})/n = K/n$, the law of large numbers and iterated expectations imply

$$\begin{aligned}\Gamma_n &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[v_i v_i' | z_i\right] - \frac{1}{n}\sum_{i=1}^{n}(1 - M_{ii})\mathbb{E}\left[v_i v_i' | z_i\right] + o_p(1) \\ &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[v_i v_i' | z_i\right] + o_p(1) = \Gamma + o_p(1).\end{aligned}$$

Next, we study

$$S_n = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\sum_{j=1}^{n} M_{ij} v_i \varepsilon_j + B_n + R_n.$$

The following lemma quantifies the magnitude of the bias term B_n as well as the additional variability arising from the (remainder) term R_n.

https://doi.org/10.1017/S026646661600013X Published online by Cambridge University Press

LEMMA 2. *If Assumption PLM is satisfied and if* $K \to \infty$, *then* $B_n = O_p(\sqrt{n}K^{-\alpha_g-\alpha_h})$ *and* $R_n = o_p(1)$.

Like the previous lemma, this lemma does not require $K/n \to 0$. Interestingly, the bias term B_n involves approximation of both unknown functions $g(z)$ and $h(z)$, implying an implicit trade-off between smoothness conditions for $g(z)$ and $h(z)$. The implied bias condition $K^{2(\alpha_g+\alpha_h)}/n \to \infty$ only requires that $\alpha_g + \alpha_h$ be large enough, but not necessarily that α_g and α_h separately be large. It follows that if this bias condition holds, then

$$S_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \sum_{j=1}^{n} M_{ij} v_i \varepsilon_j + o_p(1),$$

as argued heuristically in the previous section.

Having dispensed with asymptotically negligible contributions to S_n, we turn to its leading term. This term is shown below to be asymptotically Gaussian with asymptotic variance given by

$$\begin{aligned}\Sigma_n &= \frac{1}{n}\mathbb{V}\left[\sum_{i=1}^{n}\sum_{j=1}^{n} M_{ij} v_i \varepsilon_j \,\middle|\, Z\right] \\ &= \frac{1}{n}\sum_{i=1}^{n} M_{ii}^2 \mathbb{E}\left[v_i v_i' \varepsilon_i^2 \middle| z_i\right] + \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1, j\neq i}^{n} M_{ij}^2 \mathbb{E}\left[v_i v_i' \varepsilon_j^2 \middle| z_i, z_j\right].\end{aligned}$$

Here, the first term following the second equality corresponds to the usual asymptotic approximation, while the second term adds an additional term that accounts for large K. Once again it is interesting to consider what happens in some special cases. Under homoskedasticity of ε_i (i.e., when $\mathbb{E}[\varepsilon_i^2 | x_i, z_i] = \mathbb{E}[\varepsilon_i^2]$),

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

because $\sum_{j=1}^{n} M_{ij}^2 = M_{ii}$. If, in addition, $\mathbb{E}[v_i v_i' | z_i] = \mathbb{E}[v_i v_i']$, then $\Sigma_n = \sigma_\varepsilon^2 (1 - K/n)\Gamma$. Also, if $K/n \to 0$, then by $\sum_{i=1}^{n}\sum_{j=1, j\neq i}^{n} M_{ij}^2/n \leq K/n$ and the law of large numbers, we have

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

which corresponds to the standard asymptotics limiting variance.

https://doi.org/10.1017/S026646661600013X Published online by Cambridge University Press

The following theorem combines Lemmas 1 and 2 with a central limit theorem for quadratic forms to show asymptotic normality of $\hat{\beta}$.

THEOREM 1. *If Assumption PLM is satisfied and if* $K^{2(\alpha_g+\alpha_h)}/n \to \infty$, *then*

$$\Omega_n^{-1/2}\sqrt{n}\left(\hat{\beta}-\beta_0\right) \to_d \mathcal{N}(0, I_d), \qquad \Omega_n = \Gamma_n^{-1}\Sigma_n\Gamma_n^{-1}.$$

If, in addition, $\mathbb{E}\left[\varepsilon_i^2 | x_i, z_i\right] = \sigma_\varepsilon^2$, *then* $\Omega_n = \sigma_\varepsilon^2\Gamma_n^{-1}$.

This theorem shows that $\hat{\beta}$ is asymptotically normal when K/n need not converge to zero. An implication of this result is that inconsistent series-based nonparametric estimators of the unknown functions $g(z)$ and $h(z)$ may be employed when forming $\hat{\beta}$, that is, $K/n \nrightarrow 0$ is allowed (increasing the variability of the nonparametric estimators), provided that $K \to \infty$ (to remove nonparametric smoothing bias). This asymptotic distributional result does not rely on asymptotic linearity, nor on the actual convergence of the matrices Γ_n and Σ_n, and leads to a new (larger) asymptotic variance that captures terms that are assumed away by the classical result. The asymptotic distribution result of Donald and Newey (1994) is obtained as a special case where $K/n \to 0$. More generally, when K/n does not converge to zero, the asymptotic variance will be larger than the usual formula because it accounts for the contribution of "remainder" U_n in equation (3). For instance, when both ε_i and v_i are homoskedastic, the asymptotic variance is

$$\Gamma_n^{-1}\Sigma_n\Gamma_n^{-1} = \sigma_\varepsilon^2\Gamma_n^{-1} = \sigma_\varepsilon^2\Gamma^{-1}(1-K/n)^{-1},$$

which is larger than the usual asymptotic variance $\sigma_\varepsilon^2\Gamma^{-1}$ by the degrees-of-freedom correction $(1-K/n)^{-1}$.

3.2. Asymptotic Variance Estimation under Homoskedasticity

Consistent asymptotic variance estimation is useful for large sample inference. If the assumptions of Theorem 1 are satisfied and if $\hat{\Sigma}_n - \Sigma_n \to_p 0$, then

$$\hat{\Omega}_n^{-1/2}\sqrt{n}\left(\hat{\beta}-\beta_0\right) \to_d \mathcal{N}(0, I_d), \qquad \hat{\Omega}_n = \hat{\Gamma}_n^{-1}\hat{\Sigma}_n\hat{\Gamma}_n^{-1},$$

implying that valid large-sample confidence intervals and hypothesis tests for linear and nonlinear transformations of the parameter vector β can be based on $\hat{\Omega}_n$.[2] Under (conditional) heteroskedasticity of unknown form, constructing a consistent estimator $\hat{\Sigma}_n$ turns out to be very challenging if $K/n \nrightarrow 0$. Intuitively, the problem arises because the estimated residuals entering the construction of $\hat{\Sigma}_n$ are not consistent unless $K/n \to 0$, implying that $\hat{\Sigma}_n - \Sigma_n \nrightarrow_p 0$ in general. Solving this problem is beyond the scope of this paper; see Cattaneo, Jansson, and Newey (2015).

Under homoskedasticity of ε_i; however, the asymptotic variance Σ_n simplifies and admits a correspondingly simple consistent estimator. To describe this result,

https://doi.org/10.1017/S026646661600013X Published online by Cambridge University Press

note that if $\mathbb{E}\left[\varepsilon_i^2 \middle| x_i, z_i\right] = \sigma_\varepsilon^2$ then $\Sigma_n = \sigma_\varepsilon^2 \Gamma_n$, where $\hat{\Gamma}_n - \Gamma_n \to_p 0$ by Lemma 1. It therefore suffices to find a consistent estimator of σ_ε^2. Let

$$s^2 = \frac{1}{n-d-K} \sum_{i=1}^{n} \hat{\varepsilon}_i^2, \qquad \hat{\varepsilon}_i = \sum_{j=1}^{n} M_{ij} \left(y_j - \hat{\beta}' x_j\right),$$

denote the usual OLS estimator of σ_ε^2 incorporating a degrees-of-freedom correction.

The following theorem shows that s^2 is a consistent estimator, even when the number of terms is "large" relative to the sample size.

THEOREM 2. *Suppose the conditions of Theorem 1 are satisfied. If* $\mathbb{E}\left[\varepsilon_i^2 \middle| x_i, z_i\right] = \sigma_\varepsilon^2$, *then* $s^2 \to_p \sigma_\varepsilon^2$ *and* $\hat{\Sigma}_n^{\texttt{HOM}} - \Sigma_n \to_p 0$, *where* $\hat{\Sigma}_n^{\texttt{HOM}} = s^2 \hat{\Gamma}_n$.

This theorem provides a distribution free, large sample justification for the degrees-of-freedom correction required for exact inference under homoskedastic Gaussian errors. Intuitively, accounting for the correct degrees of freedom is important whenever the number of terms in the semilinear model is "large" relative to the sample size.

4. SMALL SIMULATION STUDY

We conducted a Monte Carlo experiment to explore the extent to which the asymptotic theoretical results obtained in the previous section are present in small samples. Using the notation already introduced, we consider the following partially linear model:

$$\begin{aligned} y_i &= x_i'\beta + g(z_i) + \varepsilon_i, & \mathbb{E}[\varepsilon_i | x_i, z_i] &= 0, & \mathbb{E}\left[\varepsilon_i^2 \middle| x_i, z_i\right] &= \sigma_\varepsilon^2, \\ x_i &= h(z_i) + v_i, & \mathbb{E}[v_i | z_i] &= 0, & \mathbb{E}\left[v_i^2 \middle| z_i\right] &= \sigma_v^2(z_i), \end{aligned}$$

where $d = 1$, $\beta = 1$, $d_z = 5$, $z_i = (z_{1i}, \cdots, z_{d_z i})'$ with $z_{\ell i} \sim$ i.i.d. Uniform$(-1, 1)$, $\ell = 1, \cdots, d_z$. The unknown regression functions are set to $g(z_i) = h(z_i) = \exp(\|z_i\|^2)$, which are not additive separable in the covariates z_i. The simulation study is based on $S = 5,000$ replications, each replication taking a random sample of size $n = 500$ with all random variables generated independently. We consider 6 data generating processes (DGPs) as follows:

Data Generating Process for Monte Carlo Experiment

	(ε_i, v_i) – Distributions		
	Gaussian	Asymmetric	Bimodal
$\sigma_v^2(z_i) = 1$	Model 1	Model 3	Model 5
$\sigma_v^2(z_i) = \varsigma\left(1 + \|z_i\|^2\right)^2$	Model 2	Model 4	Model 6

Specifically, Models 1, 3, and 5 correspond to homoskedastic (in v_i) DGPs, while Models 2, 4, and 6 correspond to heteroskedastic (in v_i) DGPs. For the latter

https://doi.org/10.1017/S026646661600013X Published online by Cambridge University Press

models, the constant ς was chosen so that $\mathbb{E}[v_i^2] = 1$. The three distributions considered for the unobserved error terms ε_i and v_i are: the standard Normal (labelled "Gaussian") and two Mixture of Normals inducing either an asymmetric or a bimodal distribution; their Lebesgue densities are depicted in Figure 1. We explored other specifications for the regression functions, heteroskedasticity form, and distributional assumptions, but we do not report these additional results because they were qualitatively similar to those discussed here.

The estimators considered in the Monte Carlo experiment are constructed using power series approximations. We do not impose additive separability on the basis, though we do restrict the interaction terms to not exceed degree 5. To be specific, we consider the following polynomial basis expansion:

Polynomial Basis Expansion: $d_z = 5$ and $n = 500$

K	$p_K(z_i)$	K/n
6	$(1, z_{1i}, z_{2i}, z_{3i}, z_{4i}, z_{5i})'$	0.012
11	$\left(p_6(z_i)', z_{1i}^2, z_{2i}^2, z_{3i}^2, z_{4i}^2, z_{5i}^2\right)'$	0.022
21	$p_{11}(z_i)$ + first-order interactions	0.042
26	$\left(p_{21}(z_i)', z_{1i}^3, z_{2i}^3, z_{3i}^3, z_{4i}^3, z_{5i}^3\right)'$	0.052
56	$p_{26}(z_i)$ + second-order interactions	0.112
61	$\left(p_{56}(z_i)', z_{1i}^4, z_{2i}^4, z_{3i}^4, z_{4i}^4, z_{5i}^4\right)'$	0.122
126	$p_{61}(z_i)$ + third-order interactions	0.252
131	$\left(p_{126}(z_i)', z_{1i}^5, z_{2i}^5, z_{3i}^5, z_{4i}^5, z_{5i}^5\right)'$	0.262
252	$p_{131}(z_i)$ + fourth-order interactions	0.504
257	$\left(p_{252}(z_i)', z_{1i}^6, z_{2i}^6, z_{3i}^6, z_{4i}^6, z_{5i}^6\right)'$	0.514
262	$\left(p_{257}(z_i)', z_{1i}^7, z_{2i}^7, z_{3i}^7, z_{4i}^7, z_{5i}^7\right)'$	0.524
267	$\left(p_{262}(z_i)', z_{1i}^8, z_{2i}^8, z_{3i}^8, z_{4i}^8, z_{5i}^8\right)'$	0.534
272	$\left(p_{267}(z_i)', z_{1i}^9, z_{2i}^9, z_{3i}^9, z_{4i}^9, z_{5i}^9\right)'$	0.544
277	$\left(p_{272}(z_i)', z_{1i}^{10}, z_{2i}^{10}, z_{3i}^{10}, z_{4i}^{10}, z_{5i}^{10}\right)'$	0.554

Thus, our simulations explore the consequences of introducing many terms in the partially linear model by varying K on the grid above from $K = 6$ to $K = 277$, which gives a range for K/n of $\{0.012, \cdots, 0.554\}$. For each point on the grid of K/n, we report average bias, average standard deviation, mean square error and average standardized bias of $\hat{\beta}$ across simulations. We also consider the coverage error rates and interval length for two asymptotic 95% confidence intervals:

$$\mathrm{CI}_0 = \left[\hat{\beta} - \Phi^{-1}_{1-\alpha/2}\frac{\hat{\sigma}\,\hat{\Gamma}_n^{-1/2}}{\sqrt{n}} \quad , \quad \hat{\beta} + \Phi^{-1}_{1-\alpha/2}\frac{\hat{\sigma}\,\hat{\Gamma}_n^{-1/2}}{\sqrt{n}}\right],$$

$$\mathrm{CI}_1 = \left[\hat{\beta} - \Phi^{-1}_{1-\alpha/2}\frac{s\hat{\Gamma}_n^{-1/2}}{\sqrt{n}} \quad , \quad \hat{\beta} + \Phi^{-1}_{1-\alpha/2}\frac{s\hat{\Gamma}_n^{-1/2}}{\sqrt{n}}\right],$$

where $\hat{\sigma}^2 = (n - d - K)s^2/n$, and $\Phi_u^{-1} = \Phi^{-1}(u)$ denotes the inverse of the Gaussian distribution function. That is, CI_0 and CI_1 are formed employing the

https://doi.org/10.1017/S026646661600013X Published online by Cambridge University Press

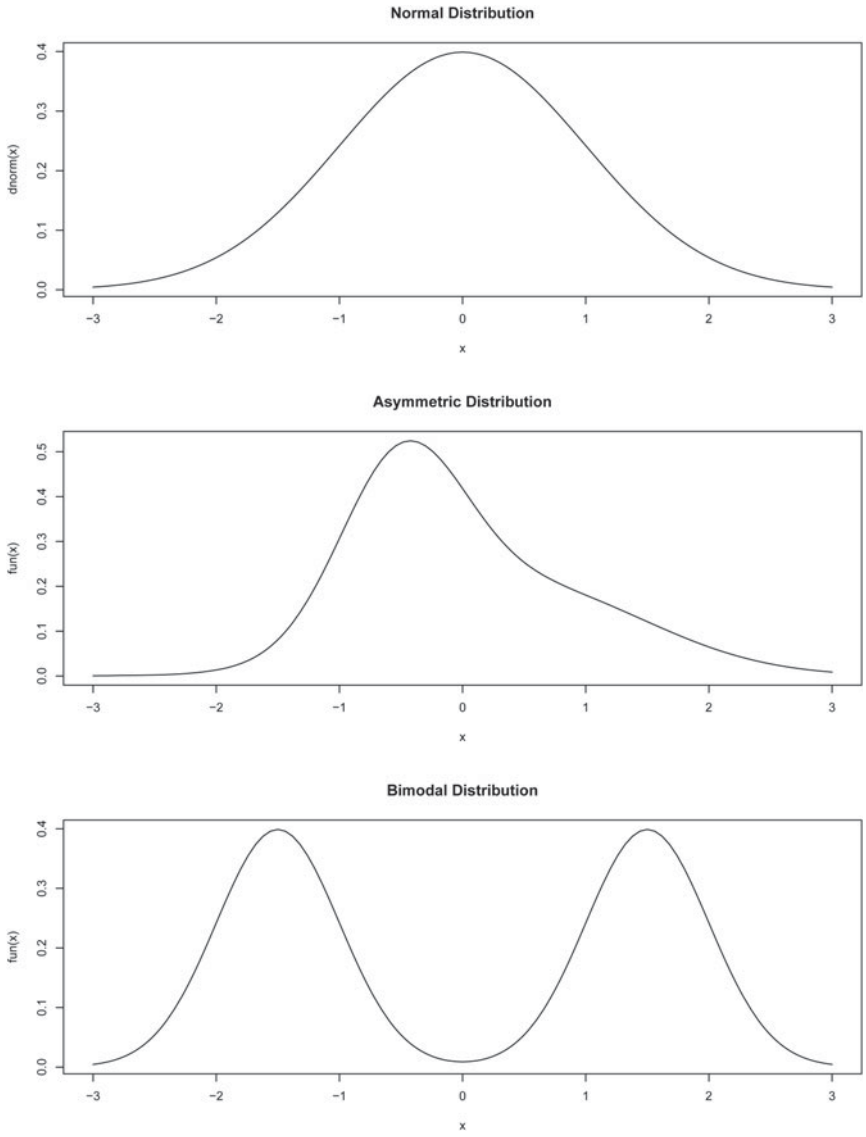


FIGURE 1. Lebesgue densities of error terms distributions.

t-statistic constructed using the homoskedasticity-consistent variance estimators without and with degrees-of-freedom correction, respectively.

The main findings from the Monte Carlo experiment are presented in Tables 1–3. All results are consistent with the theoretical conclusions presented in the previous section. First, the results for standard Normal and non-Normal errors are qualitatively similar. This indicates that the Gaussian approximation

https://doi.org/10.1017/S026646661600013X Published online by Cambridge University Press

TABLE 1. Simulation results, models 1–2, Gaussian distribution

(a) Model 1: Homoskedastic v_i

K/n	Bias	SD	RMSE	$\frac{\text{Bias}}{\text{SD}}$	CI_0	CI_1	$\hat{\sigma}$	s
0.012	0.481	0.040	0.483	11.898	0.000	0.000	0.039	0.039
0.022	0.001	0.045	0.045	0.031	0.947	0.950	0.045	0.045
0.042	0.002	0.047	0.047	0.051	0.939	0.945	0.045	0.046
0.052	0.002	0.046	0.046	0.049	0.940	0.947	0.045	0.046
0.112	0.002	0.047	0.047	0.041	0.936	0.952	0.045	0.048
0.122	0.000	0.048	0.048	0.005	0.935	0.949	0.045	0.048
0.252	0.001	0.052	0.052	0.013	0.907	0.947	0.045	0.052
0.262	0.000	0.052	0.052	−0.008	0.904	0.949	0.045	0.052
0.504	0.000	0.063	0.063	0.003	0.841	0.951	0.045	0.064
0.514	0.000	0.064	0.064	−0.002	0.828	0.947	0.045	0.064
0.524	0.000	0.064	0.064	−0.003	0.827	0.948	0.045	0.065
0.534	0.000	0.066	0.066	−0.003	0.821	0.950	0.045	0.066
0.544	0.001	0.068	0.068	0.010	0.803	0.946	0.045	0.067
0.554	0.000	0.067	0.067	0.004	0.808	0.949	0.045	0.067

(b) Model 2: Heteroskedastic v_i

K/n	Bias	SD	RMSE	$\frac{\text{Bias}}{\text{SD}}$	CI_0	CI_1	$\hat{\sigma}$	s
0.012	0.483	0.046	0.485	10.460	0.000	0.000	0.039	0.040
0.022	0.002	0.045	0.045	0.034	0.949	0.953	0.045	0.046
0.042	0.001	0.046	0.046	0.015	0.946	0.949	0.045	0.046
0.052	0.002	0.046	0.046	0.034	0.947	0.955	0.045	0.046
0.112	0.001	0.049	0.049	0.015	0.932	0.950	0.045	0.048
0.122	0.001	0.049	0.049	0.025	0.929	0.946	0.045	0.049
0.252	0.000	0.052	0.052	0.009	0.914	0.951	0.046	0.053
0.262	0.001	0.053	0.053	0.025	0.915	0.952	0.046	0.054
0.504	0.000	0.068	0.068	0.002	0.827	0.947	0.048	0.068
0.514	0.001	0.068	0.068	0.019	0.829	0.953	0.048	0.068
0.524	0.003	0.068	0.069	0.050	0.824	0.953	0.047	0.069
0.534	0.000	0.070	0.070	0.003	0.819	0.949	0.048	0.070
0.544	0.002	0.070	0.070	0.024	0.819	0.948	0.048	0.071
0.554	0.000	0.074	0.074	−0.004	0.801	0.943	0.048	0.072

Notes:
(i) columns Bias, SD, RMSE and $\frac{\text{Bias}}{\text{SD}}$ report, respectively, average bias, average standard deviation, root mean square error, and average standarized bias of the estimator $\hat{\beta}$ across simulations;
(ii) columns CI_0 and CI_1 report empirical coverage for homoskedastic-consistent confidence intervals, respectively, without and with degrees-of-freedom correction;
(iii) columns $\hat{\sigma}$ and s report the average across simulations of the standard errors estimators, respectively, without and with degrees-of-freedom correction.

obtained in Theorem 1 is a good approximation in finite samples, even when K is a nontrivial fraction of the sample size. Second, as expected, a small choice of K leads to important smoothing biases. This affects the finite sample properties

https://doi.org/10.1017/S026646661600013X Published online by Cambridge University Press

TABLE 2. Simulation results, models 3–4, asymmetric distribution

(a) Model 3: Homoskedastic v_i

K/n	Bias	SD	RMSE	$\frac{\text{Bias}}{\text{SD}}$	CI_0	CI_1	$\hat{\sigma}$	s
0.012	0.481	0.039	0.483	12.486	0.000	0.000	0.038	0.038
0.022	0.002	0.043	0.043	0.040	0.943	0.946	0.042	0.042
0.042	0.001	0.044	0.044	0.032	0.942	0.947	0.042	0.043
0.052	0.001	0.043	0.043	0.023	0.946	0.954	0.042	0.043
0.112	0.001	0.045	0.045	0.023	0.931	0.947	0.042	0.044
0.122	0.002	0.045	0.045	0.036	0.936	0.951	0.042	0.045
0.252	0.001	0.049	0.049	0.013	0.902	0.950	0.042	0.048
0.262	0.001	0.049	0.049	0.013	0.915	0.953	0.042	0.049
0.504	0.000	0.060	0.060	0.001	0.829	0.950	0.042	0.059
0.514	0.000	0.060	0.060	−0.007	0.828	0.948	0.042	0.060
0.524	0.000	0.060	0.060	−0.006	0.830	0.952	0.042	0.061
0.534	0.000	0.061	0.061	−0.001	0.819	0.950	0.042	0.061
0.544	0.000	0.062	0.062	0.000	0.809	0.951	0.042	0.062
0.554	0.001	0.064	0.064	0.009	0.794	0.944	0.042	0.063

(b) Model 4: Heteroskedastic v_i

K/n	Bias	SD	RMSE	$\frac{\text{Bias}}{\text{SD}}$	CI_0	CI_1	$\hat{\sigma}$	s
0.012	0.485	0.046	0.488	10.566	0.000	0.000	0.038	0.038
0.022	0.001	0.042	0.042	0.031	0.947	0.949	0.042	0.043
0.042	0.001	0.043	0.043	0.025	0.946	0.951	0.042	0.043
0.052	0.002	0.044	0.044	0.047	0.937	0.943	0.042	0.043
0.112	0.002	0.045	0.045	0.037	0.933	0.945	0.043	0.045
0.122	0.001	0.046	0.046	0.025	0.929	0.945	0.043	0.046
0.252	0.000	0.050	0.050	−0.004	0.910	0.949	0.043	0.050
0.262	0.001	0.050	0.050	0.020	0.907	0.951	0.043	0.050
0.504	0.000	0.064	0.064	−0.002	0.832	0.947	0.045	0.064
0.514	0.001	0.065	0.065	0.008	0.827	0.948	0.045	0.064
0.524	−0.001	0.065	0.065	−0.015	0.817	0.948	0.045	0.065
0.534	0.001	0.066	0.066	0.013	0.824	0.948	0.045	0.065
0.544	0.000	0.067	0.067	−0.002	0.799	0.951	0.045	0.066
0.554	0.000	0.067	0.067	−0.001	0.811	0.948	0.045	0.067

Notes:
(i) columns Bias, SD, RMSE and $\frac{\text{Bias}}{\text{SD}}$ report, respectively, average bias, average standard deviation, root mean square error, and average standarized bias of the estimator $\hat{\beta}$ across simulations;
(ii) columns CI_0 and CI_1 report empirical coverage for homoskedastic-consistent confidence intervals, respectively, without and with degrees-of-freedom correction;
(iii) columns $\hat{\sigma}$ and s report the average across simulations of the standard errors estimators, respectively, without and with degrees-of-freedom correction.

of the point estimators as well as the distributional approximations obtained in this paper. In particular, it affects the empirical size of all the confidence intervals. Third, in all cases the results under homoskedasticity or heteroskedasticity

https://doi.org/10.1017/S026646661600013X Published online by Cambridge University Press

TABLE 3. Simulation results, models 5–6, bimodal distribution

(a) Model 5: Homoskedastic v_i

K/n	Bias	SD	RMSE	$\frac{\text{Bias}}{\text{SD}}$	CI_0	CI_1	$\hat{\sigma}$	s
0.012	0.482	0.058	0.486	8.340	0.000	0.000	0.059	0.059
0.022	0.001	0.076	0.076	0.009	0.948	0.950	0.076	0.077
0.042	0.001	0.078	0.078	0.008	0.944	0.948	0.076	0.077
0.052	−0.001	0.078	0.078	−0.010	0.940	0.948	0.076	0.078
0.112	0.002	0.081	0.081	0.026	0.930	0.946	0.076	0.080
0.122	0.001	0.080	0.080	0.018	0.936	0.953	0.076	0.081
0.252	0.002	0.088	0.088	0.026	0.912	0.949	0.076	0.088
0.262	0.001	0.087	0.087	0.008	0.908	0.952	0.076	0.088
0.504	−0.001	0.109	0.109	−0.013	0.827	0.950	0.076	0.108
0.514	0.001	0.108	0.108	0.012	0.832	0.953	0.076	0.109
0.524	0.000	0.110	0.110	0.003	0.825	0.948	0.076	0.110
0.534	−0.004	0.110	0.110	−0.033	0.818	0.950	0.076	0.111
0.544	0.001	0.111	0.111	0.012	0.819	0.949	0.076	0.112
0.554	−0.001	0.111	0.111	−0.006	0.817	0.956	0.076	0.114

(b) Model 6: Heteroskedastic v_i

K/n	Bias	SD	RMSE	$\frac{\text{Bias}}{\text{SD}}$	CI_0	CI_1	$\hat{\sigma}$	s
0.012	0.483	0.062	0.487	7.811	0.000	0.000	0.059	0.060
0.022	0.001	0.077	0.077	0.011	0.945	0.948	0.076	0.077
0.042	0.001	0.077	0.077	0.011	0.945	0.951	0.076	0.078
0.052	−0.001	0.079	0.079	−0.009	0.941	0.948	0.077	0.079
0.112	0.000	0.082	0.082	0.001	0.938	0.954	0.077	0.082
0.122	0.004	0.080	0.080	0.046	0.942	0.955	0.077	0.082
0.252	0.000	0.092	0.092	0.002	0.904	0.946	0.078	0.090
0.262	0.002	0.089	0.089	0.026	0.910	0.957	0.078	0.091
0.504	−0.001	0.117	0.117	−0.005	0.826	0.946	0.080	0.114
0.514	−0.002	0.116	0.116	−0.017	0.828	0.951	0.081	0.116
0.524	0.000	0.118	0.118	0.003	0.821	0.945	0.081	0.117
0.534	0.001	0.118	0.118	0.010	0.815	0.953	0.081	0.119
0.544	0.000	0.119	0.119	−0.003	0.816	0.952	0.081	0.120
0.554	0.000	0.125	0.125	0.001	0.797	0.943	0.081	0.121

Notes:
(i) columns Bias, SD, RMSE and $\frac{\text{Bias}}{\text{SD}}$ report, respectively, average bias, average standard deviation, root mean square error, and average standarized bias of the estimator $\hat{\beta}$ across simulations;
(ii) columns CI_0 and CI_1 report empirical coverage for homoskedastic-consistent confidence intervals, respectively, without and with degrees-of-freedom correction;
(iii) columns $\hat{\sigma}$ and s report the average across simulations of the standard errors estimators, respectively, without and with degrees-of-freedom correction.

in v_i are qualitatively similar, showing that our theoretical results provide a good finite sample approximation in both cases, even when K is a nontrivial fraction of the sample size. Fourth, as suggested by Theorem 2, confidence intervals without

https://doi.org/10.1017/S026646661600013X Published online by Cambridge University Press

degrees-of-freedom correction (CI_0) are under-sized, while the confidence intervals with degrees-of-freedom correction (CI_1) have close-to-correct empirical size in all cases. This result shows that the degrees-of-freedom correction is crucial to achieve close-to-correct empirical size when K/n is non-negligible.

In conclusion, we found in our small-scale simulation study that our theoretical results for the partially linear model with possibly many terms provide good approximation in samples of moderate size. In particular, under homoskedasticity of ε_i, we showed that confidence intervals constructed using s^2 exhibit good empirical coverage even when K/n is "large". We also confirmed that the Gaussian distributional approximation given in Theorem 1 represents well the finite sample distribution of $\hat{\beta}$ even when K/n is "large".

In Cattaneo, Jansson, and Newey (2015) we analyze in detail the case of (conditional) heteroskedasticity in ε_i, which requires the use of a new standard error formula, and also compare those results to the case of homoskedasticity analyzed herein. The reader is referred to that work for further details.

5. CONCLUSION

This paper showed asymptotic normality and gave consistent standard errors for coefficients of interest when the number of covariates grows as fast as the sample size. It is also shown how this asymptotics has a similar structure to previously established results for many instrument asymptotics or small bandwidths. These results are all based on results for degenerate U-statistics, where asymptotic normality happens when the number of covariates diverges to infinity or the bandwidth shrinks to zero.

Our results apply to a class of semiparametric estimators $\hat{\beta}$ satisfying

$$\sqrt{n}\left(\hat{\beta}-\beta_0\right)=\hat{\Gamma}_n^{-1}S_n+o_p(1),$$

where $\hat{\Gamma}_n$ and S_n take a particular V-stastistic form, as discussed in Section 2. This class of semiparametric estimators covers several interesting problems, but it is by no means exhaustive. For example, Cattaneo and Jansson (2015) show that a large class of (kernel-based) semiparametric estimators admit an expansion of the form

$$\sqrt{n}\left(\hat{\beta}-\beta_0\right)=\hat{\Gamma}_n^{-1}S_n-\mathcal{B}_n+o_p(1),$$

where the bias term $\mathcal{B}_n$ is quantitatively and conceptually distinct from the smoothing bias B_n described in Section 2 and, crucially, dominates the quadratic term U_n arising from the V-statistic S_n; that is, $U_n=o_p(\mathcal{B}_n)$ in that setting. Nevertheless, the structure we have considered in this paper is useful, providing new results for the partially linear model and a common structure for disparate literatures on many instruments and small bandwidths.

Finally, as a reviewer pointed out, the alternative asymptotics discussed in this paper are also qualitatively distinct, but conceptually similar, to that encountered

https://doi.org/10.1017/S026646661600013X Published online by Cambridge University Press

in the recent literature on "large" panel data models where the number of units n and the number of periods T are proportional; see, for example, Alvarez and Arellano (2003), Hahn and Newey (2004) and references therein. Specifically, whereas the "large-(n, T) asymptotics" lead to the presence of a first-order bias in the distributional approximation (centering), the alternative asymptotics discussed in this paper lead to a change in the first-order variance of the distributional approximation (scale). Therefore, the "large-(n, T) asymptotics" in panel data contexts are more closely related to those obtained in Cattaneo and Jansson (2015) for nonlinear semiparametric problems, than to the distribution theory emerging from the common structure highlighted in this paper.

NOTES

1. In time series contexts, the exact decomposition is less useful, but approximations thereof with properties similar to those we discuss herein can be developed. For an example and related references see Atchadé and Cattaneo (2014).

2. Another approach to inference would be via the bootstrap. For small bandwidth asymptotics, Cattaneo, Crump, and Jansson (2014a) showed that the standard nonparametric bootstrap does not provide a valid distributional approximation in general. We conjecture that the standard nonparametric bootstrap will also fail to provide valid inference for other alternative asymptotic frameworks.

REFERENCES

Alvarez, J. & M. Arellano (2003) The time series and cross-section asymptotics of dynamic panel data estimators. *Econometrica* 71(4), 1121–1159.

Angrist, J., G.W. Imbens, & A. Krueger (1999) Jackknife instrumental variables estimation. *Journal of Applied Econometrics* 14(1), 57–67.

Aradillas–Lopéz, A., B.E. Honoré, & J.L. Powell (2007) Pairwise difference estimation with nonparametric control variables. *International Economic Review* 48, 1119–1158.

Atchadé, Y.F. & M.D. Cattaneo (2014) A martingale decomposition for quadratic forms of Markov chains (with applications). *Stochastic Processes and their Applications* 124(1), 646–677.

Bekker, P.A. (1994) Alternative approximations to the distributions of instrumental variables estimators. *Econometrica* 62, 657–681.

Belloni, A., V. Chernozhukov, & C. Hansen (2014) Inference on treatment effects after selection among high-dimensional controls. *Review of Economic Studies* 81(2), 608–650.

Cattaneo, M.D., R.K. Crump, & M. Jansson (2010) Robust data-driven inference for density-weighted average derivatives. *Journal of the American Statistical Association* 105(491), 1070–1083.

Cattaneo, M.D., R.K. Crump, & M. Jansson (2014a) Bootstrapping density-weighted average derivatives. *Econometric Theory* 30(6), 1135–1164.

Cattaneo, M.D., R.K. Crump, & M. Jansson (2014b) Small bandwidth asymptotics for density-weighted average derivatives. *Econometric Theory* 30(1), 176–200.

Cattaneo, M.D. & M. Jansson (2015) Bootstrapping Kernel-Based Semiparametric Estimators. Working paper, University of Michigan.

Cattaneo, M.D., M. Jansson, & W.K. Newey (2015) Treatment Effects With Many Covariates and Heteroskedasticity. Working paper, University of Michigan.

Chao, J.C., N.R. Swanson, J.A. Hausman, W.K. Newey, & T. Woutersen (2012) Asymptotic distribution of JIVE in a heteroskedastic IV regression with many instruments. *Econometric Theory* 28(1), 42–86.

Chen, X. (2007) Large sample sieve estimation of semi-nonparametric models. In J. Heckman, & E. Leamer (eds.), *Handbook of Econometrics*, vol. 6, pp. 5550–5632. Elsevier Science B.V.

https://doi.org/10.1017/S026646661600013X Published online by Cambridge University Press

de Jong, P. (1987) A central limit theorem for generalized quadratic forms. *Probability Theory and Related Fields* 75, 261–277.
Donald, S.G. & W.K. Newey (1994) Series estimation of semilinear models. *Journal of Multivariate Analysis* 50(1), 30–40.
El Karoui, N., D. Bean, P.J. Bickel, C. Lim, & B. Yu (2013) On robust regression with high-dimensional predictors. *Proceedings of the National Academy of Sciences* 110(36), 14557–14562.
Escanciano, J.C. & D. Jacho-Chavez (2012) $\sqrt{n}$-Uniformly consistent density estimation in nonparametric regression. *Journal of Econometrics* 167(1), 305–316.
Giné, E. & R. Nickl (2008) A simple adaptive estimator of the integrated square of a density. *Bernoulli* 14(1), 47–61.
Hahn, J. & W.K. Newey (2004) Jackknife and analytical bias reduction for nonlinear panel data models. *Econometrica* 72(4), 1295–1319.
Hansen, C., J. Hausman, & W.K. Newey (2008) Estimation with many instrumental variables. *Journal of Business and Economic Statistics* 26(4), 398–422.
Heckman, J.J. & E.J. Vytlacil (2007) Econometric evaluation of social programs, Part I: Causal models, structural models and econometric policy evaluation. In J. Heckman, & E. Leamer (eds.), *Handbook of Econometrics*, vol. 6, pp. 4780–4874. Elsevier Science B.V.
Huber, P.J. (1973) Robust regression: Asymptotics, conjectures, and Monte Carlo. *Annals of Stastistics* 1(5), 799–821.
Imbens, G.W. & J.M. Wooldridge (2009) Recent developments in the econometrics of program evaluation. *Journal of Economic Literature* 47(1), 5–86.
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), 693–700.
Li, Q. & S. Racine (2007) *Nonparametric Econometrics*. Princeton University Press.
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. (2009) Two-step series estimation of sample selection models. *Econometrics Journal* 12(1), S217–S229.
Newey, W.K., F. Hsieh, & J.M. Robins (2004) Twicing kernels and a small bias property of semiparametric estimators. *Econometrica* 72(1), 947–962.
Powell, J.L., J.H. Stock, & T.M. Stoker (1989) Semiparametric estimation of index coefficients. *Econometrica* 57(6), 1403–1430.
Schick, A. & W. Wefelmeyer (2013) Uniform convergence of convolution estimators for the response density in nonparametric regression. *Bernoulli* 19(5B), 2250–2276.
van der Vaart, A.W. (1998) *Asymptotic Statistics*. Cambridge University Press.

APPENDIX A: Proofs

All statements involving conditional expectations are understood to hold almost surely. Qualifiers such as "a.s." will be omitted to conserve space. Throughout the appendix, C will denote a generic constant that may take different values in each case.

Proof of Lemma 1. Let $X = [x_1, \ldots, x_n]'$, $H = [h_1, \ldots, h_n]'$, and $V = [v_1, \ldots, v_n]'$. By Assumption PLM and the Markov inequality,

$$\operatorname{tr}\left(\frac{1}{n} H'MH\right) = \min_{\eta_h \in \mathbb{R}^{K\times d}} \frac{1}{n}\sum_{i=1}^{n} \left\| h(z_i) - \eta_h' p_K(z_i) \right\|^2 = O_p\left(K^{-2\alpha_h}\right) \to_p 0.$$

https://doi.org/10.1017/S026646661600013X Published online by Cambridge University Press

Also, $V'V/n = O_p(1)$ by Assumption PLM and the Markov inequality, so by the Cauchy–Schwarz inequality and M idempotent, $\|H'MV/n\| \leq [\mathrm{tr}(H'MH/n)\mathrm{tr}(V'V/n)]^{1/2} \to_p 0$. By the triangle inequality, we then have

$$\hat{\Gamma}_n = \frac{1}{n}X'MX = \frac{1}{n}(V+H)'M(V+H) = \frac{1}{n}V'MV + o_p(1).$$

Next, by Lemma A1 of Chao et al. (2012),

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

Finally, by the Markov inequality and using $\mathbb{E}\left[n^{-1}\sum_{i=1}^{n} M_{ii}v_i v_i'|Z\right] = \Gamma_n$,

$$\frac{1}{n}\sum_{i=1}^{n} M_{ii}v_i v_i' - \Gamma_n \to_p 0$$

because Assumption PLM implies that $v_i v_i'$ and $v_j v_j'$ are uncorrelated conditional on Z and that $\mathbb{E}\left[M_{ii}^2\|v_i\|^4|Z\right] \leq C$. n

Proof of Lemma 2. Let $G = [g_1, \ldots, g_n]'$ and $\varepsilon = [\varepsilon_1, \ldots, \varepsilon_n]'$. By the Cauchy–Schwarz inequality, M idempotent, Assumption PLM, and the Markov inequality,

$$\left\|\frac{1}{n}G'MH\right\| \leq \sqrt{\mathrm{tr}\left(\frac{1}{n}G'MG\right)}\sqrt{\mathrm{tr}\left(\frac{1}{n}H'MH\right)} = O_p(K^{-\alpha_g-\alpha_h}),$$

which gives $B_n = G'MH/\sqrt{n} = O_p(\sqrt{n}K^{-\alpha_g-\alpha_h})$.

Also, $R_n = (V'MG + H'M\varepsilon)/\sqrt{n} = O_p(K^{-\alpha_g} + K^{-\alpha_h}) = o_p(1)$ because

$$\mathbb{E}\left[\left\|\frac{1}{\sqrt{n}}V'MG\right\|^2|Z\right] = \frac{1}{n}G'M\mathbb{E}[VV'|Z]MG \leq C\frac{1}{n}G'MG = O_p\left(K^{-2\alpha_g}\right)$$

and

$$\mathbb{E}\left[\left\|\frac{1}{\sqrt{n}}H'M\varepsilon\right\|^2|Z\right] = \mathrm{tr}\left(\frac{1}{n}H'M\mathbb{E}[\varepsilon\varepsilon'|Z]MH\right) \leq C\,\mathrm{tr}\left(\frac{1}{n}H'MH\right) = O_p\left(K^{-2\alpha_h}\right)$$

by Assumption PLM and the Markov inequality. n

Proof of Theorem 1. By Lemma A2 of Chao et al. (2012),

$$\Sigma_n^{-1/2}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\sum_{j=1}^{n} M_{ij}v_i\varepsilon_j \to_d \mathcal{N}(0, I_d)$$

under Assumption PLM. Combining this result with Lemmas 1 and 2, we obtain the results stated in the theorem. n

https://doi.org/10.1017/S026646661600013X Published online by Cambridge University Press

Proof of Theorem 2. Let $Y = [y_1, \ldots, y_n]$ and $\hat{\varepsilon} = [\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_n]' = M(Y - X\hat{\beta})$. It follows similarly to the proof of Lemma 1 that

$$\begin{aligned}\frac{1}{n}\varepsilon' M\varepsilon &= \frac{1}{n}\sum_{i=1}^{n} M_{ii}\varepsilon_i^2 + \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1, j\neq i}^{n} \varepsilon_i M_{ij}\varepsilon_j \\ &= \frac{1}{n}\sum_{i=1}^{n} M_{ii}\mathbb{E}\left[\varepsilon_i^2 \middle| z_i\right] + o_p(1) = \frac{n-K}{n}\sigma_\varepsilon^2 + o_p(1),\end{aligned}$$

so it suffices to show that $\hat{\varepsilon}'\hat{\varepsilon}/n = \varepsilon' M\varepsilon/n + o_p(1)$.

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 and $\varepsilon' M\varepsilon/n = O_p(1)$ gives

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

Similarly, $G'MG/n = o_p(1)$ together with $(Y - X\hat{\beta} - G)' M(Y - X\hat{\beta} - G)/n = O_p(1)$ and the Cauchy–Schwarz inequality gives

$$\frac{1}{n}\hat{\varepsilon}'\hat{\varepsilon} = \frac{1}{n}\left(Y - X\hat{\beta}\right)' M\left(Y - X\hat{\beta}\right) = \frac{1}{n}\left(Y - X\hat{\beta} - G\right)' M(Y - X\hat{\beta} - G) + o_p(1).$$

The conclusion follows by the triangle inequality. n

APPENDIX B: Extension to Two-step Estimation

The common structure highlighted in Section 2, and later used to study IV models with many instruments, kernel-based semiparametric estimators and the series-based semiparametric semilinear model, can be extended to account for preliminary estimation. This extension, though conceptually not difficult, may be important in series-based sample selection models as discussed in Newey (2009), or kernel-based estimators as discussed in Aradillas–Lopéz, Honoré, and Powell (2007) and Escanciano and Jacho-Chavez (2012). In this appendix we discuss this extension heuristically, but relegate a formal analysis for future work.

Following the ideas and notation introduced in Section 2, consider a generic estimator $\hat{\beta}(\hat{\theta})$ of the parameter $\beta_0 = \beta_0(\theta_0) \in \mathbb{R}^d$. In this appendix, the notation $\hat{\beta}(\hat{\theta})$ (as opposed to $\hat{\beta}$) makes explicit that the estimator depends on an estimator $\hat{\theta}$ of the unknown "parameter" $\theta_0 \in \Theta$, not necessarily finite dimensional. As a natural generalization of (2) we then assume that

$$\sqrt{n}\left(\hat{\beta}(\theta) - \beta_0\right) = \hat{\Gamma}_n(\theta)^{-1} S_n(\theta), \qquad S_n(\theta) = \sum_{i=1}^{n}\sum_{j=1}^{n} u_{ij}^n(W_i, W_j; \theta).$$

The exact form of $u_{ij}^n(W_i, W_j; \theta)$ is context specific; $u_{ij}^n(W_i, W_j) = u_{ij}^n(W_i, W_j; \theta_0)$ in Section 2 and other examples are given in the references above. Suppose, in addition, that

https://doi.org/10.1017/S026646661600013X Published online by Cambridge University Press

the estimator $\hat{\theta}$ is consistent in the sense that $\|\hat{\theta}-\theta_0\| = o_p(1)$, where $\|\cdot\|$ is some context specific norm (e.g., if $\Theta \subseteq \mathbb{R}^m$ then $\|\cdot\|$ will typically be the Euclidean norm).

It follows from the discussion in the paper, that the limiting distribution of $\sqrt{n}(\hat{\beta}(\hat{\theta}) - \beta_0)$ is determined by $S_n(\hat{\theta})$ whenever $\hat{\Gamma}_n(\theta_0)^{-1}\Gamma_n \to_p I_d$ and $\hat{\Gamma}_n(\theta_0)^{-1}\hat{\Gamma}_n(\hat{\theta}) \to_p I_d$. In many cases, the latter assumption only imposes a consistency requirement (without a rate) on the estimator $\hat{\theta}$ and is therefore not particularly restrictive. The term $S_n(\hat{\theta})$ can be handled, for example, by employing the obvious decomposition

$$S_n(\hat{\theta}) = F_n(\hat{\theta}) + S_n, \qquad F_n(\theta) = S_n(\theta) - S_n(\theta_0), \qquad S_n = S_n(\theta_0),$$

where now the asymptotic distributional approximation for $S_n(\hat{\theta})$ is explained by the first-step estimation contribution $F_n(\hat{\theta})$, and the "oracle" term S_n already studied in the main paper.

The additional term $F_n(\hat{\theta})$ may be analyzed in multiple ways. For example, if $\hat{\theta}$ is finite-dimensional, $\sqrt{n}$-consistent, and some regularity conditions hold (including $\theta \mapsto u_{ij}^n(w_1, w_2; \theta)$ sufficiently "smooth" and well-behaved), then it may be shown that

$$F_n(\hat{\theta}) - \dot{F}_n(\hat{\theta}) = o_p\left(n^{-1/2}\right), \qquad \dot{F}_n(\hat{\theta}) = \left(\sum_{i=1}^n \sum_{j=1}^n \dot{u}_{ij}^n(W_i, W_j; \theta_0)\right)\left(\hat{\theta} - \theta_0\right),$$

where $\dot{u}_{ij}^n(w_1, w_2; \theta_0)$ is some function. For instance, $\dot{u}_{ij}^n(w_1, w_2; \theta_0) = \partial u_{ij}^n(w_1, w_2; \theta_0)/\partial\theta$ if $\theta \mapsto u_{ij}^n(w_1, w_2; \theta)$ is differentiable or, otherwise, $\dot{u}_{ij}^n(w_1, w_2; \theta_0)$ may be obtained using U-process theory.

The above heuristics lead to the expansion

$$S_n(\hat{\theta}) = F_n(\hat{\theta}) + S_n = \hat{\Upsilon}_n\left(\hat{\theta} - \theta_0\right) + S_n + o_p\left(n^{-1/2}\right),$$

where

$$\hat{\Upsilon}_n = \sum_{i=1}^n \sum_{j=1}^n \dot{u}_{ij}^n(W_i, W_j; \theta_0).$$

This illustrates how the discussion given in the main text may be extended to the case of two-step estimation. Assuming the first-step estimator $\hat{\theta}$ is $\sqrt{n}$-consistent (as will be the case whenever it is regular), it follows that the first step makes a non-negligible contribution to the asymptotic distribution unless the "orthogonality" condition $\hat{\Upsilon}_n = o_p(n^2)$ is satisfied.

Formalizing the above ideas is beyond the scope of this paper, but we conjecture it can be done in fairly large generality, including some cases where $\hat{\theta}$ is infinite dimensional and (possibly) not $\sqrt{n}$-consistent.

https://doi.org/10.1017/S026646661600013X Published online by Cambridge University Press