

WEAK IDENTIFICATION AND CONDITIONAL MOMENT RESTRICTIONS*

Sung Jae Jun[†] and Joris Pinkse[‡]
 The Pennsylvania State University

March 2007

Abstract

We propose a test for the value of coefficients in models with conditional moment restrictions which is asymptotically valid regardless of identification strength. The test is in essence an Anderson–Rubin test using nonparametrically estimated instruments to which we apply a standard error correction. We show that a standard t -test using nonparametrically estimated instruments is asymptotically invalid when identification strength λ decreases at a rate of $1/\sqrt{k}$, where k is a smoothing parameter which increases at a rate less than the sample size n . The proposed test is consistent when λ goes to zero more slowly than $1/\sqrt[4]{nk}$ ($\lambda \succ 1/\sqrt[4]{nk}$) and has local power at least as good as fully parametric tests when $\lambda \succ 1/\sqrt{k}$. We show that the proposed test is asymptotically equivalent to an infeasible Anderson–Rubin test if $\lambda \succ 1/\sqrt{k}$. We allow for both two- and one-sided testing and the procedure accommodates both linear and nonlinear models. In all cases the limiting distribution under the null hypothesis is a standard normal.

*We thank seminar participants at Syracuse University, Columbia University, The Pennsylvania State University, and The Festschrift Conference for Tony Lancaster at Brown University for comments and discussions.

[†](corresponding author) Department of Economics, The Pennsylvania State University, 608 Kern Graduate Building, University Park PA 16802, sjun@psu.edu

[‡]joris@psu.edu

1 Introduction

We propose a test for the value of coefficients in models with conditional moment restrictions which is asymptotically valid regardless of identification strength. The test is in essence an Anderson–Rubin test using nonparametrically estimated instruments to which we apply a standard error correction. We show that a standard t -test using nonparametrically estimated instruments is asymptotically invalid when identification strength λ decreases at a rate of $1/\sqrt{k}$, where k is a smoothing parameter which increases at a rate less than the sample size n . The proposed test is consistent when λ goes to zero more slowly than $1/\sqrt[4]{nk}$ ($\lambda \succ 1/\sqrt[4]{nk}$) and has local power at least as good as fully parametric tests when $\lambda \succ 1/\sqrt{k}$. We show that the proposed test is asymptotically equivalent to an infeasible Anderson–Rubin test if $\lambda \succ 1/\sqrt{k}$. We allow for both two- and one-sided testing and the procedure accommodates both linear and nonlinear models. In all cases the limiting distribution under the null hypothesis is a standard normal.

Although our methods allow for general nonlinear moment conditions models we begin our discussion with a simple model in which the structural equation is linear, in which there are no nuisance parameters, i.e. $y_i = Y_i\theta_0 + u_i$, and in which there is homoskedasticity. We do so because the simple model will highlight the issues and differences across testing procedures with the least number of caveats and exceptions. The regressor Y_i is potentially endogenous and we have a vector of instruments z_i at our disposal such that $m_i(\theta) = E(y_i - Y_i\theta|z_i)$ equals zero at a single value θ_0 (unless $\lambda = 0$). The optimal instruments in this model are given by $g_i = g(z_i) = E(Y_i|z_i)$.

In parametric linear simultaneous equations models, it is assumed that $g(z) = z'\pi_0$, where π_0 can be estimated using an ordinary least squares regression of Y_i on z_i . If the predictions of this regression are subsequently used as instruments for Y_i then we have the familiar two stage least squares (2SLS) estimator. It is now well understood that the potential identification failure due to small values of π_0 can make 2SLS unreliable (see e.g., Staiger and Stock (1997), Stock, Wright, and Yogo (2002) among others). Staiger and Stock devised the notion of weak instruments to capture this phenomenon. They let $g(z) = z'\tilde{\pi}_0\lambda$ with λ equal to 0 or decreasing at a rate of $1/\sqrt{n}$. In reality λ does not vary with the sample size; weak instruments are a technical tool to study and improve on the behavior of test statistics when sample sizes are large and instruments are poor.

Under weak identification ($\lambda \preceq 1/\sqrt{n}$ in the parametric case) a parametric t -test for θ_0 suffers from serious size distortions, as is documented in Dufour (1997), Staiger and Stock (1997), and Stock, Wright, and Yogo (2002) among many others. There are several methods that address this issue in parametric models, of which we discuss the Anderson–Rubin (1949, AR) procedure, the Kleibergen (2002, 2005; K) test, and the Moreira (2003, M).¹ Andrews, Moreira and Stock (2004) found in a simulation study that the M-test approximates optimal average power in the above model with linear g and normally distributed errors which are independent of the instruments. The M-test moreover does not have the peculiar power curve of the K-test and unlike the AR-test it does not lose power fast when the dimension of z increases.

If the parametric assumption on g is not satisfied then two things could happen. First, it is possible (but unlikely) that g is such that θ_0 is identified off the conditional moment condition but not off the unconditional moment condition using instruments z_i . The other, nearly inevitable, possibility is that θ_0 is identified but that the strength of z_i as an instrument is less than that of g_i .

However, if g is estimated using nonparametric methods then a standard t -test using the resulting semiparametric estimator $\hat{\theta}$ of θ_0 breaks down earlier than if a correct parametric specification was chosen for g . Specifically, we show that if g can be written as $g(z) = \lambda \tilde{g}(z)$ with $0 < E\tilde{g}_i^2 < \infty$, and g is estimated using k -nearest neighbor estimation then $\hat{\theta}$ ceases to be a consistent estimator when $\lambda \preceq 1/\sqrt[4]{nk}$ and that the asymptotic validity² of the corresponding t -test (hereafter the N-test) is lost when $\lambda \preceq 1/\sqrt{k}$ where k must be chosen such that $k \prec n$; the corresponding parametric³ rates are $1/\sqrt{n}$ in both cases. It is our conjecture that similar conclusions obtain if other nonparametric estimators of g are used. The test proposed in this paper is asymptotically valid irrespective of identification strength and uses instruments estimated using k -nearest neighbors.

We now turn to the (asymptotic) power properties of the various tests, still in the context of the above simple model. No test can be consistent when identification is weak. In the parametric³ case this means that consistency requires that $\lambda \succ 1/\sqrt{n}$. We show that our semiparametric test requires that $\lambda \succ 1/\sqrt[4]{nk}$; it is no coincidence

¹See also Staiger and Stock (1997) for the AR-test and Andrews, Moreira, and Stock (2004, 2006) for the M-test.

²A test is asymptotically valid if its actual rejection frequency under the null (i.e. the size) equals the significance level in the limit.

³Both when g is parametrized and estimated and when it is fully known.

that this corresponds to the rate requirement for the semiparametric estimator of θ_0 (using k -nearest-neighbor estimates of g) to be consistent. The power of our test increases more slowly when $1/\sqrt[4]{nk} \prec \lambda \preceq 1/\sqrt{k}$ than does the power of the infeasible version of our test, i.e. an AR-test using the true but unknown g .

When $\lambda \succ 1/\sqrt{k}$, our test is (asymptotically) equivalent to an infeasible AR test for any value of θ (including θ_0). Parametric weak-identification-robust tests generally do not have this desirable property even if g is correctly specified.⁴ Our test is further equivalent to the N-test in terms of its (asymptotic local) power properties and hence is at least as good in terms of local power as any test using parametric g .

The asymptotic distribution of our test statistic under the null hypothesis is standard normal, as opposed to a chi-square for the AR- and K-tests and a nonstandard distribution for the M-test. Consequently, in contrast to any of the other weak identification-robust procedures, testing one-sided alternatives with our test is straightforward. Andrews, Moreira, and Stock (2004) point out that no one-sided version of the K-test is consistent and that the one-sided modification of the M test does better in terms of power than currently available competitors when errors are normal and g is indeed linear in z ; this feature is naturally lost when g is nonlinear and the same considerations apply as in the two-sided case. A further advantage of the normal limit distribution is that the usual critical values apply, whereas with the M-test critical values must be simulated and the simulation procedure is different for the one-sided case than it is for the two-sided version.

The above discussion and comparisons continue to be valid if additional regressors (endogenous or exogenous) and corresponding coefficients (nuisance parameters) β_0 enter linearly into the structural equation, provided that such nuisance parameters are identified if θ_0 is known. The presence of heteroskedasticity of unknown form does not affect the asymptotic validity or consistency of our test. However, since g_i is then no longer the optimal instrument, the optimal local power properties are lost. Likewise, the M-test is then no longer necessarily optimal, either. When nuisance parameters β_0 enter nonlinearly into the structural equation but are consistently estimable if θ_0 is known, then the M-test can no longer be used but otherwise the same considerations apply as above.

The discussion becomes more complicated when the conditional moment condition

⁴Parametric procedures that do have this property can be constructed but such tests — like our test — would give up power/lose consistency in a range of λ -rates close to $1/\sqrt{n}$.

is nonlinear. We discuss the homoskedastic case here noting that our test works when there is heteroskedasticity but that the same caveats apply as in the linear case. The presence of nuisance parameters, provided that they are estimable if θ_0 is known, does not materially affect the conclusions, so we explore the case without nuisance parameters. The M-test cannot deal with any nonlinearities so we exclude it from our discussion. Both our test and the AR- and K-tests continue to be asymptotically valid and the tradeoffs vis-a-vis local power are the same as before; our test has local power that is at least as good as that of parametric tests if $\lambda \succ 1/\sqrt{k}$. The requirements for consistency of our test are largely the same as in the linear case. But an implicit requirement for the consistency of a test is identification of the parameter being tested. Even absent concerns about the strength of one's instruments, identification can be a thorny issue in models with nonlinear moment conditions. For the discussion of consistency, we take identification of θ_0 as given.

In section 2 we demonstrate the limitations of the standard semiparametric t -test and establish the asymptotic validity of our own test. The consistency and power properties mentioned above are presented in section 3. Section 4 extends the results to models with nuisance parameters. We then subject all tests mentioned above to a simulation study (section 5) and the results are as one would expect: the N-test has poor size properties, the parametric tests do better when g is linear and our test rules the roost when g is nonlinear. Section 6 concludes.

2 Validity

Before establishing the asymptotic validity of our test in models with conditional moment conditions we first show that a semiparametric estimator $\hat{\theta}$ of θ_0 in the simple model used in the introduction is inconsistent and the corresponding t -test is asymptotically invalid when instruments are weak. A consequence of the example is that the standard first order asymptotics under strong identification can lead to misleading results and that the problem of weak instruments applies a fortiori in the case in which instruments are estimated nonparametrically. Although our example is simple, we note that the problem applies generally.

Example I *Recall the simple model from the introduction with the symbols introduced therein. Let $\hat{\theta}_I = \sum_i g_i y_i / \sum_i g_i Y_i$ and $\hat{\theta} = \sum_i \hat{g}_i y_i / \sum_i \hat{g}_i Y_i$, where \hat{g}_i is a*

nonparametric k -nearest neighbor estimator with $k \prec n$, so $\hat{\theta}_I$ and $\hat{\theta}$ are the infeasible and feasible semiparametric estimators (see Newey (1990, 1993)), respectively. Then generally

$$\begin{cases} \hat{\theta}_I - \theta_0 & \xrightarrow{d} \Psi_{\tilde{g}u}/(\tilde{C}E\tilde{g}_i^2 + \Psi_{\tilde{g}v}), & \text{if } 0 < \sqrt{n}\lambda \rightarrow \tilde{C} < \infty, \\ \sqrt{n}\lambda(\hat{\theta}_I - \theta_0) & \xrightarrow{d} \Psi_{\tilde{g}u}/E\tilde{g}_i^2, & \text{if } \sqrt{n}\lambda \rightarrow \infty, \end{cases} \quad (1)$$

$$\begin{cases} \hat{\theta} - \theta_0 & \xrightarrow{d} \Psi_{vu}/(\Psi_{vv} + \tilde{C}^2E\tilde{g}_1^2), & \text{if } \sqrt[4]{nk}\lambda \rightarrow \tilde{C} < \infty, \\ \sqrt{nk}\lambda^2(\hat{\theta} - \theta_0) & \xrightarrow{d} \Psi_{vu}/E\tilde{g}_1^2, & \text{if } 1/\sqrt[4]{nk} \prec \lambda \prec 1/\sqrt{k}, \\ \sqrt{n}\lambda(\hat{\theta} - \theta_0) & \xrightarrow{d} (\tilde{C}\Psi_{vu} + \Psi_{\tilde{g}u})/E\tilde{g}_1^2, & \text{if } 1/(\sqrt{k}\lambda) \rightarrow \tilde{C} < \infty, \end{cases} \quad (2)$$

where all Ψ 's together have a joint normal distribution with zero mean and finite variance.

An informal justification for the above results is in appendix F.⁵ \square

The behavior of the infeasible estimator as a function of λ is effectively the same as that of a standard parametric estimator. The behavior of the feasible estimator as a function of λ , however, is highly unusual and the asymptotic distribution of the two estimators coincides only when $\lambda \succ 1/\sqrt{k}$. When $1/\sqrt[4]{nk} \prec \lambda \preceq 1/\sqrt{k}$, the feasible estimator is consistent and asymptotically normal, but its variance is nonstandard.⁶ The feasible estimator loses consistency when $\lambda \preceq 1/\sqrt[4]{nk}$, compared with $\lambda \preceq 1/\sqrt{n}$ for the infeasible version. An estimator which is based on a linearity assumption for g also becomes inconsistent when $\lambda \preceq 1/\sqrt{n}$, but is of course inconsistent at all levels of λ when g_i is in fact orthogonal to z_i . Because the standard asymptotic distribution of $\hat{\theta}$ in example I ceases to be valid sooner (i.e. for larger λ) than with an estimator based on a correctly specified parametric model, correcting the asymptotic distribution is more valuable, also.

We now continue our discussion with the introduction of our test statistic \mathbf{t} . For the rest of this section we assume that the null hypothesis $H_0 : \theta_0 = \theta_H$ is satisfied and show that \mathbf{t} always has a limiting $N(0, 1)$ distribution, provided that some weak conditions are satisfied. We postpone the introduction of nuisance parameters until

⁵The justification uses results established and notation and choices used in (the proof of) theorem 1, so it is better to finish reading this section prior to exploring the justification.

⁶This is similar to the behavior of LIML estimators in linear simultaneous equations models under many instruments asymptotics; see e.g., Bekker (1994).

section 4 and assume that

$$E(m_i(\theta_0)|z_i) = E(m_i|z_i) = 0 \text{ a.s.}$$

is satisfied for some known function m .

As mentioned before, identification strength is measured by λ , which is a non-increasing (with n) sequence of numbers such that

$$M_i(\theta) = E[m_i(\theta)|z_i] = \lambda \tilde{M}_i(\theta) \text{ a.s.}, \quad (3)$$

where \tilde{M} is a function which does not depend on n . We thus follow Stock and Wright (2000) in allowing the entire conditional moment condition to vary with λ . Note that the above construction is artificial; moment conditions do not generally vary with the sample size in real applications. It is just there to study and improve on the properties of standard procedures. Our procedure will be asymptotically valid regardless of the rate at which λ declines and also when λ is fixed. A reasonable expectation is then that if instruments are poor in empirical work, reliable inferences can nonetheless be made with the testing procedure we propose.

The optimal instruments for the estimation of the scalar θ_0 under strong identification and homoskedasticity are given by $g_i(\theta_0)$ where $g_i(\theta) = E[m_{\theta i}(\theta)|z_i]$, where $m_{\theta i}$ is the partial derivative of m_i with respect to θ . The moment condition that is tested in this paper is then

$$E[g_1(\theta_H)m_1(\theta_H)] = 0. \quad (4)$$

Generally, m is nonlinear and g is unknown, but can be estimated by

$$\hat{g}_i(\theta) = \sum_j w_{ij} m_{\theta j}(\theta),$$

where the w_{ij} 's are *nearest neighbor weights*, which are chosen such that

- (i) for each observation i the k observations j that are closest to observation i are assigned a positive weight; all others are assigned a zero weight,
- (ii) in the case of ties positive weights are assigned randomly among the ties,

- (iii) for all i , $\sum_j w_{ij} = 1$,
- (iv) for all i , $w_{ii} = 0$,
- (v) for some fixed C_w^-, C_w , either $w_{ij} = 0$ or $0 < C_w^- \leq kw_{ij} \leq C_w$, and
- (vi) k is chosen such that for some $\alpha > 0$, $n^{1/2+\alpha} \prec k \prec n$.

Since the weights are chosen by us and there are always weights that satisfy the above requirements, the conditions imposed on the weights are innocuous.

Our test statistic is $\mathbf{t}(\theta_H)$ with

$$\mathbf{t}(\theta) = \frac{\sum_i m_i(\theta) \hat{g}_i(\theta)}{\sqrt{\sum_i m_i^2(\theta) \hat{g}_i^2(\theta) - n^{-1} (\sum_i m_i(\theta) \hat{g}_i(\theta))^2 + \sum_{ij} w_{ij} w_{ji} m_i(\theta) m_{\theta_i}(\theta) m_j(\theta) m_{\theta_j}(\theta)}}. \quad (5)$$

An unusual feature of the test statistic is the additional nonparametric correction term in the denominator which is asymptotically irrelevant if identification is strong. Note that the numerator of $\mathbf{t}(\theta)$ divided by n is an estimator of $E[g_1(\theta)m_1(\theta)]$.

Let $v_i(\theta) = m_{\theta_i}(\theta) - g_i(\theta)$. We make the following assumptions.

Assumption A $E[E(m_1^2(\theta)|z_1)E(v_1^2(\theta)|z_1) - \{E(m_1(\theta)v_1(\theta)|z_1)\}^2] > 0$.

If $V(m_1|z_1 = z), V(v_1|z_1 = z)$ are bounded away from zero on the support of z_1 , as is often assumed (see e.g. Robinson, 1987), then assumption **A** is implied by

$$P\left(|\text{Corr}(m_1, v_1|z_1)| = 1\right) < 1,$$

which is unlikely ever to be violated in empirical work.

Assumption B $V[\tilde{g}_1(\theta)m_1(\theta)] > 0$, $E\tilde{M}_1^2(\theta) < \infty$, $E\tilde{g}_1^4 < \infty$, and $E[E(m_1^4(\theta)|z_1)]^4 < \infty$, $E[E(v_1^4(\theta)|z_1)]^4 < \infty$.

In most applications with i.i.d. data, error distributions are not typically fat-tailed and the assumption of the existence of higher moments is then not an issue.

Theorem 1 *If assumptions A–B hold at $\theta = \theta_0$, then $\mathbf{t}(\theta_0) \xrightarrow{d} N(0, 1)$, regardless of the (nonincreasing) λ -sequence.*

Note that theorem 1 is a nonstandard result because it involves both nonparametric weights and an artificial sample-size-dependent sequence λ , both of which play an important role.

Theorem 1 establishes our first and most important desideratum, i.e. asymptotic validity. Given that both our test and the K-, M- and AR-tests are all asymptotically valid we now proceed with a comparison of their power.

3 Power

In this section we discuss both conditions under which our test is consistent and the local power properties of our test.

We begin with a discussion of consistency against any alternatives $\theta \neq \theta_0$ in a set $\Theta \ni \theta_0$. For this we need to discuss the issue of identification since consistency cannot be obtained without identification. Let \tilde{g} be defined by $g = \lambda\tilde{g}$.

Assumption C $\forall \theta \in \Theta : E[\tilde{g}_1(\theta)\tilde{M}_1(\theta)] = 0 \Leftrightarrow \theta = \theta_0$.

Without assumption C, consistency does not even obtain for fixed λ . But since consistency also depends on the rate at which the λ -sequence goes to zero, assumption C is by itself not sufficient. For parametric weak identification-robust tests, consistency requires that $\lambda \succ 1/\sqrt{n}$. In the semiparametric case, $\lambda \succ 1/\sqrt[4]{nk}$ is the best achievable rate. To see why this is so consider again the simple model of the introduction. The numerator of our test statistic for arbitrary θ is then

$$\begin{aligned} \sum_i \hat{g}_i(y_i - Y_i\theta) &= \sum_i \hat{g}_i u_i - \sum_i \hat{g}_i Y_i(\theta - \theta_0) \\ &= \sum_i \hat{g}_i \{u_i - v_i(\theta - \theta_0)\} - \sum_i (\hat{g}_i - g_i)g_i(\theta - \theta_0) + \sum_i g_i^2(\theta - \theta_0). \end{aligned} \quad (6)$$

In the proof of theorem 1 we established that the first right hand side term in (6) is $O_p(\rho)$ with $\rho = \sqrt{n}(\lambda + 1/\sqrt{k})$. The middle right hand side term in (6) can be shown to be asymptotically negligible and the last term is $O_p(n\lambda^2)$. Because the first right hand side term is (after renorming) a mean zero normal random variable in the limit, the last right hand side term must dominate the first for consistency to obtain. This only occurs if $n\lambda^2 \succ \sqrt{n/k}$, i.e. if $\lambda \succ 1/\sqrt[4]{nk}$. This rate is exactly the cutoff rate for consistency of the semiparametric estimator discussed in example I, which is no coincidence; if a parameter cannot be estimated consistently it cannot be tested consistently.

In the parametric case, \hat{g}_i is replaced with $z'_i \hat{\pi}$ and the first right hand side term becomes

$$\hat{\pi}' \sum_i z_i \{u_i - v_i(\theta - \theta_0)\} = O_p(\sqrt{n}\lambda + 1),$$

since $\hat{\pi} = \pi_0 + O_p(n^{-1/2})$, such that consistency requires that $n\lambda^2 \succ \sqrt{n}\lambda + 1$ or $\lambda \succ 1/\sqrt{n}$. So the cause for the discrepancy is the fact that the nonparametric estimator of g_i converges more slowly than does the parametric one. It should be noted that the rate at which k can increase can be chosen to increase at a rate close to n , in which case the difference is minor.

Theorem 2 *Let assumption C hold and let $\lambda \succ 1/\sqrt[4]{nk}$. Then for all $\theta \in \Theta \setminus \{\theta_0\}$ for which assumptions A–B are satisfied,*

(i) $\forall C < \infty : \lim_{n \rightarrow \infty} P[|\mathbf{t}(\theta)| > C] = 1,$

(ii) *If $E[\tilde{g}_1(\theta)\tilde{M}_1(\theta)]$ is a continuous function of θ and Θ is an interval, then (i) is also true for the one-sided version of our test, and*

(iii) *if $\lambda \succ 1/\sqrt{k}$ then*

$$\mathbf{t}(\theta) = \sqrt{n}\lambda \frac{E[\tilde{g}_1(\theta)\tilde{M}_1(\theta)]}{\sqrt{V[\tilde{g}_1(\theta)m_1(\theta)]}} + o_p(\sqrt{n}\lambda). \quad (7)$$

Part (ii) is intuitive. If the expectation in (ii) is continuous and zero at θ_0 then it is positive for all $\theta > \theta_0$ and negative for all $\theta < \theta_0$ in view of assumption C.

Part (iii) of theorem 2 implies that if $\lambda \succ 1/\sqrt{k}$ then for all $\theta \neq \theta_0$, $\mathbf{t}(\theta)$ is asymptotically equivalent to the infeasible AR-test. It also allows us to make some inferences about the local power of our test.⁷ If one considers a sequence $\theta_H = \theta_n = \theta_0 + \Delta/(\sqrt{n}\lambda)$, then

$$\frac{E[\tilde{g}_1(\theta_n)\tilde{M}_1(\theta_n)]}{\sqrt{V[\tilde{g}_1(\theta_n)m_1(\theta_n)]}} \simeq \frac{E[\tilde{g}_1^2(\theta_0)]}{\sqrt{V[\tilde{g}_1(\theta_0)m_1(\theta_0)]}}(\theta_n - \theta_0),$$

⁷We do not formally derive the local power since, given adequate differentiability conditions, it entails simple albeit messy repetitions of arguments similar to those made for the above two theorems.

and one would expect that

$$\mathbf{t} \xrightarrow{d} N \left(\frac{\Delta E[\tilde{g}_1^2(\theta_0)]}{\sqrt{V[\tilde{g}_1(\theta_0)m_1(\theta_0)]}}, 1 \right). \quad (8)$$

The rate at which the local alternative approximates the null is $1/(\sqrt{n}\lambda)$, not the usual $1/\sqrt{n}$ -rate. The $1/(\sqrt{n}\lambda)$ -rate is not specific to our test but applies to all tests under weak identification; see e.g. Andrews, Moreira and Stock (2004, 2006). It arises from the fact that the mean of the asymptotic distribution if instruments g_i are used is $(\theta_n - \theta_0)Eg_1^2/\sqrt{V(g_1m_1)}$ which is of order $(\theta_n - \theta_0)\sqrt{n}\lambda$. Further, the mean in (8) is exactly minus the mean of the limiting distribution of the infeasible version of the N-test under local alternatives and (under homoskedasticity) equals or exceeds that of any other choice of instruments. Finally, away from θ_0 the power of the proposed test — which equals the power of the infeasible AR test — could be greater than or less than that of the infeasible version of the N-test since the term corresponding to the first right hand side term in (7) for the N-test is $\sqrt{n}\lambda(\theta_0 - \theta)E[\tilde{g}_1^2(\theta_0)]/\sqrt{V[\tilde{g}_1(\theta_0)m_1(\theta_0)]}$. Even in a linear model with homoskedasticity the two are not the same.

4 Nuisance Parameters

The main results of this paper were presented in a model in which the structural equation has only a single unknown coefficient. In realistic models, the number of parameters is much greater and we now extend our analysis to the case that m_i is a function of both θ and a vector of nuisance parameters β . Under weak additional conditions all results derived earlier go through in the presence of nuisance parameters provided that $\beta(\theta)$ can be estimated.

We define $\beta(\theta)$ as

$$\beta(\theta) = \underset{\beta}{\operatorname{argmin}} E[E\{m_1(\theta, \beta)|z_1\}]^2. \quad (9)$$

Other definitions of $\beta(\theta)$ are conceivable and it is possible that (9) does not identify $\beta(\theta)$. In linear models this would occur when the number of good instruments is less than the total number of regressors minus one. If there is only one endogenous

regressor with coefficient θ then (9) requires the absence of multicollinearity among the exogenous regressors. With multiple endogenous regressors the assumption of the availability of good instruments can be unrealistic and goes against the spirit of the weak identification literature, but there are sufficiently many applications in which it is warranted to allow for this possibility here. For nonlinear models, conditions under which $\beta(\theta)$ is uniquely defined are discussed at length in the GMM literature.

Let $h_i(\theta) = h_i(\theta, \beta(\theta))$ (and similarly for all other symbols that depend on both θ and β) and $h_i(\theta, \beta) = E[m_{\beta i}(\theta, \beta)|z_i]$. Then (9) leads to the moment condition

$$E[h_1(\theta, \beta(\theta))m_1(\theta, \beta(\theta))] = 0. \quad (10)$$

Note that (10) is natural since $h_i(\theta_0)$ is exactly the vector of optimal instruments under homoskedasticity. In fact, using h_i as instruments is necessary to maintain equality of the local power of our test and the N-test.

We propose to estimate $\beta(\theta)$ on the basis of the moment condition (10), which requires its identification.

Assumption D *For all $\theta \in \Theta$ there is a unique $\beta \in \mathcal{B}$, with \mathcal{B} compact, such that (10) is satisfied. Further, at this unique $\beta = \beta(\theta)$ the Jacobian $Q = Q(\theta) = E[h_1(\theta)h_1'(\theta) + M_1(\theta)m_{\beta\beta_1}(\theta)]$ is invertible.*

Note that the invertibility requirement is automatically fulfilled in sufficiently large samples if $\lambda \prec 1$ provided that $E[h_1(\theta)h_1'(\theta)] > 0$ for all $\theta \in \Theta$.

There are situations in which h is known, e.g. when the structural equation is linear, in which case the nuisance parameters can be estimated parametrically. Here we focus on the more challenging case in which h_i itself is estimated. We define $\hat{\beta}(\theta)$ as a solution to

$$\sum_i \hat{h}_i(\theta, \hat{\beta}(\theta))m_i(\theta, \hat{\beta}(\theta)) = 0, \quad (11)$$

where \hat{h} is a k -nearest neighbor estimator of h . Under conditions to be outlined below, we will show that $\hat{\beta}(\theta)$ is a consistent estimator of $\beta(\theta)$. In fact, under the null hypothesis $\hat{\beta}(\theta) = \hat{\beta}(\theta_0)$ is a \sqrt{n} -consistent estimator achieving the semiparametric efficiency bound for the estimation of $\beta(\theta_0)$. Under the alternative hypothesis, however, $\hat{\beta}(\theta)$ generally converges at a rate slower than \sqrt{n} . This is innocuous since it converges fast enough to ensure that the consistency properties of our test statistic are not affected provided that $\lambda \succ 1/\sqrt[4]{nk}$, as was assumed in theorem 2. The

discrepancy in convergence rates under the null and alternative hypotheses is due to the fact that $M_1(\theta_0) = E[m_1(\theta_0, \beta(\theta_0)) | z_1] = 0$ a.s. but $M_1(\theta)$ is not generally zero for other values of θ .⁸ Since we maintain the setup of (3) here, the only exception is when $\lambda = 0$.

The moment condition we tested previously, i.e. (4), can continue to be used here, albeit that now $g_i(\theta) = g_i(\theta, \beta(\theta))$ and $m_i(\theta) = m_i(\theta, \beta(\theta))$. However, it is more convenient to replace (4) with an equivalent one from which the effect of the estimation of the nuisance parameters has been isolated, i.e.

$$E[q_1(\theta_0)m_1(\theta_0)] = 0, \quad (12)$$

with $q_i(\theta) = q_i(\theta, \beta(\theta)) = g_i(\theta) - \kappa'(\theta)h_i(\theta)$, where $\kappa(\theta) = \kappa(\theta, \beta(\theta)) = (E[h_1(\theta)h_1'(\theta)])^{-1}E[h_1(\theta)g_1(\theta)]$. Let $\hat{q}_i(\theta) = \hat{g}_i(\theta) - \kappa'(\theta)\hat{h}_i(\theta)$ and further $\hat{q}_i(\theta) = \hat{q}_i(\theta, \hat{\beta}(\theta))$ and similarly for \hat{g} and \hat{h} . Condition (12) has the appealing feature that for $\hat{m}_i(\theta) = m_i(\theta, \hat{\beta}(\theta))$, the difference between $\sum_i \hat{q}_i(\theta)\hat{m}_i(\theta)$ and $\sum_i \hat{q}_i(\theta)m_i(\theta)$ is asymptotically negligible. The transition from g to q is similar to ‘partialing out’ other regressors in a linear regression model.

Our test statistic then becomes

$$\hat{\mathbf{t}}(\theta) = \frac{\sum_i \hat{q}_i(\theta)\hat{m}_i(\theta)}{\sqrt{\sum_i \hat{q}_i^2(\theta)\hat{m}_i^2(\theta) + \sum_{ij} w_{ij}w_{ji}\hat{m}_i(\theta)\hat{m}_{\theta i}(\theta)\hat{m}_j(\theta)\hat{m}_{\theta j}(\theta)}}, \quad (13)$$

where $\hat{m}_{\theta i}(\theta) = m_{\theta i}(\theta, \hat{\beta}(\theta))$. The only difference between \mathbf{t} and $\hat{\mathbf{t}}$ is that we now use q instead of g and that the right hand side quantities in (13) depend on $\hat{\beta}$. All assumptions made earlier will hence now be applied to q instead of g .

Previously g , being M ’s derivative, varied proportional to λ . Now the entire q -function varies with λ . To see this, note that $M_{\theta i}(\theta) = g_i(\theta) + h_i'(\theta)\beta_\theta(\theta)$, and that by the implicit function theorem it follows from (9) that for $Q(\theta) = E[h_1(\theta)h_1'(\theta) + M_1(\theta)m_{\beta\beta 1}(\theta)]$,

$$\begin{aligned} \beta_\theta(\theta) &= -(Q(\theta))^{-1}E[h_1(\theta)g_1(\theta) + M_1(\theta)m_{\beta\theta 1}(\theta)] \\ &= -\kappa(\theta) - (E[h_1(\theta)h_1'(\theta)])^{-1}E[M_1(\theta)(m_{\beta\theta 1}(\theta) - m_{\beta\beta 1}(\theta)\kappa(\theta))] + O(\lambda^2). \end{aligned}$$

⁸ $n^{-1} \sum_i (\hat{h}_i - h_i)m_i = o_p(n^{-1/2})$ if $E(m_i | z_i) = 0$ a.s. but this is not true if h_i is merely orthogonal to m_i .

Hence since

$$M_{\theta_i}(\theta) = q_i(\theta) + h'_i(\theta)[\beta_\theta(\theta) + \kappa(\theta)],$$

q_i varies proportional to λ up to terms of order λ^2 .

We are now in a position to state our assumptions and formulate our nuisance parameter theorem. Most assumptions made previously carry over albeit that they are now made with respect to q instead of g . The main implication of this shift is that assumption **C** requires (i) that there is a unique solution $\beta(\theta)$ to (10), which was assumed in assumption **D**, and (ii) that there is a unique combination (θ_0, β_0) that zeroes both $E[\tilde{g}_1(\theta, \beta)m_1(\theta, \beta)]$ and $E[h_1(\theta, \beta)m_1(\theta, \beta)]$. Therefore, (ii) requires that (θ_0, β_0) are identified for fixed λ if g_i, h_i are used as instruments. The only additional assumptions we make here relate to the smoothness of m .

Assumption E For $f = m, m_\theta, m_\beta, m_{\theta\beta}$ we have

$$E[E(\sup_{\beta \in \mathcal{B}} \|f_1(\theta, \beta)\|^2 | z_1)]^2 < \infty.$$

Further,

$$E[E(\sup_{\beta \in \mathcal{B}} \|m_{\beta\beta_1}(\theta, \beta)\| | z_1)]^2 < \infty.$$

Assumption **E** is strong. It assumes the existence of at least two partial derivatives and moreover assumes that a uniform bound of these derivatives is finite in expectation. The assumption of the existence of derivatives excludes interesting applications such as quantile regression models. There are also nonpathological situations in which assumption **E** would be violated when the derivatives do exist; an example can be found in van der Vaart (1998), pp. 48–49.

Theorem 3 *If assumptions **A–C** hold when g is replaced with q and moreover assumptions **D–E** are satisfied then theorems 1 and 2 hold when g is replaced with q and \mathbf{t} with $\hat{\mathbf{t}}$.*

5 Simulations

We now compare several tests proposed in the literature with ours using simulation experiments. Because not all tests can be used when the structural equation is non-

linear, we use the simple model from the introduction, i.e.

$$\begin{cases} y_i = Y_i\theta_0 + u_i \\ Y_i = g(z_i) + v_i \end{cases}.$$

All eight instruments ($z_i \in \mathbb{R}^8$) are independent standard normals and there is homoskedasticity throughout. We consider the following data generating processes (DGP's). DGP1 was motivated by Heckman (1978).

DGP1: Nonlinear IV:
$$\begin{cases} y_i = Y_i\theta_0 + u_i \\ Y_i = I\{\epsilon_i \leq \frac{1}{2} + \Phi(z_i'\iota)\lambda\} - \frac{1}{2}, \end{cases}$$

where $u_i = 5(\epsilon_i - \frac{1}{2}) + \eta_i$ with ϵ_i and η_i drawn from standard uniform and normal distributions, respectively.⁹ Note that the correlation between u_i and ϵ_i is about 0.8, and hence Y_i is highly endogenous. Y_i is a binary variable taking values $\pm 1/2$ with conditional mean (given z_i)

$$g(z) \equiv E(Y_i|z_i = z) = \Phi(z_i'\iota)\lambda I\{\Phi(z_i'\iota)\lambda \leq 1/2\} + I\{\Phi(z_i'\iota)\lambda > 1/2\}/2,$$

which are the optimal instruments. For sufficiently small λ , we note that

$$g(z) = \Phi(z_i'\iota)\lambda \rightarrow 0 \text{ as } \lambda \rightarrow 0,$$

which causes identification failure. When λ is big, the nonlinearity of g makes using a linear probability model for the estimation of g inefficient.

DGP2: Linear IV:
$$\begin{cases} y_i = Y_i\theta_0 + u_i \\ Y_i = z_i'\iota\lambda + v_i, \end{cases}$$

where $\lambda \geq 0$, and u_i, v_i are drawn from the joint normal distribution with mean, variance, covariance equal to 0, 1, and 0.8, respectively. This is the case where the linear specification for the endogenous regressor is correct, which means that the ‘optimality’ results of Andrews, Moreira, and Stock (2004, 2006) apply.

Recall that the parameter λ indicates the strength of identification; we considered the following values for λ : 0, 0.05, 0.1, 0.3, 0.5, 0.7, 1 and 2. In all cases we test $H_0 : \theta_0 = \theta_H$ versus $H_1 : \theta_0 \neq \theta_H$ (or $H_0 : \theta_0 = \theta_H$ versus $H_1 : \theta_0 < \theta_H$ for one-sided alternatives). The primary goal is for the tests to have correct size properties irrespective of identification strength (as measured by λ), but we also want them to have good power. In all experiments, we use $n = 200$, $k = 70$, and 1,000 replications.

⁹ $\Phi(\cdot)$ is the distribution function of $N(0, 1)$, and ι is a vector of ones.

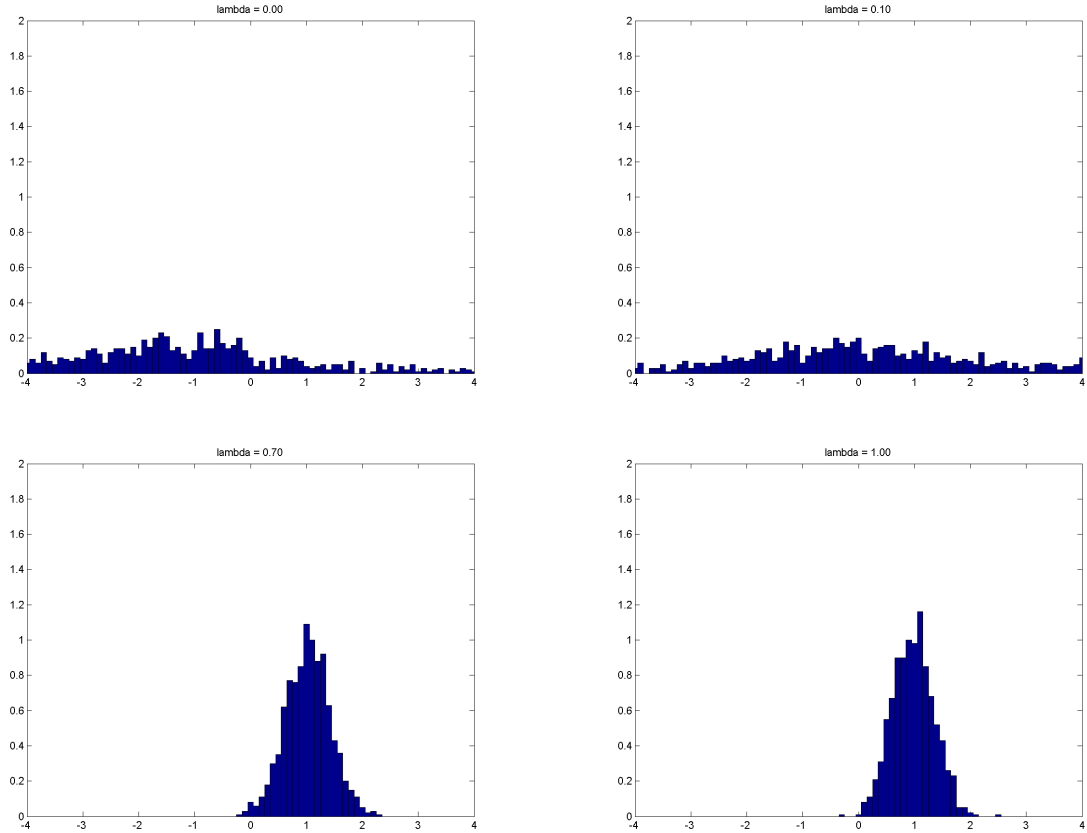


Figure 1: Histograms of the semiparametric estimator $\hat{\theta}$ under DGP1

We now illustrate our earlier finding that the semiparametric estimator $\hat{\theta}$ does not behave as strong identification asymptotics would indicate when λ is small; the results are depicted in figures 1 and 2, which contain histograms of its distribution. For both DGP's, the histograms are not nicely concentrated around the true $\theta_0 = 1$, when the value of λ is small. This is not surprising in view of the inconsistency of $\hat{\theta}$ when $\lambda \leq 1/\sqrt[4]{nk}$.

Figures 3 and 4 show the rejection rates of a number of test statistics as a function of θ_H for various values of λ ; figure 3 for DGP1 and figure 4 for DGP2. In all cases $\theta_0 = 1$ and in all cases except for the N-test we used heteroskedasticity-robust versions of the test statistics.¹⁰ The nominal size of all tests is 0.05. As expected, the t -test based on an efficient point estimator suffers from serious size distortions,

¹⁰The exact forms of the statistics used in the experiments are available upon request. For one-sided M-tests, see e.g., Andrews, Moreira, and Stock (2004).

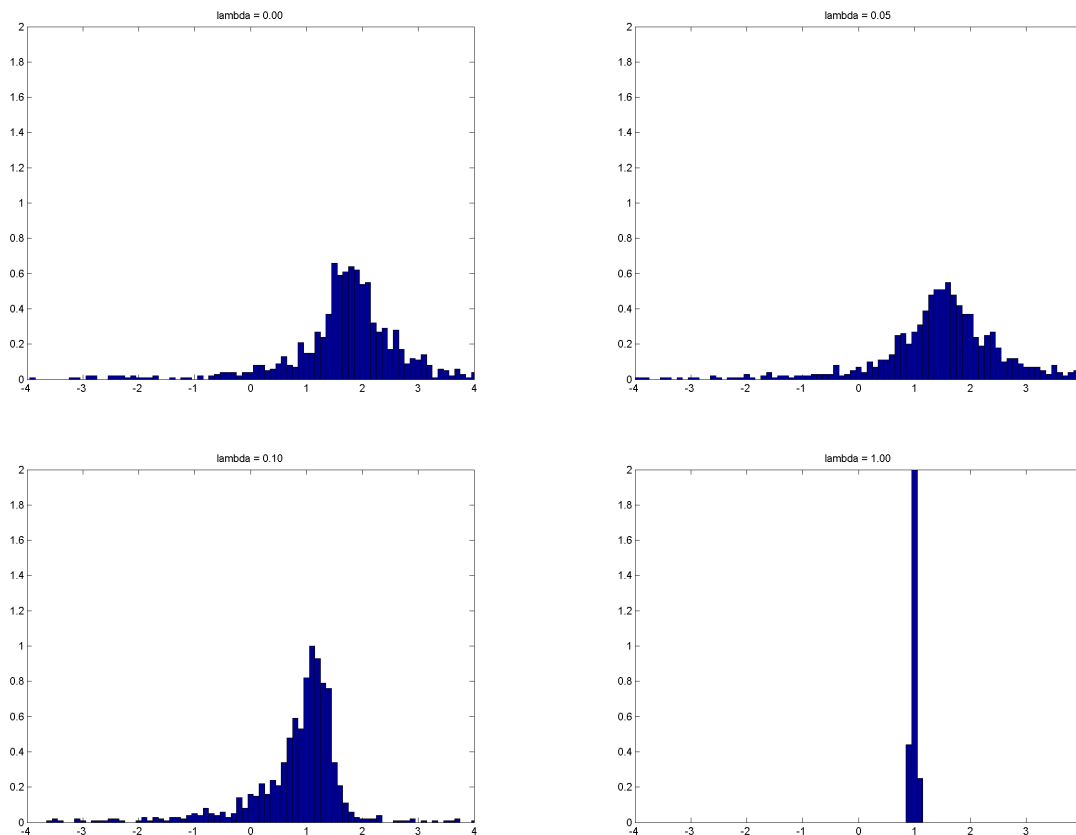


Figure 2: Histograms of the semiparametric estimator $\hat{\theta}$ under DGP2

when the values of λ are small. But note that the power curves of the proposed tests are close to those of the t -tests based on the efficient estimator when λ is relatively large.

The size of the M-test is moderately distorted, which may be attributable to the use of the heteroskedasticity-robust version of the test and would gradually disappear with an increase of the sample size. We feel encouraged that our test (which is also robust to heteroskedasticity) appears to have better size properties in these experiments. As we increase λ , the gain from using nonparametrically estimated instruments becomes apparent under DGP1 and our test is comparable to the N-test in this case.

Figure 4 shows that, again as expected, our test has less power than the M-test in a fully linear specification. When λ increases, however, the power curves of both tests are almost the same, as our theoretical results would indicate. Note that the power

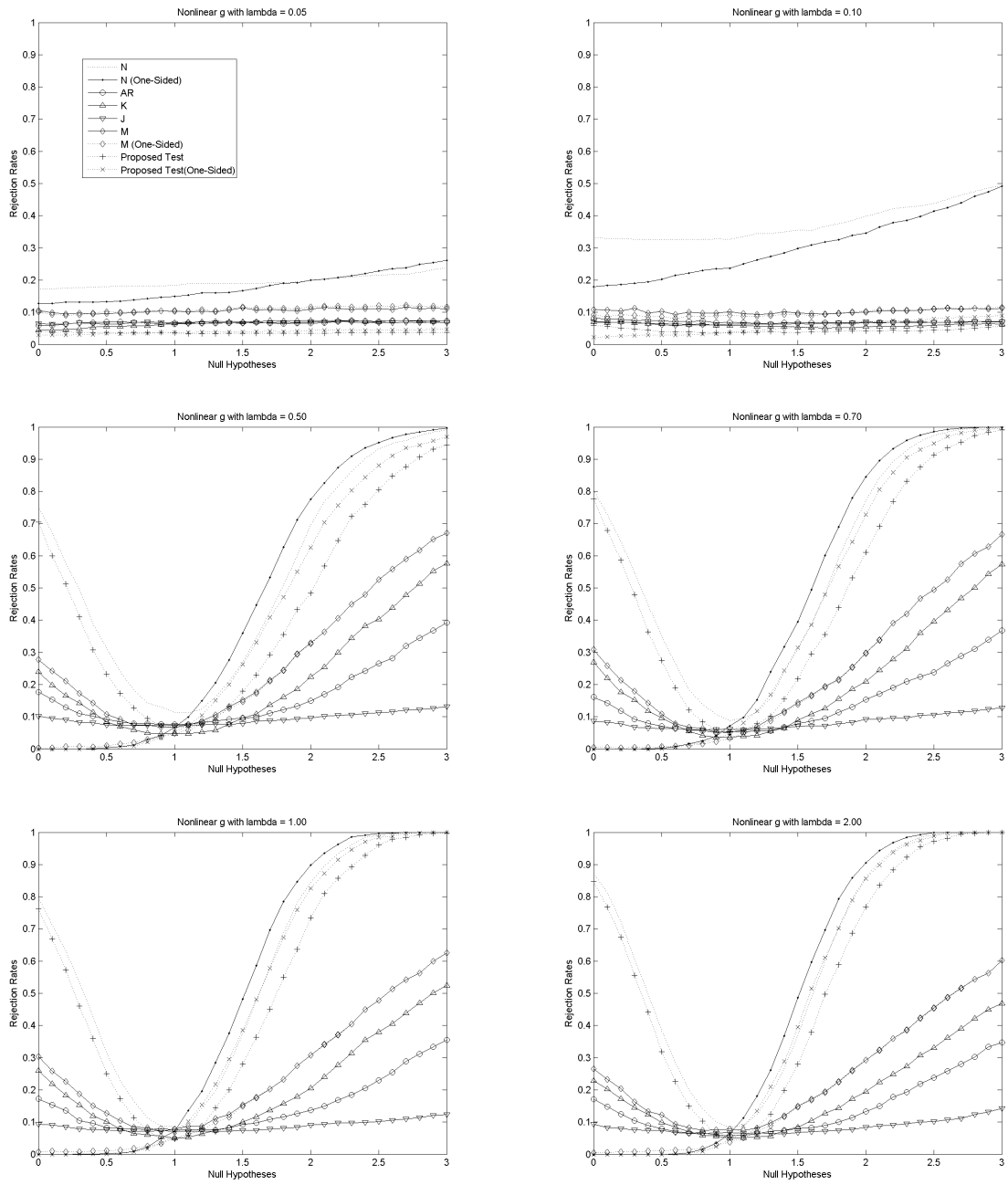


Figure 3: Monte Carlo rejection rates for nonlinear g (DGP1)

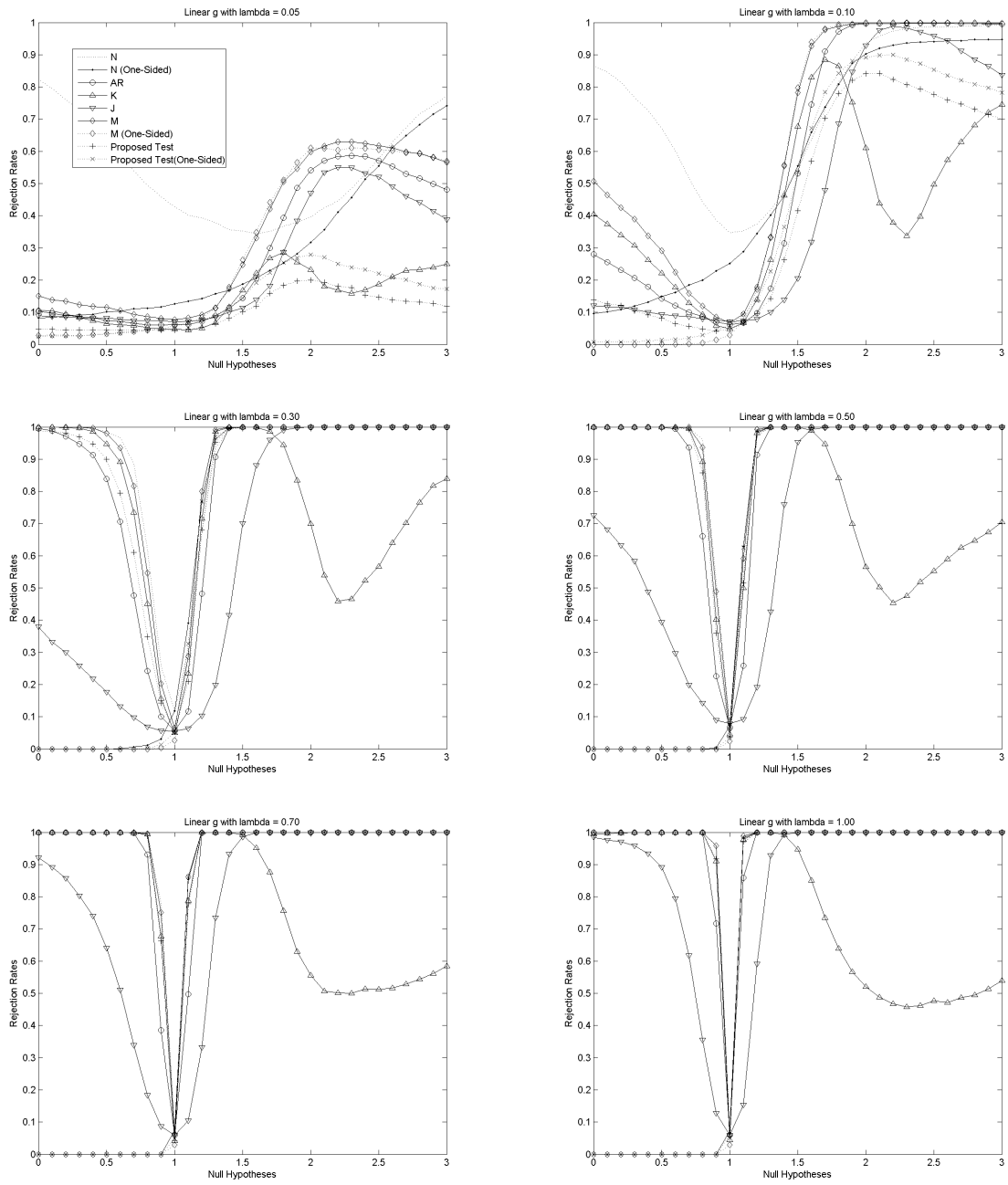


Figure 4: Monte Carlo rejection rates for linear g (DGP2)

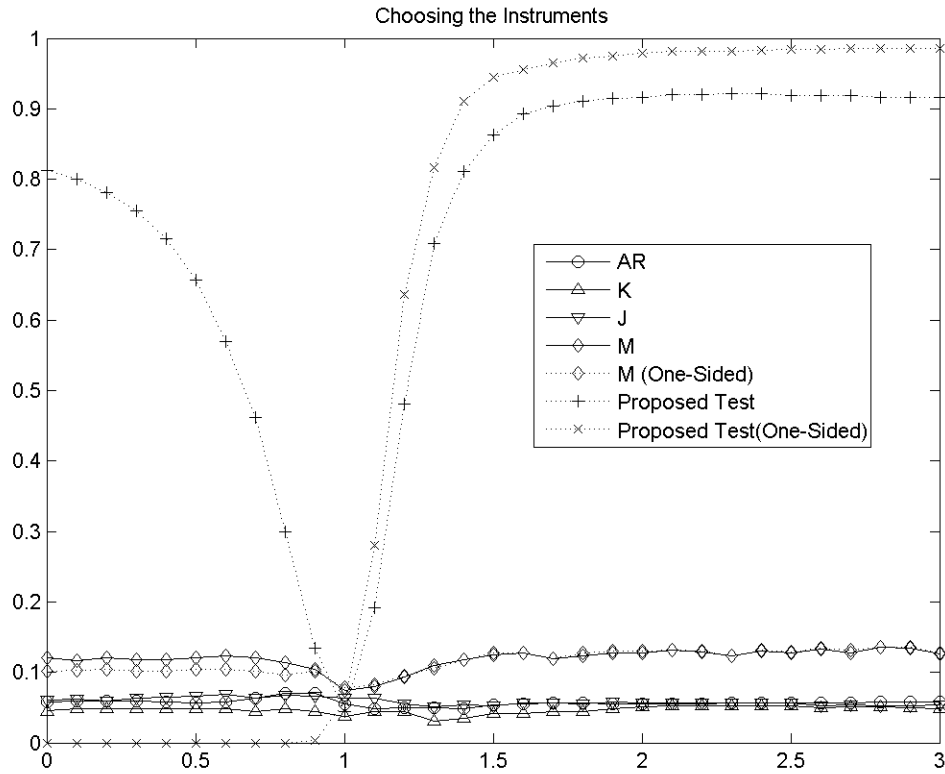


Figure 5: Monte Carlo rejection rates for quadratic g (DGP3)

curves of several tests are decreasing in the right tail and that the power curve of the K-test has a peculiar shape. It is straightforward to show that the AR-test has a (theoretical) nonmonotone power curve and since our test is in effect a semiparametric version of the AR-test, it inherits this feature. As for the K-test, it is well-known that when $\theta_H - \theta_0 = \sigma_{uu}/\sigma_{uv}$ ($= 1/0.8 \approx 1.25$ in DGP2), the K-test suffers from spurious power-declines.

There is any number of nonlinear specifications we could have chosen for DGP1 and depending on our choice the relative performance of the tests considered will vary. We feel that DGP1 is favoring the parametric tests somewhat since g is a function of a linear index. To emphasize this point we now consider an extreme example to illustrate the fact that using arbitrary unconditional moment conditions generated from unconditional moment conditions can lead to poor inference.

$$\mathbf{DGP3: Nonlinear IV 2:} \quad \begin{cases} y_i = Y_i\theta_0 + u_i \\ Y_i = \|z_i\|^2 - 8 + v_i, \end{cases}$$

where u_i, v_i are drawn from the same joint normal distribution as in DGP2. Note that here $E(z_i(y_i - Y_i\theta)) = -E(z_i(\|z_i\|^2 - 8))(\theta - \theta_0) = 0$ for any $\theta \in \mathbb{R}$ because the distribution of the z_i 's is even. Therefore, the unconditional moment condition using linear instruments, $E(z_i(y_i - Y_i\theta))$, does not identify the parameter of interest, while using $g_i = \|z_i\|^2 - 8$ does identify θ_0 . Figure 5 shows that the power curves of the parametric tests are indeed almost flat, while the proposed test performs well.

In summary, parametric tests appear to perform better when g is linear and our test is preferable when g is nonlinear.

6 Conclusions

We have proposed a weak identification–robust test which can be applied in models with conditional moment restrictions. The test has attractive properties and requires weak assumptions, which do however exclude nonsmooth moment conditions and fat-tailed distributions. We have shown it to be asymptotically valid regardless of identification strength yet to have the same local power as the corresponding semi-parametric t–test if identification is sufficiently strong. Our simulation experiments suggest that the test works well in practice.

References Cited

- Anderson, Theodore W., and Herman Rubin (1949), “Estimation of the parameters of a single equation in a complete system of stochastic equations,” *Annals of Mathematical Statistics* 20, 46–63.
- Andrews, Donald W.K., Marcelo J. Moreira, and James H. Stock (2004), “Optimal invariant similar tests for instrumental variables regression,” *Cowles Foundation Discussion Paper* No. 1476.
- Andrews, Donald W.K., Marcelo J. Moreira, and James H. Stock (2006), “Optimal two-sided invariant similar tests for instrumental variables regression,” *Econometrica* 74, 715–752.
- Bekker, Paul A. (1994), “Alternative approximations to the distribution of instrumental variables estimators,” *Econometrica* 62, 657–681.

- Davidson, James (1994), *Stochastic limit theory*, Oxford University Press, Oxford, UK.
- Dufour, Jean-Marie (1997), “Some impossibility theorems in econometrics with applications to structural and dynamic models,” *Econometrica* 65, 1365–1387.
- Heckman, James J. (1978), “Dummy endogenous variables in a simultaneous equations system,” *Econometrica* 46, 931–959.
- Kleibergen, Frank (2002), “Pivotal statistics for testing structural parameters in instrumental variables regression,” *Econometrica* 70, 1781–1083.
- Kleibergen, Frank (2005), “Testing parameters in GMM without assuming that they are identified,” *Econometrica* 73, 1103–1123.
- Moreira, Marcelo J. (2003), “A conditional likelihood ratio test for structural models,” *Econometrica* 71, 1027–1048.
- Newey, Whitney K. (1990), “Efficient instrumental variables estimation of nonlinear models,” *Econometrica* 58, 809–837.
- Newey, Whitney K. (1993), “Efficient estimation of models with conditional moment restrictions,” in vol. 11 of *Handbook of Statistics*, 419–454, North Holland, Amsterdam, Netherlands.
- Robinson, Peter M. (1987), “Asymptotically efficient estimation in the presence of heteroskedasticity of unknown form,” *Econometrica* 55, 875–891.
- Staiger, Douglas, and James H. Stock (1997), “Instrumental variables regression with weak instruments,” *Econometrica* 65, 557–586.
- Stock, James H., and Jonathan H. Wright (2000), “GMM with weak identification,” *Econometrica* 68, 1055–1096.
- Stock, James H., Jonathan H. Wright, and Motohiro Yogo (2002), “A survey of weak instruments and weak identification in generalized method of moments,” *Journal of Business and Economic Statistics* 20, 518–529.
- van der Vaart, Aad W. (1998), *Asymptotic statistics*, Cambridge University Press, Cambridge UK.

A Proofs of Theorems

Throughout the appendix we use the abbreviations RHS and LHS for *right hand side* and *left hand side*, respectively. Further, RHS1 means the first right hand side term, and similarly for RHS2, RHS3, etcetera.

A.1 Proof of theorem 1

Omitting the θ_0 argument, express the test statistic in (5) as \hat{N}/\hat{D} . Let $\rho = \sqrt{n}(\lambda + 1/\sqrt{k})$. We proceed as follows. Define

$$\tilde{D}^2 = nV(g_1 m_1) + \sum_{ij} w_{ij}^2 E(m_i^2 | z_i) E(v_j^2 | z_j) + \sum_{ij} w_{ij} w_{ji} E(m_i v_i | z_i) E(m_j v_j | z_j). \quad (14)$$

Let further $B_i = \zeta_i + \sum_{j < i} \xi_{ij}$ where $\zeta_i = m_i g_i$ and $\xi_{ij} = h_{ij} + h_{ji}$ with $h_{ij} = w_{ij} m_i v_j$.

1. We show that $\hat{N}/\tilde{D} \xrightarrow{d} N(0, 1)$.

(a) We show that $\hat{N} - \tilde{N} = o_p(\rho)$, where $\tilde{N} = \sum_i B_i$, such that $\{B_i/\tilde{D}\}$ is a martingale difference array with respect to the natural filtration.

- i. Noting that $M_i(\theta_0) = 0$ a.s., lemma C1 establishes that $\hat{N} = \sum_i \zeta_i + \sum_{ij} h_{ij}$.
- ii. The result then follows from the fact that $\sum_i \sum_{j \neq i} h_{ij} = \sum_i \sum_{j < i} (h_{ij} + h_{ji})$.

(b) We show that $\tilde{N}/\tilde{D} \xrightarrow{d} N(0, 1)$. For this we establish the conditions of lemma D1

- i. We show that $\max_i |B_i| = o_p(\rho)$; see lemma D2
- ii. We show in lemma D3 that $\sum_i B_i^2 - \tilde{D}^2 = o_p(\rho^2)$.
- iii. $\tilde{D}^{-1} = O_p(\rho^{-1})$ is shown in lemma C11
- iv. We show in lemma D4 that $\sup_n E[\max_{i \leq n} B_i^2 / \tilde{D}^2] < \infty$.
- v. Then requirement (i) follows from 1(b)iv, (ii) from 1(b)ii and 1(b)iii and (iii) from 1(b)i and 1(b)iii.

(c) So $\hat{N}/\tilde{D} = (\hat{N} - \tilde{N})/\tilde{D} + \tilde{N}/\tilde{D} = \tilde{N}/\tilde{D} + o_p(1) \xrightarrow{d} N(0, 1)$.

2. We show that $\hat{D}/\tilde{D} \xrightarrow{p} 1$.

(a) $\hat{D}^2 - \tilde{D}^2 = o_p(\rho^2)$ is established in lemma C8.

(b) $\tilde{D} = O_p(\rho)$ is shown in lemma C10.

(c) So by Slutsky's theorem and 1(b)iii, noting that $\hat{D}, \tilde{D} \geq 0$ by definition, we have $\hat{D}/\tilde{D} - 1 = \sqrt{1 + (\hat{D}^2 - \tilde{D}^2)/\tilde{D}^2} - 1 = o_p(1)$ by Slutsky's theorem.

3. Combining the results established in items 2 and 1 above it follows by Cramér's theorem that $\hat{N}/\hat{D} = (\hat{N}/\tilde{D})/(\hat{D}/\tilde{D}) \xrightarrow{d} N(0, 1)$.

A.2 Proof of theorem 2

Omitting the θ argument, express the test statistic once again as \hat{N}/\hat{D} . We proceed as follows.

1. We show that $\hat{N} = nE(g_1M_1) + o_p(n\lambda^2)$, which is established in lemma C3, noting that $n\lambda^2 \succ \rho$ by assumption.
2. We show in lemma C8 that $\hat{D}^2 = \tilde{D}^2 + o_p(\rho^2)$, such that $\hat{D} - \tilde{D} = o_p(\rho)$; see item 2c of the proof of theorem 1.
3. We show in lemma C10 that $\tilde{D} = O_p(\rho)$.
4. Therefore by Slutsky's theorem for $\theta \neq \theta_0$,

$$\frac{\hat{D}}{\hat{N}} = \frac{\tilde{D} + o_p(\rho)}{nE(g_1M_1) + o_p(n\lambda^2)} = \frac{O_p(\rho/(n\lambda^2))}{E(\tilde{g}_1\tilde{M}_1)} = o_p(1).$$

So part (i) of the theorem holds.

5. For part (ii), note that $E(\tilde{g}_1\tilde{M}_1)$ has the same sign for all $\theta > \theta_0$ and has the opposite sign for all $\theta < \theta_0$ since $E(\tilde{g}_1\tilde{M}_1)$ is a continuous function and cannot be zero except at θ_0 .
6. Finally, part (iii). By lemma C9, $\tilde{D}^2 = nV(g_1m_1) + o_p(n\lambda^2)$, and hence

$$\frac{\hat{N}}{\hat{D}} = \frac{nE(g_1m_1) + o_p(n\lambda^2)}{\sqrt{nV(g_1m_1) + o_p(\sqrt{n}\lambda)}} = \sqrt{n}\lambda \frac{E(\tilde{g}_1\tilde{M}_1)}{\sqrt{V(\tilde{g}_1m_1)}} + o_p(\sqrt{n}\lambda).$$

A.3 Proof of theorem 3

Express the test statistic as \hat{N}/\hat{D} and let \hat{N}, \hat{D} be as in the previous theorems with g_i replaced with q_i . We proceed as follows. The θ_0 -argument is dropped in the list below.

1. First asymptotic validity.
 - (a) Lemma E4 shows that $\hat{\hat{N}} - \hat{N} = o_p(\rho)$.
 - (b) Lemma E8 establishes that $\hat{\hat{D}}^2 - \hat{D}^2 = o_p(\rho^2)$.

(c) Consequently,

$$\frac{\hat{N}}{\hat{D}} = \frac{(\hat{N} - \hat{N})/\rho + \hat{N}/\rho}{(\hat{D} - \hat{D})/\rho + \hat{D}/\rho} = \mathbf{t} + o_p(1).$$

2. For consistency:

(a) Lemma E4 shows that $\hat{N} - \hat{N} = O_p(\rho) + o_p(n\lambda^2)$.

(b) Lemma E8 establishes that $\hat{D}^2 - \hat{D}^2 = o_p(\rho^2)$.

(c) Consequently,

$$\frac{\hat{D}}{\hat{N}} = \frac{\rho}{n\lambda^2} \frac{(\hat{D} - \hat{D})/\rho + \hat{D}/\rho}{(\hat{N} - \hat{N})/(n\lambda^2) + \hat{N}/(n\lambda^2)} = 1/\mathbf{t} + o_p(\rho/(n\lambda^2)).$$

B Technical Lemmas

Lemma B1 *Let $f_i = f(z_i)$ be such that $E\|f_i\|^{p_1 p_2} < \infty$ and $a_i > 0$ be such that $E a_i^{p_2/(p_2-1)} < \infty$ for some $p_2 \geq 1$.¹¹ Then*

$$E\left(w_{ij} a_i |f_i - f_j|^{p_1}\right) = o(n^{-1}). \quad (15)$$

Proof: Note first that by the Jensen inequality,

$$\left(\sum_j w_{ij} |f_i - f_j|^{p_1}\right)^{p_2} \leq \sum_j w_{ij} |f_i - f_j|^{p_1 p_2}. \quad (16)$$

The expectation of the RHS in (16) is $o(1)$ by lemma 1 of Robinson (1987). Hence the LHS in (15) is by the Hölder inequality bounded by

$$\left(E|a_1|^{p_2/(p_2-1)}\right)^{(p_2-1)/p_2} \left[E\left(\sum_j w_{ij} |f_i - f_j|^{p_1}\right)^{p_2}\right]^{1/p_2} = O(1)o(1) = o(1). \quad \square$$

Let $\tau(a, b) = P(\|z_1 - b\| \leq \|a - b\|)$ and $A(a) = \{b : \tau(a, b) \leq 2k/n\}$.

Lemma B2 $\sup_a P[z_1 \in A(a)] = O(k/n)$.

Proof: We establish the result for $z_i \in \mathbb{R}^2$; the case in which z_i is scalar-valued

¹¹ $p_2 = 1$ means that a_i is bounded a.s..

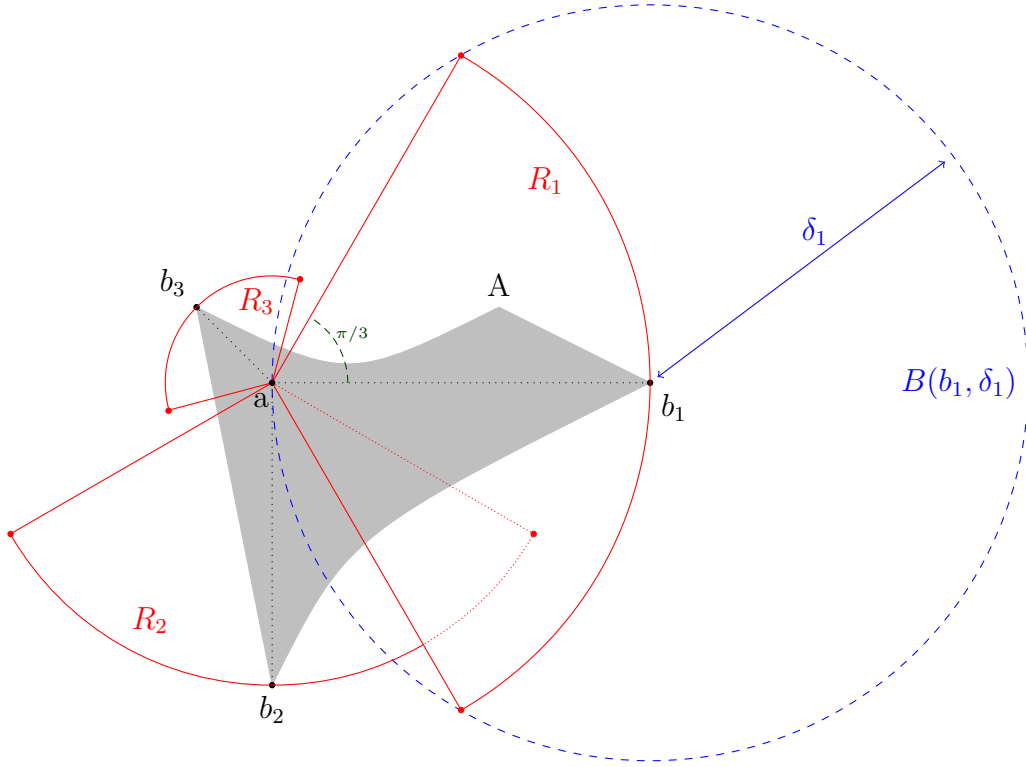


Figure 6: Proof of lemma B2

is simple and the case in which z_i has dimension greater than two is only more complicated in terms of exposition.

Let $A = A(a)$ and $b(t, \omega) = (a_1 + t \cos \omega, a_2 + t \sin \omega)$ and further $\delta_1 = \sup_{\omega, t: b(t, \omega) \in A} t$, ω_1 a corresponding ω and $b_1 = b(\delta_1, \omega_1)$; see figure 6. If $\delta_1 = 0$ there is nothing to prove so suppose that $\delta_1 > 0$. Then let $R_1 = \{b = b(t, \omega) : |\omega - \omega_1| < \pi/3, 0 \leq t \leq \delta_1\}$. So R_1 is a slice of a pie with radius δ_1 extending $\pi/3$ in both directions from ω ; see again figure 6. By construction for all $|\omega - \omega_1| < \pi/3$ and $t \geq 0$: $b(t, \omega) \in A \Rightarrow b(t, \omega) \in R_1$.

Now let $\delta_2 = \sup_{\omega, t: b(t, \omega) \in A \setminus R_1} t$ and define ω_2, b_2, R_2 accordingly. Repeat this procedure until A is covered by at most 10 R_j -sets. By symmetry it suffices to show that $P(z_2 \in R_1) \leq 2k/n$. Now let $A_1 = \{b(t, \omega_1) : 0 \leq t < \delta_1\}$. Then if \mathcal{B} denotes a closed ball and

$$B(b_1, \delta_1) = \bigcup_{b \in A_1} \mathcal{B}(b, \|b - a\|),$$

then because $A_1 \subset A$,

$$P(z_2 \in R_1) \leq P[z_2 \in B(b_1, \delta_1)] = \lim_{t \uparrow \delta_1} P[z_2 \in \mathcal{B}(b(t, \omega_1), t)] \leq \lim_{t \uparrow \delta_1} 2k/n = 2k/n. \quad \square$$

Lemma B3 $\sup_a E(w_{21}w_{31}|z_1 = a) = O(n^{-2})$.

Proof: Let \mathcal{N}_i denote the set of neighbors of i . Since for any a ,

$$E(w_{21}w_{31}|z_1 = a) \leq \frac{C_w^2}{k^2} E[I(z_1 \in \mathcal{N}_2)I(z_1 \in \mathcal{N}_3)|z_1 = a],$$

it suffices to show that $\sup_a E[I(z_1 \in \mathcal{N}_2)I(z_1 \in \mathcal{N}_3)|z_1 = a] = O(k^2/n^2)$. First, suppose that $P(z_1 = a) = c > 0$. Let $S_j^* = \sum_{i \neq j, 1} I(z_i = a)$. Then using the randomization scheme for assigning weights in the case of ties we have

$$\begin{aligned} & E[I(1 \in \mathcal{N}_2)I(1 \in \mathcal{N}_3)|z_1 = a] \\ &= E[I(1 \in \mathcal{N}_2)I(1 \in \mathcal{N}_3)\{I(S_2^* \leq c(n-2)/2) + I(S_2^* > c(n-2)/2)\}|z_1 = a] \\ &\leq P[S_2^* \leq c(n-2)/2] + \frac{2k}{c(n-2)} E[I(1 \in \mathcal{N}_3)|z_1 = a] \\ &\leq P[S_2^* \leq c(n-2)/2] + \frac{2k}{c(n-2)} P[S_3^* \leq c(n-2)/2] + \frac{4k^2}{c^2(n-2)^2}. \end{aligned}$$

Now by the Hoeffding inequality,

$$P(S_2^* \leq c(n-2)/2|z_1 = a) \leq P(|S_2^* - c(n-2)| \geq c(n-2)/2) \leq \exp(-c^2(n-2)/2),$$

which decreases exponentially fast. Now suppose that $P(z_1 = a) = 0$. Let $S_j(a, b) = \sum_{i \neq j, 1} I(\|z_i - b\| \leq \|a - b\|)$ and let τ be as defined prior to lemma B2. Then because

$$\begin{aligned} I(1 \in \mathcal{N}_2) &\leq I(\tau(z_1, z_2) \leq 2k/n) + I(z_1 \in \mathcal{N}_2)I(\tau(z_1, z_2) > 2k/n) \\ &\leq I(\tau(z_1, z_2) \leq 2k/n) + I(S_2^*(z_1, z_2) < k)I(\tau(z_1, z_2) > 2k/n) \\ &\leq I(\tau(z_1, z_2) \leq 2k/n) + I(|S_2(z_1, z_2) - (n-2)\tau(z_1, z_2)| > k), \end{aligned}$$

we have

$$E[I(1 \in \mathcal{N}_2)I(1 \in \mathcal{N}_3)|z_1 = a] \leq E[I(\tau(a, z_2) \leq 2k/n)I(\tau(a, z_3) \leq 2k/n)|z_1 = a] \\ + P[|S_2(a, z_2) - (n-2)\tau(a, z_2)| > k] + P[|S_3(a, z_3) - (n-2)\tau(a, z_3)| > k]. \quad (17)$$

RHS1 in (17) is $\sup_a [P(z_1 \in A(a))]^2 = O(k^2/n^2)$ by lemma B2. RHS2 and RHS3 in (17) are the same and are by the Hoeffding inequality bounded by $\exp(-2k^2/(n-2))$, which goes to zero at an exponential rate. \square

Lemma B4 *Suppose that for some $p \geq 1$, $E|E(b_1|z_1)|^p < \infty$ and $E|E(a_1|z_1)|^{p/(p-1)} < \infty$. Then (i) $E(w_{12}a_1b_2) = n^{-1}E[a_1E(b_1|z_1)] + o(n^{-1}) = O(n^{-1})$, for any $1 < p < \infty$, (ii) $E(w_{12}^*a_1b_2) = O(n^{-1}k^{1-p^*})$. If $E|E(a_1|z_1)|^{2p} < \infty$, $E|E(b_1|z_1)|^{2p} < \infty$ and $E|E(c_1|z_1)|^{p/(p-1)} < \infty$ for some $p \geq 1$ then*

$$(iii) \ E(w_{12}w_{13}c_1a_2b_3) = O(n^{-2}),$$

$$(iv) \ \sum_t E(w_{12}w_{1t}c_1a_2b_t) = O(n^{-1}). \text{ Further, if for some } p_1 > 4 \text{ and } p_2 \geq 1,$$

$$E|E(a_1|z_1)|^{p_1p_2} < \infty, \ E|E(b_2|z_2)|^{p_1p_2/(p_2-1)} < \infty \text{ and } E|c_1| < \infty, \text{ then}$$

$$(v) \ E(w_{21}w_{31}c_1a_2b_3) = o(n^{-3/2}k^{-1/2}), \text{ and}$$

$$(vi) \ \sum_t E(w_{21}w_{t1}c_1a_2b_t) = o(1/\sqrt{nk}).$$

Proof: First (i). Note that

$$E(w_{12}a_1b_2) = E(w_{12}E(a_1|z_1)E(b_2|z_2)) \\ = E(w_{12}E(a_1|z_1)E(b_1|z_1)) + E[w_{12}E(a_1|z_1)\{E(b_2|z_2) - E(b_1|z_1)\}] \\ = n^{-1}E[E(a_1|z_1)E(b_1|z_1)] + o(n^{-1}),$$

by lemma B1. (ii) follows from (i) noting that $w_{12}^{p^*-1} \leq C_w^{p^*-1}k^{1-p^*}$ by construction.

For (iii), note that

$$|E(w_{12}w_{13}c_1a_2b_3)| \leq E[w_{12}w_{13}|c_1|(a_2^2 + b_3^2)] \\ = n^{-1}E[w_{12}|c_1|a_2^2] + n^{-1}E[w_{12}|c_1|b_3^2] = O(n^{-2}),$$

by (i). For (iv) note that

$$\sum_t E(w_{12}w_{1t}c_1a_2b_t) = E(w_{12}^2c_1a_2b_2) + (n-1)E(w_{12}w_{13}c_1a_2b_3) \\ = O(n^{-1}k^{-1}) + O(n^{-1}) = O(n^{-1}),$$

by parts (ii) and (iii). Further, (v) follows since for $\tilde{a}_i = E(a_i|z_i)$, $\tilde{b}_i = E(b_i|z_i)$,

$$\begin{aligned} |E(w_{21}w_{31}c_1a_2b_3)| &= |E[w_{21}w_{31}c_1\tilde{a}_2\tilde{b}_3]| = |E\{E[w_{21}w_{31}\tilde{a}_2\tilde{b}_3|z_1]c_1\}| \\ &\stackrel{\text{Hölder}}{\leq} E\{(E[(w_{21}w_{31})^{p_1/(p_1-1)}|z_1])^{(p_1-1)/p_1}c_1\}(E|\tilde{a}_2\tilde{b}_3|_1^p)^{1/p_1} \\ &\stackrel{\text{B3,Hölder}}{\leq} C_w^2 n^{-2}(n/k)^{2/p_1} E|c_1|(E|\tilde{a}_2|^{p_1 p_2})^{1/(p_1 p_2)}(E|\tilde{b}_3|^{p_1 p_2/(p_2-1)})^{(p_2-1)/(p_1 p_2)} \prec n^{-3/2}k^{-1/2}. \end{aligned}$$

where the first equality follows from lemma B3 and the second from part (i). Finally, (vi) follows from (ii) and (v). \square

Lemma B5 *Let a_i, b_i be such that for some $p \geq 1$, $E|E(b_i|z_1)|^p < \infty$ and $E|E(a_1|z_1)|^{p/(p-1)} < \infty$ and let p_1, p_2 be such that $\min(p_1, p_2) \geq 0$ and $\max(p_1, p_2) \geq 1$. Let furthermore $|c_{ij}| \leq 1$ be constants. Then*

- (i) *If $E(a_i|z_i) = E(b_i|z_i) = 0$ a.s. then $\sum_{ij} w_{ij}^{p_1} w_{ji}^{p_2} c_{ij} a_i b_j = O_p(n^{1/2} k^{1/2 - p_1 - p_2})$,*
 - (ii) *If $E(a_i|z_i) = 0$ a.s. and $p_1 \geq 1$ then $\sum_{ij} w_{ij}^{p_1} w_{ji}^{p_2} c_{ij} a_i b_j = O_p(n^{1/2} k^{1 - p_1 - p_2})$,*
 - (iii) *if $\min(p_1, p_2) \geq 1$ then $\sum_{ij} w_{ij}^{p_1} w_{ji}^{p_2} c_{ij} [a_i b_j - E(a_i|z_i)E(b_j|z_j)] = O_p(n^{1/2} k^{1 - p_1 - p_2})$,*
- If furthermore, for some $p_3 > 8$, $E|E(b_i|z_i)|^{p_3} < \infty$ then*
- (iv) *if $E(a_i|z_i) = 0$ a.s. then $\sum_{ij} w_{ij}^{p_2} c_{ij} a_i b_j = o_p(n^{3/4} k^{3/4 - p_2})$,*
 - (v) $\sum_{ij} w_{ij}^{p_2} c_{ij} [a_i b_j - E(a_i|z_i)E(b_j|z_j)] = o_p(n^{3/4} k^{3/4 - p_2})$,
 - (vi) $\sum_{ij} w_{ij}^{p_1} w_{ji}^{p_2} a_i b_j = \sum_{ij} w_{ij}^{p_1} w_{ji}^{p_2} E(a_i|z_i)E(b_i|z_i) + o_p(nk^{1 - p_1 - p_2})$.

Proof: For (i), suppose that $p_1 \geq 1$ (the case $p_2 \geq 1$ is symmetric), square and take expectations to obtain

$$\begin{aligned} \sum_{ij} c_{ij} E[w_{ij}^{2p_1} w_{ji}^{2p_2} a_i^2 b_j^2 + (w_{ij} w_{ji})^{p_1 + p_2} a_i b_i a_j b_j] \\ \leq (C_w/k)^{2p_1 + 2p_2 - 1} \sum_{ij} w_{ij} (a_i^2 b_j^2 + |a_i b_i a_j b_j|) \stackrel{\text{B4}}{=} O(nk^{1 - 2p_1 - 2p_2}). \end{aligned}$$

For (ii) likewise square and take expectations which yields

$$\begin{aligned} \sum_{ij,t} c_{ij} c_{it} E[(w_{ij} w_{it})^{p_1} (w_{ji} w_{ti})^{p_2} a_i^2 b_j b_t] \\ \leq (C_w/k)^{2p_1 + 2p_2 - 2} \sum_{ij,t} E(w_{ij} w_{it} a_i^2 b_j b_t) \stackrel{\text{B4}}{=} O(nk^{2(1 - p_1 - p_2)}). \end{aligned}$$

For (iii) write $a_i b_j - E(a_i|z_i)E(b_j|z_j) = (a_i - E(a_i|z_i))(b_j - E(b_i|z_i)) + E(a_i|z_i)(b_i -$

$E(b_i|z_i)) + (a_i - E(a_i|z_i))E(b_i|z_i)$ and apply parts (i) and (ii). For (iv) we get

$$\sum_{ijt} c_{ij} c_{it} E[(w_{ji} w_{ti})^{p_2} a_i^2 b_j b_t] \leq (C_w/k)^{2p_2-2} \sum_{ijt} E(w_{ji} w_{ti} a_i^2 b_j b_t) \stackrel{\text{B4}}{=} o(n^{3/2} k^{3/2-2p_2}).$$

For (v) use the same expansion as for (iii) but apply parts (i) and (iv). Finally, for (vi) use either (iii) or (v) and (supposing $p_1 \geq 1$; $p_2 \geq 1$ is similar)

$$\begin{aligned} & \left| \sum_{ij} w_{ij}^{p_1} w_{ji}^{p_2} E(a_i|b_i) [E(b_j|z_j) - E(b_i|z_i)] \right| \\ & \leq n^2 (C_w/k)^{p_1+p_2-1} E|w_{12} E(a_1|z_1) [E(b_2|z_2) - E(b_1|z_1)]| \stackrel{\text{B1}}{=} o(nk^{1-p_1-p_2}). \quad \square \end{aligned}$$

C Approximations

Lemma C1 $\sum_i \hat{g}_i(\theta) [m_i(\theta) - M_i(\theta)] = \sum_i g_i(\theta) [m_i(\theta) - M_i(\theta)] + \sum_{ij} w_{ij} v_j(\theta) [m_i(\theta) - M_i(\theta)] + o_p(\sqrt{n}\lambda)$.

Proof: Since $\hat{g}_i = \sum_j w_{ij} m_{\theta j} = g_i + \sum_j w_{ij} (g_j - g_i) + \sum_j w_{ij} v_j$, we consider (omitting the θ argument)

$$\begin{aligned} E \left[\sum_{ij} w_{ij} (g_j - g_i) (m_i - M_i) \right]^2 &= \sum_{ijs} E [w_{ij} w_{is} (g_j - g_i) (g_s - g_i) (m_i - M_i)^2] \\ &\leq 2^{-1} \sum_{ijs} E [w_{ij} w_{is} ((g_j - g_i)^2 + (g_s - g_i)^2) (m_i - M_i)^2] \\ &= \sum_{ij} E [w_{ij} (g_j - g_i)^2 (m_i - M_i)^2] \stackrel{\text{B1}}{=} o(n\lambda^2). \quad \square \end{aligned}$$

Lemma C2 $\sum_i \hat{g}_i(\theta) M_i(\theta) = nE[g_1(\theta) M_1(\theta)] + o_p(n\lambda^2 + \rho)$.

Proof: We have (omitting the θ argument),

$$\sum_i \hat{g}_i M_i - nE[g_i M_i] = \sum_i (\hat{g}_i - g_i) M_i + \left(\sum_i g_i M_i - nE[g_i M_i] \right). \quad (18)$$

Square RHS2 in (18) and take expectations to obtain $\sum_i V(g_i M_i) = O(n\lambda^4)$. So RHS2 in (18) is $O_p(\sqrt{n}\lambda^2)$.

RHS1 in (18) can be written as

$$\sum_{ij} w_{ij}(g_j - g_i)M_i + \sum_{ij} w_{ij}v_jM_i. \quad (19)$$

Squaring the second term in (19) and taking expectations yields

$$\sum_{ijt} E(w_{ij}w_{tj}v_j^2M_iM_t) \stackrel{\text{B4}}{=} o(n^{3/2}k^{-1/2}\lambda^2) = o(n^2\lambda^4 + \rho^2).$$

For the first term in (19) we get

$$\sum_{ij} E|w_{ij}(g_j - g_i)M_i| \stackrel{\text{B1}}{=} o(n\lambda^2). \quad \square$$

Lemma C3 $\sum_i \hat{g}_i(\theta)m_i(\theta) - nE[g_1(\theta)m_1(\theta)] = o_p(n\lambda^2) + O_p(\rho)$.

Proof: The LHS is (omitting θ)

$$\begin{aligned} \sum_i \hat{g}_i(m_i - M_i) + \left[\sum_i \hat{g}_iM_i - nE(g_1M_1) \right] \\ \stackrel{\text{C1,C2}}{=} \sum_i g_i(m_i - M_i) + \sum_{ij} w_{ij}(m_i - M_i)v_j + o_p(n\lambda^2 + \rho). \end{aligned}$$

Now,

$$\begin{aligned} E\left(\sum_i g_i(m_i - M_i)\right)^2 &= \sum_i E[g_1^2(m_1 - M_1)^2] = O(n\lambda^2), \\ \sum_{ij} w_{ij}(m_i - M_i)v_j &\stackrel{\text{B5(i)}}{=} O_p(\sqrt{n/k}). \quad \square \end{aligned}$$

Lemma C4 $\left[\sum_i \hat{g}_i(\theta)m_i(\theta)\right]^2 - (nE[g_1(\theta)m_1(\theta)])^2 = o_p(\rho^4)$.

Proof: Omitting the θ -argument we have (using $a^2 - b^2 = (a - b)^2 + 2b(a - b)$)

$$\begin{aligned} \left| \left(\sum_i \hat{g}_im_i\right)^2 - (nE(g_1m_1))^2 \right| &\stackrel{\text{triangle,Schwarz}}{\leq} \left(\sum_i \hat{g}_im_i - nE(g_1m_1)\right)^2 \\ &+ 2\left| \left(\sum_i \hat{g}_im_i - nE(g_1m_1)\right)nE(g_1m_1) \right| \stackrel{\text{C3}}{=} o_p(n^2\lambda^4) + O_p(n\lambda^2\rho) = o_p(\rho^4). \quad \square \end{aligned}$$

Lemma C5 $\sum_i \hat{g}_i^2(\theta)m_i^2(\theta) = \sum_i g_i^2(\theta)m_i^2(\theta) + \sum_i m_i^2(\theta)(\sum_j w_{ij}v_j(\theta))^2 + o_p(\rho^2)$.

Proof: Omit the θ -argument throughout and let $\hat{g}_i^* = g_i + \sum_j w_{ij}v_j$. We show separately that

$$\sum_i (\hat{g}_i^2 - (\hat{g}_i^*)^2) m_i^2 = o_p(\rho^2), \quad (20)$$

$$\sum_i (\hat{g}_i^* m_i)^2 - \sum_i g_i^2 m_i^2 - \sum_i m_i^2 \left(\sum_j w_{ij} v_j \right)^2 = o_p(\rho^2). \quad (21)$$

First (21). The LHS is $2 \sum_{ij} w_{ij} g_i m_i^2 v_j$. Divide by two, square and take expectations to obtain (using $2|ab| \leq a^2 + b^2$)

$$\sum_{ijt} E(w_{ij} w_{it} g_i m_i^2 g_t m_t^2 v_j^2) \stackrel{\text{B4}}{=} o(n^{3/2} k^{-1/2} \lambda^2) = o(\rho^3).$$

Now (20). We have

$$\begin{aligned} \left| \sum_i (\hat{g}_i^2 - (\hat{g}_i^*)^2) m_i^2 \right| &= \left| 2 \sum_i \hat{g}_i^* (\hat{g}_i - \hat{g}_i^*) m_i^2 + \sum_i (\hat{g}_i^* m_i)^2 \right| \\ &\stackrel{\text{triangle, Schwarz}}{\leq} 2 \sqrt{\sum_i (\hat{g}_i^* m_i)^2} \sqrt{\sum_i (\hat{g}_i - \hat{g}_i^*)^2 m_i^2} + \sum_i (\hat{g}_i - \hat{g}_i^*)^2 m_i^2. \end{aligned}$$

First,

$$\begin{aligned} \sum_i (\hat{g}_i - \hat{g}_i^*)^2 m_i^2 &= \sum_i \left(\sum_j w_{ij} (g_j - g_i) \right)^2 m_i^2 \\ &\stackrel{\text{Schwarz}}{\leq} \sum_{ij} w_{ij} (g_j - g_i)^2 m_i^2 \stackrel{\text{B1}}{=} o_p(n\lambda^2) = o_p(\rho^2). \end{aligned}$$

So it remains to be shown that $\sum_i (\hat{g}_i^* m_i)^2 = O_p(\rho^2)$, which follows by combining (21) with

$$\begin{aligned} \sum_i E(g_i^2 m_i^2) + \sum_i E \left(\sum_j w_{ij} m_i v_j \right)^2 \\ = O(n\lambda^2) + \sum_{ij} E(w_{ij}^2 m_i^2 v_j^2) = O(n\lambda^2 + n/k) = O(\rho^2). \quad \square \quad (22) \end{aligned}$$

Lemma C6 $\sum_i m_i^2(\theta) \left(\sum_j w_{ij} v_j(\theta) \right)^2 = \sum_{ij} w_{ij}^2 E[m_i^2(\theta) | z_i] E[v_j^2(\theta) | z_j] + o_p(\rho^2)$.

Proof: Omitting the θ argument,

$$\begin{aligned} & \sum_i m_i^2 \left(\sum_j w_{ij} v_j \right)^2 = \\ & \sum_{ij} w_{ij}^2 E[m_i^2 | z_i] E[v_j^2 | z_j] + \sum_{ij} w_{ij}^2 (m_i^2 v_j^2 - E[m_i^2 | z_i] E[v_j^2 | z_j]) + \sum_{ij} \sum_{t \neq j} w_{ij} w_{it} m_i^2 v_j v_t. \end{aligned} \quad (23)$$

RHS2 in (23) is $o_p(n^{3/4} k^{-5/4}) = o_p(\rho^2)$ by lemma B5(v). Square RHS3 and take expectations to obtain

$$\begin{aligned} & 2 \sum_{ij} \sum_{t \neq j} E[(w_{ij}^2 w_{it}^2 m_i^4 v_j^2 v_t^2)] + 2 \sum_{ij} \sum_{s \neq i} \sum_{t \neq j} E[w_{ij} w_{it} w_{sj} w_{st} m_i^2 m_s^2 v_j^2 v_t^2] \\ & + 4 \sum_{ij} \sum_{s \neq i} \sum_{t \neq j} E[w_{ij} w_{it} w_{si} w_{st} m_i^2 v_i m_s^2 v_s v_t^2] \stackrel{\text{B4}}{=} o(n^2 k^{-2}) = o(\rho^4). \quad \square \end{aligned}$$

Lemma C7 $\sum_{ij} w_{ij} w_{ji} m_i(\theta) m_{\theta i}(\theta) m_j(\theta) m_{\theta j}(\theta) - \sum_{ij} w_{ij} w_{ji} E[\{m_i(\theta) - M_i(\theta)\} v_i(\theta) | z_i] E[\{m_j(\theta) - M_j(\theta)\} v_j(\theta) | z_j] = o_p(\rho^2)$.

Proof: Write the LHS as (omitting the θ argument)

$$\begin{aligned} & \left[\sum_{ij} w_{ij} w_{ji} m_i m_{\theta i} m_j m_{\theta j} - \sum_{ij} w_{ij} w_{ji} E[m_i m_{\theta i} | z_i] E[m_j m_{\theta j} | z_j] \right] \\ & + 2 \sum_{ij} w_{ij} w_{ji} M_i g_i E[(m_j - M_j) v_j | z_j] + \sum_{ij} w_{ij} w_{ji} M_i g_i M_j g_j. \end{aligned} \quad (24)$$

By lemma B5(iii) the first term in (24) is $O_p(n^{1/2} k^{-1}) = o_p(\rho^2)$. Now, the third term in (24) is $o_p(\rho^2)$ because

$$E \left| \sum_{ij} w_{ij} w_{ji} M_i g_i M_j g_j \right| \leq (n^2 C_w \lambda^4 / k) E |w_{12} \tilde{M}_1 \tilde{g}_1 \tilde{M}_2 \tilde{g}_2| \stackrel{\text{B4}}{=} O(n \lambda^4 / k) = o(\rho^2).$$

And the second term in (24) is $O_p(n \lambda^2 / k) = o_p(\rho^2)$, which can be established in the same way. \square

Lemma C8 $\hat{D}^2(\theta) - \tilde{D}^2(\theta) = o_p(\rho^2)$.

Proof: Omitting the θ -argument,

$$\begin{aligned}
 & \hat{D}^2 - \tilde{D}^2 \\
 &= \underbrace{\left[\sum_i \hat{g}_i^2 m_i^2 - \sum_i g_i^2 m_i^2 - \sum_i m_i^2 \left(\sum_j w_{ij} v_j \right)^2 \right]}_{\stackrel{\text{C5}}{=} o_p(\rho^2)} - \underbrace{\left[n^{-1} \left(\sum_i \hat{g}_i m_i \right)^2 - n \left(E[g_1 M_1] \right)^2 \right]}_{\stackrel{\text{C4}}{=} o_p(\rho^4/n) = o_p(\rho^2)} \\
 &+ \underbrace{\left[\sum_i g_i^2 m_i^2 - n E(g_1^2 m_1^2) \right]}_{= O_p(\sqrt{n} \lambda^2)} + \underbrace{\left[\sum_i m_i^2 \left(\sum_j w_{ij} v_j \right)^2 - \sum_{ij} w_{ij}^2 E(m_i^2 | z_i) E(v_j^2 | z_j) \right]}_{\stackrel{\text{C6}}{=} o_p(\rho^2)} \\
 &+ \underbrace{\left[\sum_{ij} w_{ij} w_{ji} m_i m_{\theta_i} m_j m_{\theta_j} - \sum_{ij} w_{ij} w_{ji} E(m_i v_i | z_i) E(m_j v_j | z_j) \right]}_{\stackrel{\text{C7}}{=} o_p(\rho^2)} = o_p(\rho^2). \quad \square \quad (25)
 \end{aligned}$$

Lemma C9 $\tilde{D}^2(\theta) = nV[g_1(\theta)m_1(\theta)] + O_p(n/k)$.

Proof: Omitting θ ,

$$\begin{aligned}
 E \left[\sum_{ij} w_{ij}^2 E(m_i^2 | z_i) E(v_j^2 | z_j) \right] &\leq (C_w/k) n^2 E(w_{12} m_1^2 v_2^2) \stackrel{\text{B4}}{=} O(n/k), \\
 E \left| \sum_{ij} w_{ij} w_{ji} E(m_i v_i | z_i) E(m_j v_j | z_j) \right| &\leq (C_w/k) n^2 E(w_{12} | m_1 v_1 m_2 v_2) \stackrel{\text{B4}}{=} O(n/k). \quad \square
 \end{aligned}$$

Lemma C10 $\tilde{D}(\theta) = O_p(\rho)$.

Proof: Apply lemma C9 and note that $V[g_1(\theta)m_1(\theta)] = O(n\lambda^2)$, such that $\tilde{D}^2(\theta) = O_p(\rho^2)$. \square

Lemma C11 $\tilde{D}^{-1}(\theta) = O_p(\rho^{-1})$.

Proof: Since for generic $a, b \geq 0$, $1/(a+b) \leq 1/\max(a, b) = \min(1/a, 1/b)$, sufficient conditions are (omitting θ)

$$1/V(\tilde{g}_1 m_1) = O(1), \quad (26)$$

$$1/\left[\sum_{ij} w_{ij}^2 E(m_i^2 | z_i) E(v_i^2 | z_i) - \sum_{ij} w_{ij} w_{ji} E(m_i v_i | z_i) E(m_j v_j | z_j) \right] = O_p(k/n). \quad (27)$$

(26) holds by assumption **B**. Now, since for generic a, b , $ab \leq (a^2 + b^2)/2$,

$$\left| \sum_{ij} w_{ij} w_{ji} E(m_i v_i | z_i) E(m_j v_j | z_j) \right| \leq \sum_{ij} w_{ij}^2 (E(m_i v_i | z_i))^2.$$

Thus, the LHS denominator in (27) is bounded below by

$$\begin{aligned} \sum_{ij} w_{ij}^2 [E(m_i^2 | z_i) E(v_i^2 | z_i) - \{E(m_i v_i | z_i)\}^2] &\geq (C_w^- / k) \sum_i [E(m_i^2 | z_i) E(v_i^2 | z_i) \\ &- \{E(m_i v_i | z_i)\}^2] = (nC_w^- / k) E[E(m_i^2 | z_i) E(v_i^2 | z_i) - \{E(m_i v_i | z_i)\}^2] + O_p(\sqrt{n}/k), \end{aligned}$$

where the last inequality follows from the fact that the sample and population means of i.i.d. samples differ by $O_p(1/\sqrt{n})$. The result stated in (27) then follows from assumption **A**. \square

D Martingale Differences

Lemma D1 *Let $\{X_{ni}, \mathcal{F}_{ni}\}$ be a martingale difference array for which*

(i) $\sup_n E[\max_{i \leq n} X_{ni}^2] < \infty$, (ii) $\sum_{i=1}^n X_{ni}^2 \xrightarrow{p} 1$ and (iii) $\max_{i \leq n} |X_{ni}| \xrightarrow{p} 0$. Then $\sum_{i=1}^n X_{ni} \xrightarrow{d} N(0, 1)$.

Proof: This is Davidson (1994), theorem 24.3, where the only difference is that Davidson's sufficient condition that the unconditional variances sum to one is replaced with condition (i). Since in Davidson's proof of theorem 24.3 the finite unconditional variances requirement is only used to verify (i), the lemma statement holds trivially. \square

Lemma D2 $\max_i |B_i| = o_p(\rho)$.

Proof: Pick any $\epsilon > 0$. Then for any $2 < p \leq 4$ (suppressing θ_0),

$$P(\max_i |B_i| > \epsilon \rho) \leq \sum_i P(|B_i| > \epsilon \rho) \stackrel{\text{Markov}}{\leq} \rho^{-p} \epsilon^{-p} \sum_i E|B_i|^p. \quad (28)$$

So by the Minkowski inequality it suffices to show that $\sum_i E|B_i|^p = o(\rho^p)$ or indeed that

$$nE|\zeta_1|^p = o(\rho^p), \quad \sum_i E \left| \sum_{j < i} h_{ij} \right|^p = o(\rho^p), \quad \sum_i E \left| \sum_{j < i} h_{ji} \right|^p = o(\rho^p). \quad (29)$$

The LHS in the first condition in (29) is $O(n\lambda^p) = o(\rho^p)$. The second and third conditions in (29) are similar; we only show the second. We have for some fixed $C_p < \infty$,

$$\begin{aligned} \sum_i E \left| \sum_{j<i} h_{ij} \right|^p &= \sum_i E \left| m_i \sum_{j<i} w_{ij} v_j \right|^p \stackrel{\text{Burkholder}^{12}}{\leq} C_p \sum_i E \left[m_i^2 \sum_{j<i} w_{ij}^2 v_j^2 \right]^{p/2} \\ &\stackrel{\text{Jensen}}{\leq} C_p n^{p/2+1} E |w_{12} m_1 v_2|^p \stackrel{\text{B4}}{=} O(n^{p/2} k^{1-p}) = o(\rho^p). \quad \square \end{aligned}$$

Lemma D3 $\sum_i B_i^2 - \tilde{D}^2(\theta_0) = o_p(\rho^2)$.

Proof: Noting that $\tilde{D}^2(\theta_0) = nE\zeta_1^2 + \sum_i \sum_{j<i} E(\xi_{ij}^2 | z_i, z_j)$, we have (omitting the θ_0 -argument)

$$\begin{aligned} \sum_i B_i^2 &= \sum_i \zeta_i^2 + 2 \sum_i \zeta_i \eta_i + \sum_i \eta_i^2 \\ &= \tilde{D}^2 + \sum_i (\zeta_i^2 - E\zeta_i^2) + 2 \sum_i \zeta_i \eta_i + \sum_i \sum_{j<i} \{ \xi_{ij}^2 - E(\xi_{ij}^2 | z_i, z_j) \} + 2 \sum_i \sum_{j<i} \sum_{t<j} \xi_{ij} \xi_{it}. \end{aligned} \quad (30)$$

Square RHS2 in (30) and take its expectation to obtain $\sum_i V\zeta_i^2 = O(n\lambda^4) = o(\rho^4)$. RHS3 in (30) is twice

$$\sum_i \sum_{j<i} w_{ij} m_i \zeta_i v_j + \sum_i \sum_{j<i} w_{ji} \zeta_i v_i m_j \stackrel{\text{B5(ii),(iv)}}{=} o_p(n^{3/4} k^{-1/4} \lambda) + O_p(n^{1/2} \lambda) = o_p(\rho^2),$$

where the c_{ij} -weights in lemma B5 can be taken as $c_{ij} = I(j < i)$. RHS4 in (30) can be further expanded as

$$\begin{aligned} &\sum_i \sum_{j<i} w_{ij}^2 \{ m_i^2 v_j^2 - E(m_i^2 | z_i) E(v_j^2 | z_j) \} \\ &\quad + 2 \sum_i \sum_{j<i} w_{ij} w_{ji} \{ m_i v_i m_j v_j - E(m_i v_i | z_i) E(m_j v_j | z_j) \} \\ &\quad + \sum_i \sum_{j<i} w_{ji}^2 \{ m_j^2 v_i^2 - E(m_j^2 | z_j) E(v_i^2 | z_i) \} \stackrel{\text{B5(iii)}}{=} O_p(n^{1/2}/k) = o_p(\rho^2). \end{aligned}$$

¹²See theorem 15.18 of Davidson (1994) and comments following it.

Finally, square RHS5 in (30) and take expectations to obtain

$$\begin{aligned}
\sum_i \sum_{j<i} \sum_{t<j} E(\xi_{ij}^2 \xi_{it}^2) &= \sum_i \sum_{j<i} \sum_{t<j} [E(w_{ij}^2 w_{it}^2 m_i^4 v_j^2 v_t^2) + E(w_{ij}^2 w_{it}^2 m_i^2 v_i^2 m_j^2 v_j^2) \\
&\quad + E(w_{ji}^2 w_{it}^2 m_i^2 v_i^2 m_j^2 v_t^2) + E(w_{ji}^2 w_{it}^2 v_i^4 m_j^2 v_t^2)] \\
&\leq n^3 (C_w/k)^2 [E(w_{12} w_{13} m_1^4 v_2^2 v_3^2) + E(w_{12} w_{31} m_1^2 v_1^2 m_3^2 v_2^2) + E(w_{21} w_{13} m_1^2 v_1^2 m_2^2 v_3^2) \\
&\quad + E(w_{21} w_{31} v_1^4 m_2^2 v_3^2)] \stackrel{\text{B4}}{=} o(\sqrt{n^3/k^5}) = o(\rho^4). \quad \square
\end{aligned}$$

Lemma D4 $\sup_n E[\max_{i \leq n} B_i^2 / \tilde{D}^2(\theta_0)] < \infty$.

Proof: Let $\mathcal{Z} = \{z_1, \dots, z_n\}$. We have (omitting θ_0)

$$E(B_i^2 | \mathcal{Z}) = E(\zeta_i^2 | \mathcal{Z}) + E\left[\left(\sum_{j<i} \xi_{ij}\right)^2 | \mathcal{Z}\right]$$

and

$$\begin{aligned}
\sum_i E\left[\left(\sum_{j<i} \xi_{ij}\right)^2 | \mathcal{Z}\right] &\leq \sum_i \sum_{j<i} \sum_{t<i} E(\xi_{ij} \xi_{it}) | \mathcal{Z} \sum_i \sum_{j<i} E(\xi_{ij}^2 | \mathcal{Z}) \\
&= \sum_{ij} E(\xi_{ij}^2 | \mathcal{Z}) / 2 = \sum_{ij} (h_{ij}^2 + h_{ij} h_{ji}).
\end{aligned}$$

Hence

$$\begin{aligned}
E\left[\max_{i \leq n} \frac{B_i^2}{\tilde{D}^2}\right] &\leq \sum_i E\left[\frac{B_i^2}{\tilde{D}^2}\right] = \sum_i E\left[\frac{E(B_i^2 | \mathcal{Z})}{\tilde{D}^2}\right] \\
&\leq E\left[\frac{\sum_i E(\zeta_i^2 | \mathcal{Z}) + \sum_{ij} (h_{ij}^2 + h_{ij} h_{ji})}{nE(\zeta_1^2) + \sum_{ij} (h_{ij}^2 + h_{ij} h_{ji})}\right] \leq E\left[\frac{\sum_i E(\zeta_i^2 | \mathcal{Z})}{E\zeta_1^2}\right] + 1 = 2. \quad \square
\end{aligned}$$

E Nuisance Parameters

Lemma E1 *If f, f^* are such that for a neighborhood \aleph of $\tilde{\beta}$,*

$E(E[\sup_{\beta \in \aleph} |f_1(\theta, \beta)| | z_1])^2 < \infty$ and $E(E[\sup_{\beta \in \aleph} |f_1^(\theta, \beta)| | z_1])^2 < \infty$, then for any β^* converging to $\tilde{\beta}$, $\sum_{ij} w_{ij} f_i^*(\theta, \beta^*) f_j(\theta, \beta^*) = O_p(n)$.*

Proof: Let n be big enough to ensure that $\beta^* \in \aleph$. Then

$$\begin{aligned} \left| \sum_{ij} w_{ij} f_j(\theta, \beta^*) f_i^*(\theta, \beta^*) \right| &\leq \sup_{\beta \in \aleph} \left| \sum_{ij} w_{ij} f_j(\theta, \beta) f_i^*(\theta, \beta) \right| \\ &\leq \sum_{ij} w_{ij} \sup_{\beta \in \aleph} |f_j(\theta, \beta)| \sup_{\beta \in \aleph} |f_i^*(\theta, \beta)| = O_p(n), \end{aligned}$$

by lemma B4. \square

Lemma E2 $\hat{\beta}(\theta) \xrightarrow{p} \beta(\theta)$.

Proof: Since $E(h_1 m_1)$ is continuous and yields a unique solution $\beta(\theta)$ by assumption D, we are left to show uniform convergence (in β at θ) of $n^{-1} \sum_i \hat{h}_i m_i$ to $E(h_1 m_1)$. We first show pointwise convergence and then stochastic equicontinuity. By lemma B5(vi) at fixed θ, β ,

$$n^{-1} \sum_i \hat{h}_i m_i = n^{-1} \sum_{ij} w_{ij} m_{\theta j} m_i = n^{-1} \sum_i h_i E(m_i | z_i) + o_p(1) = E(h_1 m_1) + o_p(1).$$

Stochastic equicontinuity then follows from the fact that

$$\begin{aligned} \sup_{\tilde{\beta}} \sup_{\beta: \|\beta - \tilde{\beta}\| \leq \delta} \left| n^{-1} \sum_i [\hat{h}_i(\theta, \tilde{\beta}) m_i(\theta, \tilde{\beta}) - \hat{h}_i(\theta, \beta) m_i(\theta, \beta)] \right| \\ \leq \left(n^{-1} \sum_{ij} w_{ij} \sup_{\beta} \|m_{\beta \beta j}(\theta, \beta)\| \sup_{\beta} |m_i(\theta, \beta)| \right. \\ \left. + n^{-1} \sum_{ij} w_{ij} \sup_{\beta} \|m_{\beta j}(\theta, \beta)\| \sup_{\beta} |m_{\beta i}(\theta, \beta)| \right) \times \sup_{\tilde{\beta}} \sup_{\beta: \|\beta - \tilde{\beta}\| \leq \delta} \|\tilde{\beta} - \beta\| \stackrel{\text{B4,E}}{=} O_p(\delta). \end{aligned}$$

Choose δ proportional to ε in (21.42) of Davidson (1994). \square

Lemma E3 $\hat{\beta}(\theta) - \beta(\theta) = \begin{cases} O_p(n^{-1/2}), & \theta = \theta_0, \\ O_p(n^{-1/2}) + o_p(\lambda), & \theta \neq \theta_0. \end{cases}$

Proof: Let $S_n(\theta, \beta^*)$ be the vector with t -th element

$$\begin{aligned} (\hat{\beta}(\theta) - \beta(\theta))' \sum_i \left(\hat{h}_{\beta \beta i t}(\theta, \beta^*) m_i(\theta, \beta^*) + 2 \hat{h}_{\beta i t}(\theta, \beta^*) m'_{\beta i}(\theta, \beta^*) \right. \\ \left. + \hat{h}_{i t}(\theta, \beta^*) m_{\beta \beta i}(\theta, \beta^*) \right) (\hat{\beta}(\theta) - \beta(\theta)) / 2, \end{aligned}$$

where $\hat{h}_{i t}$ is the t -th element of \hat{h}_i and $\hat{h}_{\beta i t}, \hat{h}_{\beta \beta i t}$ are its first and second partial

derivatives with respect to β . Then by lemmas E1 and E2, $S_n(\theta, \beta^*) = O_p(n\|\hat{\beta}(\theta) - \beta(\theta)\|^2)$ for any β^* between $\hat{\beta}(\theta)$ and $\beta(\theta)$. Hence by the mean value theorem (omitting θ from hereon),

$$0 = \sum_i \hat{h}_i \hat{m}_i = \sum_i \hat{h}_i m_i + \sum_i m_i \hat{h}'_{\beta i} (\hat{\beta} - \beta) + \sum_i \hat{h}_i m'_{\beta i} (\hat{\beta} - \beta) + O_p(n\|\hat{\beta} - \beta\|^2). \quad (31)$$

Now, noting that $M_i(\theta_0) = 0$ a.s. and that $E(h_1 M_1) = E(h_1 m_1) = 0$ by definition,

$$\begin{aligned} \sum_i \hat{h}_i m_i &= \underbrace{\sum_i \hat{h}_i (m_i - M_i)}_{\stackrel{\text{B5(ii)}}{=} O_p(\sqrt{n})} + \underbrace{\lambda \sum_i (\hat{h}_i - h_i) \tilde{M}_i}_{\stackrel{\text{B5(vi)}}{=} o_p(n\lambda)} + \underbrace{\lambda \sum_i h_i \tilde{M}_i}_{=O_p(\sqrt{n\lambda})} \\ &= \begin{cases} O_p(\sqrt{n}), & \theta = \theta_0, \\ O_p(\sqrt{n}) + o_p(n\lambda), & \theta \neq \theta_0. \end{cases} \quad (32) \end{aligned}$$

Further,

$$\begin{aligned} \sum_i \hat{h}_{\beta i} m_i &= \underbrace{\sum_i \hat{h}_{\beta i} (m_i - M_i)}_{\stackrel{\text{B5(ii)}}{=} O_p(\sqrt{n})} + \underbrace{\lambda \sum_i (\hat{h}_{\beta i} - h_{\beta i}) \tilde{M}_i}_{\stackrel{\text{B5(vi)}}{=} o_p(n\lambda)} + \underbrace{\lambda \sum_i (h_{\beta i} \tilde{M}_i - E[h_{\beta 1} \tilde{M}_1])}_{=O_p(\sqrt{n\lambda})} + nE(h_{\beta 1} M_1) \\ &= nE(m_{\beta\beta 1} M_1) + \begin{cases} O_p(\sqrt{n}), & \theta = \theta_0, \\ O_p(\sqrt{n}) + o_p(n\lambda), & \theta \neq \theta_0. \end{cases} \quad (33) \end{aligned}$$

And

$$\sum_i \hat{h}_i m'_{\beta i} \stackrel{\text{B5(vi)}}{=} \sum_i h_i h'_i + o_p(n) = nE(h_1 h'_1) + o_p(n). \quad (34)$$

Plugging (32)–(34) into (31) we get

$$n(Q + o_p(1))(\hat{\beta} - \beta) = \begin{cases} O_p(\sqrt{n}), & \theta = \theta_0, \\ O_p(\sqrt{n}) + o_p(n\lambda), & \theta \neq \theta_0. \end{cases}$$

The stated result then follows from the fact that Q is invertible by assumption D. \square

Lemma E4 $\sum_i \hat{q}_i(\theta) \hat{m}_i(\theta) = \sum_i \hat{q}_i(\theta) m_i(\theta) + \begin{cases} o_p(\rho), & \theta = \theta_0, \\ O_p(\rho) + o_p(n\lambda^2), & \theta \neq \theta_0. \end{cases}$

Proof: Note that for any β^* between $\hat{\beta}(\theta)$ and $\beta(\theta)$,

$$\begin{aligned} & \left\| (\hat{\beta}(\theta) - \beta(\theta))' \sum_i \left[\hat{g}_{\beta\beta_i}(\theta, \beta^*) m_i(\theta, \beta^*) + 2\hat{g}_{\beta_i}(\theta, \beta^*) m'_{\beta_i}(\theta, \beta^*) \right. \right. \\ & \quad \left. \left. + \hat{g}_i(\theta, \beta^*) m_{\beta\beta_i}(\theta, \beta^*) \right] (\hat{\beta}(\theta) - \beta(\theta)) \right\| \stackrel{\text{E1}}{=} O_p(n \|\hat{\beta}(\theta) - \beta(\theta)\|^2). \end{aligned}$$

Hence by the mean value theorem (omitting arguments from hereon)

$$\begin{aligned} \sum_i \hat{q}_i \hat{m}_i &= \sum_i \hat{g}_i \hat{m}_i \\ &= \sum_i \hat{g}_i [m_i + m'_{\beta_i}(\hat{\beta} - \beta)] + \sum_i m_i \hat{g}'_{\beta_i}(\hat{\beta} - \beta) + O_p(n \|\hat{\beta} - \beta\|^2) \\ &= \sum_i \hat{q}_i m_i + \sum_i \hat{q}_i m'_{\beta_i}(\hat{\beta} - \beta) + \kappa' \sum_i \hat{h}_i [m_i + m'_{\beta_i}(\hat{\beta} - \beta)] \\ & \quad + \sum_i m_i \hat{g}'_{\beta_i}(\hat{\beta} - \beta) + \begin{cases} O_p(1), & \theta = \theta_0, \\ O_p(1) + o_p(n\lambda^2), & \theta \neq \theta_0. \end{cases} \end{aligned} \quad (35)$$

We deal with each of the RHS2–RHS4 terms separately. For RHS3 in (35) note that by (31),

$$\begin{aligned} & \left\| \sum_i \hat{h}_i (m_i + m'_{\beta_i}(\hat{\beta} - \beta)) \right\| \leq \left\| \sum_i m_i \hat{h}_{\beta_i} \right\| \cdot \|\hat{\beta} - \beta\| + O_p(n \|\hat{\beta} - \beta\|^2) \\ & \stackrel{\text{E3}, (33)}{=} \begin{cases} O_p(n^{1/2}) O_p(n^{-1/2}) + O_p(1), & \theta = \theta_0, \\ \{O_p(n^{1/2}) + o_p(n\lambda)\} \{O_p(n^{-1/2}) + o_p(\lambda)\} + O_p(1) + o_p(n\lambda^2), & \theta \neq \theta_0. \end{cases} \end{aligned}$$

which is $\begin{cases} O_p(1), & \theta = \theta_0, \\ o_p(\rho + n\lambda^2), & \theta \neq \theta_0. \end{cases}$. For RHS4 in (35) we have

$$\begin{aligned} \sum_i \hat{g}_{\beta_i} m_i &= \underbrace{\sum_i \hat{g}_{\beta_i} (m_i - M_i)}_{\stackrel{\text{B5(ii)}}{=} O_p(\sqrt{n})} + \underbrace{\lambda \sum_i (\hat{g}_{\beta_i} - g_{\beta_i}) \tilde{M}_i}_{\stackrel{\text{B5(vi)}}{=} o_p(n\lambda)} + \underbrace{\lambda \sum_i g_{\beta_i} \tilde{M}_i}_{=O_p(n\lambda)} \\ &= \begin{cases} O_p(\sqrt{n}), & \theta = \theta_0, \\ O_p(\sqrt{n} + n\lambda), & \theta \neq \theta_0. \end{cases} \end{aligned} \quad (36)$$

Now combine (36) with lemma E3 to obtain rates for RHS4 in (35) of $O_p(1)$ if $\theta = \theta_0$

and $O_p(\rho) + o_p(n\lambda^2)$ if $\theta \neq \theta_0$. Finally RHS2 in (35). We have

$$\begin{aligned} \sum_i \hat{q}_i m_{\beta i} &\stackrel{\text{B5(ii)}}{=} \sum_i \hat{q}_i h_i + O_p(\sqrt{n}) \stackrel{E(q_1 h_1) = 0}{=} \sum_i (\hat{q}_i - q_i) h_i + O_p(\sqrt{n}) \\ &= \lambda \underbrace{\sum_{ij} w_{ij} (\tilde{q}_j - \tilde{q}_i) h_i}_{\stackrel{\text{B5(vi)}}{=} o_p(n\lambda)} + \underbrace{\sum_{ij} w_{ij} (m_{\theta j} - \kappa' m_{\beta j} - q_j) h_i}_{\stackrel{\text{B5(iv)}}{=} o_p(n^{3/4} k^{-1/4})} + O_p(\sqrt{n}) \\ &= O_p(\sqrt{n}) + o_p(\sqrt{n\rho}) + o_p(n\lambda), \end{aligned}$$

such that RHS2 in (35) is $\begin{cases} o_p(\rho), & \theta = \theta_0, \\ o_p(\rho + n\lambda^2), & \theta \neq \theta_0. \end{cases}$ \square

Lemma E5 $\hat{\kappa}(\theta) - \kappa(\theta) = o_p(\rho/\sqrt{n})$.

Proof: Note that (omitting θ)

$$\hat{\kappa} - \kappa = \left(n^{-1} \sum_i \hat{h}_i \hat{h}'_i \right)^{-1} n^{-1} \sum_i \hat{h}_i (\hat{g}_i - \hat{h}'_i \kappa).$$

By Slutsky it suffices to show that (i) $n^{-1} \sum_i \hat{h}_i \hat{h}'_i - E(h_1 h'_1) = o_p(1)$ and (ii) $n^{-1} \sum_i \hat{h}_i (\hat{g}_i - \hat{h}'_i \kappa) = o_p(\rho/\sqrt{n})$. Establishing (i) is similar to but simpler than showing (ii) so we only show (ii) here. We show that

$$\sum_i [\hat{h}_i (\hat{g}_i - \hat{h}'_i \kappa) - \hat{h}_i \hat{q}_i] = o_p(\sqrt{n\rho}), \quad (37)$$

$$\sum_i (\hat{h}_i \hat{q}_i - h_i q_i) = o_p(\sqrt{n\rho}), \quad (38)$$

$$\sum_i h_i q_i = o_p(\sqrt{n\rho}). \quad (39)$$

Since $E(h_1 q_1) = 0$ by construction, the LHS in (39) is $O_p(\sqrt{n\lambda})$. For (38) note that

$$\sum_i (\hat{h}_i \hat{q}_i - h_i q_i) = \sum_i (\hat{h}_i - h_i) (\hat{q}_i - q_i) + \sum_i (\hat{h}_i - h_i) q_i + \sum_i h_i (\hat{q}_i - q_i). \quad (40)$$

First RHS3 in (40). Let $\epsilon_i = m_{\theta i} - m'_{\beta i} \kappa - q_i$. Then

$$\sum_i h_i (\hat{q}_i - q_i) = \lambda \sum_{ij} w_{ij} h_i (\tilde{q}_j - \tilde{q}_i) + \sum_{ij} w_{ij} h_i \epsilon_j \stackrel{\text{B1}}{=} o_p(n\lambda) + \sum_{ij} w_{ij} h_i \epsilon_j.$$

But

$$\begin{aligned} & E \left\| \sum_{ij} w_{ij} h_i \epsilon_j \right\|^2 \\ & \leq \sum_{ij} E(w_{ij}^2 \|h_i\|^2 \epsilon_j^2) + \sum_{ij} \sum_{t \neq i} E(w_{ij} w_{tj} \|h_i\| \|h_t\| \epsilon_j^2) \stackrel{\text{B4}}{=} o(n^{3/2} k^{-1/2}) = o(n\rho). \end{aligned}$$

The derivation for RHS2 in (40) is similar to that for RHS3 and is omitted. By the Schwarz inequality, sufficient conditions for RHS1 in (40) to be $o_p(\sqrt{n}\rho)$ are that (a) $\sum_i \|\hat{h}_i - h_i\|^2 = o_p(n)$ and (b) $\sum_i (\hat{q}_i - q_i)^2 = O_p(\rho^2)$. First (b). Since $\hat{q}_i - q_i = \sum_j w_{ij}(q_j - q_i + \epsilon_j)$, we have

$$\begin{aligned} 2^{-1} \sum_i E(\hat{q}_i - q_i)^2 & \leq \sum_i E \left(\sum_j w_{ij}(q_j - q_i) \right)^2 + \sum_i E \left(\sum_j w_{ij} \epsilon_j \right)^2 \\ & \stackrel{\text{Schwarz}}{\leq} n^2 \lambda^2 E[w_{12}(\tilde{q}_2 - \tilde{q}_1)^2] + n^2 E(w_{12}^2 \epsilon_2^2) \stackrel{\text{B1,B4}}{=} o(n\lambda^2) + O(n/k) = O(\rho^2). \end{aligned}$$

Requirement (a) follows with the same derivation resulting in a rate of $o(n) + O(n/k) = o(n)$. Finally apply the mean value theorem to the LHS in (37) and obtain for some $\hat{\beta}^*$ between $\hat{\beta}$ and β

$$\sum_i \left[\hat{h}_{\beta i}(\hat{g}_i - \hat{h}'_i \kappa) + \hat{h}_i(\hat{g}_{\beta i} - \hat{h}'_{\beta i} \kappa)' \right] (\hat{\beta} - \beta) \quad \text{all at } (\theta, \hat{\beta}^*). \quad (41)$$

By lemma E3, $\|\hat{\beta} - \beta\| = o_p(\rho/\sqrt{n})$. It is hence sufficient to show that the sum in (41) is $O_p(n)$, which follows from repeated application of lemma E1. \square

Lemma E6 $\sum_i \hat{q}_i^2(\theta) \hat{m}_i^2(\theta) = \sum_i \hat{q}_i^2(\theta) m_i^2(\theta) + o_p(\rho^2)$.

Proof: Since for generic a, b , $|a^2 - b^2| \leq (a-b)^2 + 2|b(a-b)|$, and $\sum_i \hat{q}_i^2 m_i^2 = O_p(\rho^2)$ by lemmas C8 and C10 (substituting \hat{q} for \hat{g}), we need to show that $\sum_i (\hat{q}_i \hat{m}_i - \hat{q}_i m_i)^2 = o_p(\rho^2)$. Now,

$$\hat{q}_i \hat{m}_i - \hat{q}_i m_i = (\hat{g}_i \hat{m}_i - \hat{g}_i m_i) - (\hat{\kappa} - \kappa)' (\hat{h}_i \hat{m}_i - \hat{h}_i m_i) - \kappa' (\hat{h}_i \hat{m}_i - \hat{h}_i m_i) - (\hat{\kappa} - \kappa)' \hat{h}_i m_i.$$

By lemma E5 it hence suffices to show that

$$\sum_i (\hat{g}_i \hat{m}_i - \hat{g}_i m_i)^2 = o_p(\rho^2), \quad \sum_i \|\hat{h}_i \hat{m}_i - \hat{h}_i m_i\|^2 = o_p(\rho^2), \quad \sum_i \|\hat{h}_i m_i\|^2 = O_p(n). \quad (42)$$

The third condition in (42) follows from lemma B5. Now the first condition. We have for some β^* between $\hat{\beta}$ and β ,

$$\begin{aligned} \sum_i (\hat{g}_i \hat{m}_i - \hat{g}_i m_i)^2 &= \sum_i \left(\sum_j w_{ij} (\hat{m}_{\theta_j} \hat{m}_i - m_{\theta_j} m_i) \right)^2 \\ &\leq \|\hat{\beta} - \beta\|^2 \sum_{ijt} w_{ij} w_{it} \left\| m_t(\theta, \beta^*) m_{\theta_{\beta_j}}(\theta, \beta^*) + m_{\theta_j}(\theta, \beta^*) m_{\beta t}(\theta, \beta^*) \right\|. \end{aligned} \quad (43)$$

Now since for generic a_i, b_i

$$\sum_{ijt} w_{ij} w_{it} |a_j b_t| \leq \sum_{ijt} w_{ij} w_{it} (a_j^2 + b_t^2) = \sum_{ij} w_{ij} (a_j^2 + b_j^2),$$

it follows from lemmas E1 and E3 that the RHS in (43) is $o_p(\rho^2/n) O_p(n) = o_p(\rho^2)$. So we have established the first condition in (42), where the derivation for the second condition follows similarly. \square

Lemma E7 $\sum_{ij} w_{ij} w_{ji} \hat{m}_i(\theta) \hat{m}_{\theta_i}(\theta) \hat{m}_j(\theta) \hat{m}_{\theta_j}(\theta) = \sum_{ij} w_{ij} w_{ji} m_i(\theta) m_{\theta_i}(\theta) m_j(\theta) m_{\theta_j}(\theta) + o_p(\rho^2)$.

Proof: We have

$$\begin{aligned} &\left\| \sum_{ij} w_{ij} w_{ji} (\hat{m}_i \hat{m}_{\theta_i} \hat{m}_j \hat{m}_{\theta_j} - m_i m_{\theta_i} m_j m_{\theta_j}) \right\| \\ &= \left\| (\hat{\beta} - \beta)' \sum_{ij} w_{ij} w_{ji} (m_{\beta_i}(\theta, \beta^*) m_{\theta_i}(\theta, \beta^*) m_j(\theta, \beta^*) m_{\theta_j}(\theta, \beta^*) + \dots + \right. \\ &\quad \left. m_i(\theta, \beta^*) m_{\theta_i}(\theta, \beta^*) m_j(\theta, \beta^*) m_{\theta_{\beta_j}}(\theta, \beta^*)) \right\| \\ &\leq (C_w/k) \|\hat{\beta} - \beta\| \sum_{ij} w_{ij} \left\| m_{\beta_i}(\theta, \beta^*) m_{\theta_i}(\theta, \beta^*) m_j(\theta, \beta^*) m_{\theta_j}(\theta, \beta^*) + \dots + \right. \\ &\quad \left. m_i(\theta, \beta^*) m_{\theta_i}(\theta, \beta^*) m_j(\theta, \beta^*) m_{\theta_{\beta_j}}(\theta, \beta^*) \right\|. \end{aligned}$$

Apply the triangle inequality and lemmas E1 and E3 to obtain a rate of $o_p(\rho/\sqrt{n}) O_p(n/k) = o_p(\rho^2)$. \square

Lemma E8 $\hat{D}^2(\theta) = \hat{D}^2(\theta) + o_p(\rho^2)$.

Proof: By lemmas E4, E6 and E7 (omitting θ),

$$\begin{aligned} \hat{D}^2(\theta) - \hat{D}^2(\theta) &= \underbrace{\left[\sum_i \hat{q}_i^2 \hat{m}_i^2 - \sum_i \hat{q}_i^2 m_i^2 \right]}_{\stackrel{\text{E6}}{=} o_p(\rho^2)} - n^{-1} \left[\left(\sum_i \hat{q}_i \hat{m}_i \right)^2 - \left(\sum_i \hat{q}_i m_i \right)^2 \right] \\ &\quad + \underbrace{\left[\sum_{ij} w_{ij} w_{ji} \hat{m}_i \hat{m}_{\theta_i} \hat{m}_j \hat{m}_{\theta_j} - \sum_{ij} w_{ij} w_{ji} m_i m_{\theta_i} m_j m_{\theta_j} \right]}_{\stackrel{\text{E7}}{=} o_p(\rho^2)} \end{aligned} \quad (44)$$

Finally, RHS2 in (25) is in absolute value bounded by

$$\underbrace{n^{-1} \left(\sum_i (\hat{q}_i \hat{m}_i - \hat{q}_i m_i) \right)^2}_{\stackrel{\text{E4}}{=} O_p(\rho^2/n) + o_p(n\lambda^4)} + 2n^{-1} \underbrace{\left| \sum_i \hat{q}_i m_i \right|}_{=O_p(\rho)} \underbrace{\left| \sum_i (\hat{q}_i \hat{m}_i - \hat{q}_i m_i) \right|}_{\stackrel{\text{E4}}{=} O_p(\rho) + o_p(n\lambda^2)} = o_p(\rho^2). \quad \square$$

F Justification of example I

We first justify that

$$\begin{bmatrix} n^{-1/2} \sum_i \tilde{g}_i u_i \\ n^{-1/2} \sum_i \tilde{g}_i v_i \\ (n/k)^{-1/2} \sum_{ij} w_{ij} v_j u_i \\ (n/k)^{-1/2} \sum_{ij} w_{ij} v_j v_i \end{bmatrix} \xrightarrow{d} \begin{bmatrix} \Psi_{\tilde{g}u} \\ \Psi_{\tilde{g}v} \\ \Psi_{vu} \\ \Psi_{vv} \end{bmatrix} \sim N(0, V_{\Psi}), \quad (45)$$

for some $0 < V_{\Psi} < \infty$. By the Cramér–Wold device, it suffices to show the normality of an arbitrary linear combination, which can be expressed as

$$S_n = \sum_i \left(\frac{\Xi_i}{\sqrt{n}} + \frac{\sqrt{k}}{\sqrt{n}} \sum_{j < i} \Upsilon_{ij} \right),$$

where $\Xi_i = a_1 \tilde{g}_i u_i + a_2 \tilde{g}_i v_i$, and $\Upsilon_{ij} = w_{ij}(a_3 v_j u_i + a_4 v_j v_i) + w_{ji}(a_3 v_i u_j + a_4 v_i v_j)$ with a_1, a_2, a_3, a_4 being arbitrary constants. Since $X_i = \frac{\Xi_i}{\sqrt{n}} + \frac{\sqrt{k}}{\sqrt{n}} \sum_{j < i} \Upsilon_{ij}$ is a martingale difference array, so is X_i/ς where $\varsigma^2 = \sum_i EX_i^2$. The necessary conditions of theorem 24.3 of Davidson (1994) can be verified similar to the way in which the conditions of lemma D1 were verified in the proof of theorem 1.

We now analyze $\hat{\theta}_I - \theta_0 = \hat{N}_I/\hat{D}_I$ for implicitly defined \hat{N}_I, \hat{D}_I . We have

$$n^{-1/2}\lambda^{-1}\hat{N}_I = n^{-1/2}\sum_i \tilde{g}_i u_i, \quad (46)$$

$$\begin{aligned} n^{-1/2}\lambda^{-1}\hat{D}_I &= n^{-1/2}\sum_i \tilde{g}_i Y_i = \sqrt{n}\lambda n^{-1}\sum_i \tilde{g}_i^2 + n^{-1/2}\sum_i \tilde{g}_i v_i \\ &= (\sqrt{n}\lambda)E\tilde{g}_i^2 + n^{-1/2}\sum_i \tilde{g}_i v_i + O_p(\lambda), \end{aligned} \quad (47)$$

where we used the fact that $n^{-1}\sum_i \tilde{g}_i^2 = E(\tilde{g}_i^2) + O_p(n^{-1/2})$. The result stated in (1) then follows directly from combining (45)–(47) and applying the continuous mapping theorem. We now turn to $\hat{\theta} - \theta_0 = \hat{N}/\hat{D}$, for which

$$\begin{aligned} \hat{N} &= \sum_i \hat{g}_i u_i = \sum_{ij} w_{ij} Y_j u_i = \lambda \sum_{ij} w_{ij} \tilde{g}_j u_i + \sum_{ij} w_{ij} v_j u_i = \lambda \sum_{ij} w_{ij} (\tilde{g}_j - \tilde{g}_i) u_i \\ &\quad + \lambda \sum_i \tilde{g}_i u_i + \sum_{ij} w_{ij} v_j u_i = \lambda \underbrace{\sum_i \tilde{g}_i u_i}_{O_p(\sqrt{n}\lambda)} + \underbrace{\sum_{ij} w_{ij} v_j u_i}_{O_p(\sqrt{n/k})} + o_p(\sqrt{n}\lambda). \end{aligned} \quad (48)$$

Similarly,

$$\begin{aligned} \hat{D} &= \sum_i \hat{g}_i Y_i = \lambda \sum_i \hat{g}_i \tilde{g}_i + \sum_i \hat{g}_i v_i \\ &= \lambda^2 \underbrace{\sum_i \tilde{g}_i^2}_{O_p(n\lambda^2)} + \lambda^2 \underbrace{\sum_{ij} w_{ij} (\tilde{g}_j - \tilde{g}_i) \tilde{g}_i}_{o_p(n\lambda^2)} + \lambda \underbrace{\sum_{ij} w_{ij} \tilde{g}_i v_j}_{o_p(n\lambda^2 + \sqrt{n}\lambda + \sqrt{n/k})} \\ &\quad + \lambda \underbrace{\sum_i \tilde{g}_i v_i}_{O_p(\sqrt{n}\lambda)} + \lambda \underbrace{\sum_{ij} w_{ij} (\tilde{g}_j - \tilde{g}_i) v_i}_{o_p(\sqrt{n}\lambda)} + \underbrace{\sum_{ij} w_{ij} v_i v_j}_{O_p(\sqrt{n/k})} \end{aligned} \quad (49)$$

For RHS3 in (49), note that by lemma B4,

$$E\left(\sum_{ij} w_{ij} \tilde{g}_i v_j\right)^2 = \sum_{ijt} E(w_{ij} w_{it} \tilde{g}_i \tilde{g}_t v_j^2) = o(n^{3/2} k^{-1/2}),$$

which implies

$$\lambda \sum_{ij} w_{ij} \tilde{g}_i v_j = o_p(\lambda n^{3/4} k^{-1/4}) = o_p(\sqrt{n} \lambda \sqrt{\rho}) = o_p(n \lambda^2 + \rho)$$

where the last equality used the generic inequality $2ab \leq a^2 + b^2$. Therefore, we have

$$\hat{D} = \underbrace{\lambda^2 \sum_i \tilde{g}_i^2}_{O_p(n \lambda^2)} + \underbrace{\lambda \sum_i \tilde{g}_i v_i}_{O_p(\sqrt{n} \lambda)} + \underbrace{\sum_{ij} w_{ij} v_i v_j}_{O_p(\sqrt{n/k})} + o_p(n \lambda^2 + \sqrt{n} \lambda + \sqrt{n/k}).$$

The dominant numerator terms converge at the same rate if $\lambda \sim 1/\sqrt{k}$ and the denominator terms when $\lambda \sim 1/\sqrt[4]{nk}$. Combining (48) and (49) with (45) and applying the continuous mapping theorem for each of the three cases in (2) gives the results stated therein. \square