Semiparametric Tests of Conditional Moment Restrictions under Weak or Partial Identification*

Sung Jae Jun† and Joris Pinkse‡
The Center for the study of Auctions, Procurements and Competition Policy
Department of Economics
The Pennsylvania State University

October 2008

forthcoming in the Journal of Econometrics

Abstract

We propose two new semiparametric specification tests which test whether a vector of conditional moment conditions is satisfied for any vector of parameter values θ_0. Unlike most existing tests, our tests are asymptotically valid under weak and/or partial identification and can accomodate discontinuities in the conditional moment functions. Our tests are moreover consistent provided that identification is not too weak. We do not require the availability of a consistent first step estimator. Like Robinson (1987) and many others in similar problems subsequently, we use k–nearest neighbor (knn) weights instead of kernel weights. The advantage of using knn weights is that local power is invariant to transformations of the instruments and that under strong point identification computation of the test statistic yields an efficient estimator of θ_0 as a byproduct.

*We thank the co–editor and two anonymous referees for their comments and suggestions. We thank the Human Capital Foundation for their support.

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

‡joris@psu.edu; Joris Pinkse is an extramural fellow at Tilburg University

1 Introduction

We propose two specification tests for models defined by conditional moment restrictions (CMR). The null hypothesis of interest is that the model is correctly specified in the sense that there exists a (set of) parameter value(s) θ_0 that makes the conditional moment restrictions satisfied with probability one. Note that we do not impose any identification assumption, and the null hypothesis allows for the possibilities of *weak* and *partial identification*. We moreover do not require any smoothness or continuity assumptions under the null hypothesis.

Although there are quite a few specification tests that can be used in a similar context, virtually all such tests impose strong point identification under the null hypothesis and differentiability of the moment conditions Moreover, they typically exploit the fact that the unique θ_0 is $\sqrt{n}$–consistently estimable under the null (and sometimes even under the alternative). See e.g. Bierens (1990), Zheng (1996), Fan and Li (1996, 2000), Koul and Ni (2004), Delgado, Domínguez and Lavergne (2006); see Robinson (1991) and others for nonparametric independence tests. However, there are many economic examples in which a model does not provide point identification even when the specification is correct; see e.g. Chernozhukov, Hong, and Tamer (2007). Also, as Staiger and Stock (1997) and Stock and Wright (2000) have shown, even point–identified models can be difficult to deal with when identification is too weak. For example, tests of overidentifying restrictions based on two–step methods may suffer from dramatic size distortions under weak identification.

One test statistic that does allow for identification failure was proposed by Guerre and Lavergne (2005). Their test is similar to Zheng's except that the conditional error variance function (CEVF) is estimated fully nonparametrically, i.e. without estimating the (potentially nonidentified) parameter vector. Guerre and Lavergne moreover introduce a data–dependent (optimal) choice of smoothing parameter and justify a bootstrap approximation. Estimating the CEVF nonparametrically resolves the identification problem naturally and the Guerre–Lavergne test, unlike ours, is *similar*. But nonparametric CEVF estimation rules out a large class of interesting models commonly used in economics that are prone to causing identification problems, including all models with endogenous regressors. Further, Guerre and Lavergne require smoothness under the null and assume that the conditioning variables have bounded support. In short, estimating the CEVF nonparametrically (Guerre–Lavergne) is preferable in smooth models containing only exogenous regressors with bounded support, but does not apply to models which feature more general moment conditions models considered here and in e.g. Delgado, Domínguez and Lavergne (2006).

Our proposed test statistics are the minimum values of objective functions $\hat{T}_1, \hat{T}_2$, and they do not rely on the availability of $\sqrt{n}$–consistent estimators under the null. We show that our tests have correct size regardless of the identification situation and that they are consistent as long as the degree of misspecification is not too *weak*. We also analyze the power of our tests under local alternatives.

The cost of the robustness properties of our tests is twofold. First, since our statistics are global minima, they can be conservative in the sense that the rejection rates under the null can be smaller than the nominal levels. However, when the model is correctly specified with the parameter θ_0 being uniquely and strongly identified, the asymptotic size corresponds to the nominal size; our simulation experiments show that the rejection rates are then close to the nominal levels in moderate size samples, also. This is not surprising, because the global minimizer of our objective function is a $\sqrt{n}$–consistent estimator in this case. In fact, the global minimizer of $\hat{T}_2$ is then an efficient

estimator of θ_0 under homoskedasticity.[1] A second, more serious, problem is that the power of our test is necessarily no greater than that of the two–step version because our test statistic minimizes $\hat{T}_2$ (say) while the two–step version evaluates $\hat{T}_2$ at the first step estimate. The difference between the minimum of $\hat{T}_2$ and the value of $\hat{T}_2$ at the first–step estimate can be large under weak identification when the two–step test is invalid. But it can also be large if the conditional variance of the specified moment function varies more under the alternative than under the null, because the minimizer of $\hat{T}_2$ and the first step estimator in the two–step procedure may have different probability limits under the alternative.

The paper most closely related to ours is Zheng (1996). Zheng proposed a two–step approach using kernel weights with scalar moment conditions. Extending his results to the vector–valued continuous–updating case would be similar to the tests we propose in this paper. Like Robinson (1987), we use k–nearest neighbor (knn) weights instead of kernel weights. Using the knn method has two advantages. First, unlike the Zheng test, the local power of our test is invariant to transformations of exogenous conditioning variables. Zheng's test can probably be changed to achieve invariance at the cost of an unnatural and undesirable trimming procedure or (like Guerre and Lavergne (2005)) by assuming that the regressors have compact support. The second advantage of our procedure over a version of Zheng's is that under strong point identification the minimizer of our objective function is, as mentioned earlier, efficient.

The moments used here are based on parametric functions and our test does hence not cover semiparametric models like the partial linear model of Robinson (1988). For such models, the Fan and Li (1996) test is a natural choice.

The paper is organized as follows. We first define our statistics and discuss their properties under the null hypothesis. We then show their consistency properties under the traditional fixed alternative. Section 4 studies the behavior of our test under classical local alternatives, and we compare our test with Zheng (1996). Section 5 contains a modest simulation study and section 6 concludes.

2 Hypotheses and Test Statistics

Let $\{\omega_i, z_i\}$ be an i.i.d. random sequence and $\tilde{m}_i(\theta) = \tilde{m}(\omega_i, \theta)$ be a d–dimensional vector–valued function, where z_i is a d_z–dimensional vector of exogenous variables that may be contained in ω_i. We are interested in testing the null hypothesis that for some value θ_0, $E[\tilde{m}_i(\theta_0)|z_i] = 0$ a.s., i.e.

$$H_0 : \exists \theta_0 \in \Theta \quad s.t. \quad P\big(\, \|E(\tilde{m}_i(\theta_0)|z_i)\| = 0 \, \big) = 1 \tag{1}$$

$$H_1 : P\big(\inf_{\theta \in \Theta} \|E(\tilde{m}_i(\theta)|z_i)\| \neq 0 \, \big) > 0. \tag{2}$$

Before we describe our statistics in detail, we provide a few examples of situations in which our test is useful. Example I is a standard situation in which under the null hypothesis the regression function of interest has a prespecified parametric functional form. Example II deals with identification failure due to weak instruments, which is not covered by standard nonparametric testing procedures.

[1] We do not show efficiency of the estimator in this paper. Note that an estimator derived from a kernel–based specification test is known to be inefficient even under homoskedasticity; see Linton (1997, 1998). The estimator minimizing $\hat{T}_2$ is comparable with Koul and Ni's (2004); their minimum distance estimator has the same variance as the standard nonlinear least squares estimator, but does not allow for endogeneity.

Example I (Testing a Functional Form in Regression Models) *Suppose that*

$$\begin{aligned} H_0: &\quad P\big(\ E(y_i|z_i) = f(z_i,\theta_0)\ \big) = 1 \quad \text{for some} \quad \theta_0 \in \Theta \\ H_1: &\quad P\big(\ E(y_i|z_i) = f(z_i,\theta)\ \big) < 1 \quad \text{for all} \quad \theta \in \Theta, \end{aligned}$$

where $f(z_i,\theta)$ is a known function. See e.g. Zheng (1996), Fan and Li (1996, 2000), and Guerre and Lavergne (2005).

Example II (Unidentified IV Models) *Consider a simple IV model given by*

$$E(y_i - Y_i\theta_0|z_i) = 0,$$

where y_i is a scalar outcome variable, Y_i is a scalar endogenous variable, and z_i is a vector of instruments. Note that the (conditional) variance of $u_i = y_i - Y_i\theta_0$ cannot be estimated without estimating θ_0. Therefore, the method of Guerre and Lavergne (2005) cannot be applied. Suppose that the instruments are not relevant such that $E(Y_i|z_i) = 0$ a.s.. If $P\big(E(y_i|z_i) \neq 0\big) > 0$, then the specification is incorrect, because there is no parameter value that achieves the moment condition. If $E(y_i|z_i) = 0$ a.s., then the specification is correct but the parameter of interest is not identified. See e.g. Staiger and Stock (1997). In this case, we cannot reject the model on the basis of the available data.

We now proceed by describing our testing procedures. Under the null hypothesis there exists a θ_0 such that

$$E\Big(\big\|E\big(\tilde{m}_i(\theta_0)|z_i\big)\big\|^2\Big) = 0. \tag{3}$$

One possibility is to use a $\sqrt{n}$–consistent plugin estimator $\hat{\theta}$ of θ_0 and estimate the left hand side in (3) evaluated at $\hat{\theta}$. We instead consider continuous–updating type statistics, i.e. our test statistic evaluates an estimator of the left hand side in (3) at its minimizing value in order to achieve validity in situations like example II.

As mentioned in the introduction, we estimate the expectation in (3) using knn estimation, which is similar to kernel (regression) estimation in that it estimates a conditional mean by taking a weighted average over nearby observations (see e.g. Stone (1977), Robinson (1987)). The knn weights w_{ij} we use are determined as follows; see Robinson (1987) for a similar definition.

Definition A *Let k be such that $1 \prec k \prec n$, where $\prec$ ($\succ$) means that the left hand side converges faster (slower) than the right hand side. Let further $c_i(j)$ be any chosen constant, such that for all i, (i) $c_i(j) = 0$ for $j > k$, (ii) for fixed positive C_w^-, C_w independent of n, $C_w^- \leq c_i(j) \leq C_w$ for $1 \leq j \leq k$, and (iii) $\sum_{j=1}^k c_i(j) = 1$. Define $\rho_{ij} = \|z_i - z_j\|$, let $\{\upsilon_{ij}\}$ be independent random numbers drawn from a standard uniform distribution, and*

$$\zeta_{ij} = \sum_{t \neq i} I(\rho_{it} < \rho_{ij}), \quad \Psi_{ij} = \{t \neq i : \rho_{it} = \rho_{ij}\}, \quad \psi_{ij} = \sum_{t \in \Psi_{ij}} I(\upsilon_{it} < \upsilon_{ij}).$$

Then $w_{ij} = 0$ if $j = i$ and $w_{ij} = c_i(\zeta_{ij} + \psi_{ij} + 1)$ if $j \neq i$.

We make the following assumptions. Let $V(\theta) = E\big(V(\tilde{m}_i(\theta)|z_i)\big)$.

Assumption A *$V(\theta_0) = \mathrm{Var}\big(\tilde{m}_i(\theta_0)|z_i\big)$ with $0 < ||V(\theta_0)|| < \infty$.*

Assumption B *There is an M^* such that $P\big(E(\|\tilde{m}_i(\theta_0)\|^4|z_i) > M^*\big) = 0$.*

Assumption A says that $\tilde{m}_i(\theta_0)$ is homoskedastic under the null. Although it is restrictive, it can be relaxed at the expense of longer proofs; we do this for the scalar moment case in theorem 2. With homoskedasticity, $V(\theta_0) = V(\tilde{m}_i(\theta_0))$ under the null hypothesis. Note however that $V(\theta)$ is less than $V(\tilde{m}_i(\theta))$ when θ is different from θ_0. Assumption B is a restriction on the distribution of $\tilde{m}_i(\theta_0)$. Unlike Guerre and Lavergne (2005), we do not require z_i to have compact support.

We explicitly allow for the possibility of weak instruments by permitting $\tilde{\mu}_i(\theta) = E(\tilde{m}_i(\theta)|z_i)$ to vary with the sample size n, i.e. we impose the existence of a function $\tilde{\mu}_i^*$ which does not depend on n and a nonincreasing deterministic sequence of numbers λ, which can depend on n, such that

$$\tilde{\mu}_i(\theta) = \lambda\tilde{\mu}_i^*(\theta) \quad a.s.. \tag{4}$$

This allows for the possibility that instruments are so bad and the degree of misspecification is so minor that it is not meaningful to distinguish between poor identification and correct specification in finite samples. Note that this setup resembles but is different from classical local alternatives; see e.g. Eubank and Spiegelman (1990), Härdle and Mammen (1993) and Zheng (1996), and also section 4. They are similar in the sense that both of them are *asymptotic* tools designed to improve our understanding of the *finite sample* properties of tests. While classical local alternatives are intended to study finite sample power with identification being imposed, we introduce λ because we are concerned about the behavior of the tests under the null with weak or partial identification.

We now define our test statistics. Since $V(\theta)$ is unknown, we estimate it by

$$\hat{V}(\theta) = n^{-1}\sum_{ij_1j_2} w_{ij_1}w_{ij_2}\big(\tilde{m}_i(\theta) - \tilde{m}_{j_1}(\theta)\big)\big(\tilde{m}_i(\theta) - \tilde{m}_{j_2}(\theta)\big)'.$$

Letting $\hat{m}_i(\theta) = \hat{V}(\theta)^{-1/2}\tilde{m}_i(\theta)$, define

$$\hat{T}_1(\theta) = \frac{\sum_{ij} a_{ij}\hat{m}_i(\theta)'\hat{m}_j(\theta)}{\sqrt{2d\sum_{i\neq j} a_{ij}^2}} - \frac{d\sum_i a_{ii}}{\sqrt{2d\sum_{i\neq j} a_{ij}^2}} = \frac{\mathrm{tr}(\hat{\mathbb{M}}(\theta)'A\hat{\mathbb{M}}(\theta))}{\sqrt{2d\sum_{i\neq j} a_{ij}^2}} - \frac{d\,\mathrm{tr}(A)}{\sqrt{2d\sum_{i\neq j} a_{ij}^2}} \tag{5}$$

$$\hat{T}_2(\theta) = \frac{\sum_{ij} w_{ij}\hat{m}_i(\theta)'\hat{m}_j(\theta)}{\sqrt{d\sum_{ij} w_{ij}(w_{ij}+w_{ji})}} = \frac{\mathrm{tr}(\hat{\mathbb{M}}(\theta)'W\hat{\mathbb{M}}(\theta))}{\sqrt{d\sum_{ij} w_{ij}(w_{ij}+w_{ji})}}, \tag{6}$$

where k is chosen to satisfy $n^{3/4} \prec k \prec n$ for $\hat{T}_1$ and $1 \prec k \prec n$ for $\hat{T}_2$; $\hat{\mathbb{M}}(\theta)$ is an $n\times d$ matrix whose i^{th} row is $\hat{m}_i(\theta)'$, and $a_{ij} = \sum_{t=1}^n w_{ti}w_{tj}$ is the i,j element of $A = W'W > 0$ with W containing w_{ij}. The denominators can be shown to be bounded away from zero such that $\hat{T}_1(\theta)$ and $\hat{T}_2(\theta)$ are always well-defined. The test statistics are then

$$\hat{T}_s = \inf_{\theta\in\Theta} \hat{T}_s(\theta), \qquad \text{for } s = 1, 2. \tag{7}$$

The numerator of $\hat{T}_1$ is n times $n^{-1}\sum_{i=1}^n \|\hat{\mu}_i(\theta)\|^2$, with $\hat{\mu}_i$ a knn estimator of μ_i. $\hat{T}_2$ is a leave–one–out version of $\hat{T}_1$ and is similar to the statistics considered in Zheng (1996) and Fan and Li (1996), but using knn instead of kernel weights and continuous updating instead of plugin.[2]

[2]Assuming strong point identification of θ_0, the minimizer of (7) can be shown to be an efficient estimator for θ_0. Under partial identification, the limit function of $\hat{T}_s(\theta)$ (after rescaling) is $E\big(\|\mu_i(\theta)\|^2\big)$. An interesting estimator of the identified set is the collection of points θ for which $\hat{T}_s(\theta)$ is below some sample size–dependent number, much in the spirit of Chernozhukov, Hong, and Tamer (2007).

Theorem 1 *Suppose that assumptions A – B hold and let q_α be the $1-\alpha$ quantile of the standard normal. Then, under H_0, the test statistics $\hat{T}_1$ and $\hat{T}_2$, with k chosen as indicated below equation (6), satisfy*

$$\lim_{n\to\infty} P(\hat{T}_1 > q_\alpha) \leq \alpha \quad \text{and} \quad \lim_{n\to\infty} P(\hat{T}_2 > q_\alpha) \leq \alpha,$$

regardless of λ and uniqueness of θ_0.

Theorem 1 shows that the proposed statistics always have correct size and that they do not rely on the availability of a consistent estimator of θ_0. Note also that theorem 1 does not require any assumption on $\tilde{m}_i(\theta)$ except that they have uniformly bounded conditional fourth moments at θ_0. Therefore, our statistics can also be used for (instrumental) quantile models. Note that assumption A is always satisfied in (instrumental) quantile models.

We now extend the results of theorem 1 to allow for heteroskedasticity under the null. To conserve space, we only do this for scalar–valued $\tilde{m}_i(\theta)$, i.e. $d = 1$, and only for $\hat{T}_2$. Define $\sigma_i^2(\theta) = E\big(\tilde{m}_i^2(\theta)|z_i\big)$ such that $\sigma_i^2(\theta_0)$ is the conditional variance of $\tilde{m}_i(\theta_0)$ under the null.[3] Since $\sigma_i^2(\theta)$ is unknown, it should be nonparametrically estimated, $\hat{\sigma}_i^2(\theta) = \sum_j w_{ij}\tilde{m}_j^2(\theta)$. Letting $\hat{m}_i^H(\theta) = \tilde{m}_i(\theta)/\hat{\sigma}_i(\theta)$, define

$$\hat{T}_2^H = \inf_{\theta\in\Theta} \hat{T}_2^H(\theta) = \frac{\sum_{ij} w_{ij}\hat{m}_i^H(\theta)\hat{m}_j^H(\theta)}{\sum_{ij} w_{ij}(w_{ij}+w_{ji})}.$$

We then have the following theorem, which is an extension of theorem 1 to heteroskedasticity.

Theorem 2 *Suppose that assumption B holds and that $E\big(|\tilde{m}_i^2(\theta_0) - \sigma_i^2(\theta_0)|^{p^*}\big) < \infty$ for some $p^* > 6$. Suppose that there is a constant $C_s > 0$ such that $C_s \leq \sigma_i^2(\theta_0) < \infty$ a.s.. Let $k \succ n^{1/3+2/p^*}$, and let q_α be the $1-\alpha$ quantile of the standard normal. Then, under H_0,*

$$\lim_{n\to\infty} P(\hat{T}_2^H > q_\alpha) \leq \alpha,$$

regardless of λ and uniqueness of θ_0.

Theorems 1 and 2 show that the rejection probability under the null is asymptotically no greater than the nominal size. Since the tests are not similar absent strong point identification, there could be local alternatives under weak or partial identification for which our tests have power smaller than α. Although our tests are hence biased, we note that it is not generally possible to construct an asymptotically pivotal specification test that is similar under weak or partial identification. Indeed, tests of conditional moment restrictions require estimates of both the conditional moments and the conditional error variance. Unless the conditional error variance does not depend on any unidentified or weakly identified parameters, the conditional error variance cannot be consistently estimated. One case in which the conditional error variance function can be estimated is a regression model with only exogenous regressors, because in that case the conditional error variance is the same as the conditional variance of dependent variable given regressors; see e.g. Guerre and Lavergne (2005). Under partial identification, Guggenberger, Hahn, and Kim (2007) showed that specification tests of nonlinear moment inequalities have the form of nonlinear one–sided hypothesis tests of which no implementable asymptotically exact size ones are known (see also Wolak (1991)).

[3] Using $\sigma_i^*(\theta)^2 = V(\tilde{m}_i(\theta)|z_i)$ will increase power, although $\sigma_i^{*\,2}(\theta_0) = \sigma_i^2(\theta_0)$ under the null. We consider $\sigma_i^2(\theta)$ here for simplicity.

3 Strong Misspecification and Consistency of the Tests

In this section, we show that the probability of rejecting the null under the alternative converges to 1, when $\lambda \neq 0$ is fixed. As explained in section 2, this analysis establishes consistency against a fixed alternative. Local alternatives are considered in section 4.

We make the following assumptions. Let $m_i(\theta) = V(\theta)^{-1/2}\tilde{m}_i(\theta)$ and $\mu_i(\theta) = V^{-1/2}(\theta)\tilde{\mu}_i(\theta)$.

Assumption C *The parameter space Θ is compact.*

Assumption D $\inf_{\theta\in\Theta} \|V(\theta)\| > 0$ *and* $\sup_{\theta\in\Theta} \|V(\theta)\| < \infty$.

Assumption E *For some $1 < p < \infty$, $E(\sup_{\theta\in\Theta} \|\mu_i(\theta)\|^p) < \infty$.*

Assumption C is standard and so is assumption D. In fact, assumption D is related to, but weaker than, uniform boundedness (away from zero) of the conditional variance, as in Robinson (1987). Note that assumption E does not impose conditions on the error distribution. For instance, if $m_i = y_i - Y_i'\theta$, then assumption E imposes conditions on the moments of $E(y_i|z_i)$ and $E(Y_i|z_i)$, not on those of y_i, Y_i.

Definition B *Let $\mathcal{Y}$ be the support of y_i. Let $\mathcal{F}$ be a collection of functions defined by*

$$\mathcal{F} = \{f(y,\theta) : \mathcal{Y}\times\Theta \to \mathbb{R}^d \text{ s.t. } \forall\eta > 0\, \exists\delta > 0 : \|\theta-\tilde{\theta}\| < \delta \Rightarrow P(\|f(y_i,\theta)-f(y_i,\tilde{\theta})\| > \eta) < \eta\}. \quad (8)$$

Assumption F $m, \mu \in \mathcal{F}$

$\mathcal{F}$ is a collection of functions that are uniformly equicontinuous in θ with probability arbitrarily close to 1. If f is Lipschitz in θ with probability one, then it belongs to $\mathcal{F}$. Therefore, if m_i and μ_i are differentiable in θ with bounded derivatives, then they belong to $\mathcal{F}$. However, differentiability of m_i is not needed to satisfy assumption F. For example, $I(y_i \leq \theta)$, with I the indicator function, also belongs to $\mathcal{F}$, as long as y_i has a well–defined density. Note also that if $f(y_i,\theta) \in \mathcal{F}$, then $g(y_i)f(y_i,\theta)$ also belongs to $\mathcal{F}$ as long as $g(y_i)$ is bounded in probability.[4] Therefore, assumption F in fact implies that $\tilde{m}, \tilde{m}^2, \tilde{m}^4$ are all in $\mathcal{F}$.

Theorem 3 *Suppose that assumptions C – F hold. If $\lambda \neq 0$ is fixed, then under H_1, for any $C > 0$, we have*

$$\lim_{n\to\infty} P(\hat{T}_1 > C) = 1 \quad \text{and} \quad \lim_{n\to\infty} P(\hat{T}_2 > C) = 1.$$

Below we extend theorem 3 to allow for heteroskedasticity but, like before, only for $\hat{T}_2$ and $d = 1$, again to conserve space. Let $\sigma_i^2(\theta)$, $\hat{\sigma}_i^2(\theta)$, and $\hat{T}_2^H$ be defined as in section 2.

Theorem 4 *Suppose that assumptions C, E, and F hold and that $\sigma_i^2 \in \mathcal{F}$. Suppose that $\sup_\theta E\left(|\tilde{m}_i^2(\theta) - \sigma_i^2(\theta)|^{p^{**}}\right) < \infty$ for some $p^{**} > 6$ and that $C_s \leq \inf_\theta \sigma_i^2(\theta) \leq \sup_\theta \sigma_i^2(\theta) \leq C^s$ a.s. for some constants $C_s, C^s > 0$. Let $k \succ n^{1/2+1/p^{**}}$. If $\lambda \neq 0$ is fixed, then under H_1, for any $C > 0$, we have*

$$\lim_{n\to\infty} P(\hat{T}_2^H > C) = 1.$$

[4] Note that $P\big(|g(y_i)||f(y_i,\theta) - f(y_i,\tilde{\theta})| > \eta\big) \leq P\big(\eta_0|f(y_i,\theta) - f(y_i,\tilde{\theta})| > \eta\big) + P\big(|g(y_i)| > \eta_0\big)$.

4 Local Alternatives

In this section, we impose strong point identification (θ_0 is unique and $\lambda \neq 0$ is fixed) and analyze the power of our statistics under a sequence of local alternatives to the null.

The sequence of local alternatives takes the form

$$H_{1L} : P\big(\mu_i(\theta) = \mu_i^o(\theta) + \delta_n q_i\big) = 1, \tag{9}$$

where $q_i = q(z_i)$, $\mu_i^o(\theta) = \mu^o(z_i, \theta)$, $\mu_i^o(\theta_0) = 0$ a.s., and $\{\delta_n\}$ is a sequence that goes to zero at a rate of $(nk)^{-1/4}$, Define $\theta_n = \arg\min_{\theta\in\Theta} E\big(||\mu_i(\theta)||^2\big)$. We now make a few assumptions on μ_i^o, q_i. Let Γ_i denote the Jacobian of μ_i^o at θ_0.

Assumption G *(i) θ_0 is in the interior of Θ, (ii) $P(\mu_i^o(\theta) = 0) = 1 \Leftrightarrow \theta = \theta_0$, and (iii) $E(\Gamma_i'\Gamma_i)$ is invertible.*

Let $\iota^2 = E(\|q_i\|^2) - E(q_i'\Gamma_i)E(\Gamma_i'\Gamma_i)^{-1}E(\Gamma_i'q_i)$.

Assumption H *q_i is such that (i) $0 < E(q_i q_i') < \infty$, and (ii) $\iota^2 > 0$.*

Assumption G guarantees that $\theta_n = \theta_0 - \delta_n E(\Gamma_i'\Gamma_i)^{-1}E(\Gamma_i'q_i) + o(\delta_n)$ under (9). Assumption H is needed to ensure that $0 < E\big(||\mu_i(\theta_n)||\big) = O(\delta_n)$ under (9). It excludes the case that the local deviation q_i is in the linear span of Γ_i a.s.. If it is violated, then $\mu_i(\theta_n) = 0$ a.s., also. The same local alternatives were considered in e.g. Zheng (1996), but only in the context of regression estimation with exogenous regressors.[5] Let $m_{(1)i} = m_{\theta i} = \partial m_i/\partial\theta'$, $m_{(2)i} = m_{\theta\theta i} = \partial \operatorname{vec}(m_{\theta i})/\partial\theta'$, $m_{(3)i} = m_{\theta\theta\theta i} = \partial \operatorname{vec}(m_{\theta\theta i})/\partial\theta'$, and likewise for $\mu_{(1)i} = \mu_{\theta i}$, $\mu_{(2)i} = \mu_{\theta\theta i}$ and $\mu_{(3)i} = \mu_{\theta\theta\theta i}$.

Assumption I *$m_i(\theta)$ is three times differentiable such that*
(i) $\forall\theta \in \Theta : E\big(m_{(j)i}(\theta)|z_i\big) = \mu_{(j)i}(\theta)$ a.s. and $E\big(\sup_\theta ||\mu_{(j)i}(\theta)||^2\big) < \infty$ for $j = 1, 2, 3$,
(ii) $E\big(||\mu_{(1)i}(\theta_0)||^2\big) > 0$ and $E(||\mu_{(1)i}(\theta_0)||^4) < \infty$,
(iii) there is an M^{**} such that $P\big(E(||m_i(\theta_0) - \mu_i(\theta_0)||^4|z_i) > M^{**}\big) = 0$.

Assumption I imposes sufficient smoothness on m_i, which simplifies our discussion of classical local alternatives. Let $\hat\theta_s = \arg\min_{\theta\in\Theta} \hat T_s(\theta)$, such that $\hat T_s = \hat T_s(\hat\theta_s)$.

Theorem 5 *Suppose that assumptions A, B, D and G – I hold.*[6] *Then, under (9),*

$$\hat T_s - \frac{\sqrt{nk}\delta_n^2\iota^2}{\sqrt{k/n}D_s} \xrightarrow{d} N\big(0, 1\big), \quad s = 1, 2, \tag{10}$$

where D_s is such that $D_s^2 = O_p(n/k)$ and $1/D_s^2 = O_p(k/n)$.

Theorem 5 shows that our statistics have power to detect local alternatives approaching to the null at a rate of $\delta_n = (nk)^{-1/4}$ when strong point identification of θ_0 is assumed. Similar local power analyses can be found in Eubank and Spiegelman (1990), Härdle and Mammen (1991), Guerre and

[5] Zheng (1996) implicitly assumes that $E(\Gamma_i'q_i) = 0$. Otherwise, theorem 3 in Zheng (1996) would need a correction in the mean term of the limiting normal distribution.

[6] Note that earlier θ_0 was defined such that $\mu_i(\theta_0) = 0$ a.s.; here $\mu_i^o(\theta_0) = 0$ a.s.; assumptions A and B now apply to the current definition of θ_0.

Lavergne (2005), Zheng (1996), and many others. We focus in our comparison on Zheng (1996); Guerre and Lavergne (2005), theorem 3, achieve the same rate as Zheng (1996) and better than Horowitz and Spokoiny (2001).

Our result is equivalent to Zheng (1996) in terms of the rate of the local alternatives. Zheng shows that his test has some power detecting local alternatives approaching to the null at the rate of $n^{-1/2}h^{-d_z/4}$, where h is a bandwidth. For fixed k the local power of our test does not depend on d_z. However, a knn estimator can be thought of as a kernel estimator with bandwidth h_k equal to the distance to the k–th nearest neighbor, which suggests $h_k^{d_z} \sim k/n$.[7] Thus, $(nk)^{-1/4} \sim n^{-1/2}h_k^{-d_z/4}$, which is essentially Zheng's result. So we can always choose k to achieve the same rate as Zheng and vice versa. In particular, k can be chosen to increase at a rate arbitrarily close to n, in which case our test can detect alternatives arbitrarily close to $n^{-1/2}$, in contrast to e.g. Wooldridge (1992); see Zheng (1996) for a more in–depth discussion.

Zheng assumes that θ_0 can be $\sqrt{n}$–consistently estimable under the local alternatives, which is only true for alternatives for which $E(\Gamma_i' q_i) = 0$.[8] Under the local alternatives, θ_n can be $\sqrt{n}$–consistently estimated, but the difference between θ_n and θ_0 disappears at the rate at which the local alternatives approach the null. Since the local alternatives approach the null at a slower rate than $1/\sqrt{n}$, the difference between θ_n and θ_0 should be taken into account. Note that assuming $E(\Gamma_i' q_i) = 0$ makes the expectation of the numerator in (10) equal to $E(\|q_i\|^2)$ and that of Zheng equal to $E(\|\sqrt{\phi_i} q_i\|^2)$, where $\phi_i = \phi(z_i)$ denotes the density of z_i.

The fact that Zheng's local power depends on the density of z_i implies that taking transformations of z_i without changing the local alternatives will affect local power. In contrast to Zheng, the numerator in (10) does not contain the density and therefore it is invariant to transformations of z_i. The following theorem shows that the denominator in (10) is also invariant when z_i is scalar–valued; the case for vector–valued z_i is similar but is omitted to conserve space.

Theorem 6 *Suppose that assumptions A, B, D, and G – I hold. Suppose further that z_i is a scalar continuous random variable whose density has a uniformly bounded derivative and that $\{w_{ij}\}$ are uniform nearest neighbor weights (i.e. $w_{ij} \in \{0, 1/k\}$). Then under (9),*

$$\hat{T}_2 \xrightarrow{d} N\left(\frac{\iota^2}{\sqrt{2}}, 1\right).$$

To see the implications of the invariance property, consider the ratio of the mean of the local power of the Zheng test to that of our test for $d_z = 1$ and scalar–valued q_i with fixed k, h, and δ_n, assuming $E(\Gamma_i q_i) = 0$ and homoskedasticity, i.e.

$$\sqrt{nh\check{c}/k} \times \frac{E(q_i^2 \phi_i)/\sqrt{E(\phi_i)}}{E(q_i^2)}, \tag{11}$$

where $\check{c}$ is the square integral of the kernel. The first factor in (11) consists entirely of variables which do not depend on and are not chosen as a function of the shape of the local alternative. For a given density ϕ and fixed choices of smoothing parameters and kernel, there always exists some local alternative for which the local power of the Zheng test is negligble compared to ours. Indeed, if ϕ has unbounded support then for $q \propto 1/\sqrt{\phi}$ the ratio is zero.[9]

[7] This analogy is imprecise since h_k depends on i.

[8] This condition is not explicitly imposed by Zheng (1996), but it is necessary.

[9] We assumed $E(q_i^2)$ to be finite, but only to keep ι^2 well–defined; otherwise some limit argument can be used.

Since the null hypothesis allows for weak or partial identification, the local alternatives (9) are not the only interesting sequence of hypotheses. For instance, we may consider the alternative (2) with sequential λ shrinking to 0, which is an example of local alternatives that can accomodate weak identification; in the limit they become the null with no identification. This idea can be generalized to a sequence of uniform local alternatives such as

$$P\big(\mu_i(\theta) = \mu_i^o(\theta) + \delta_n q_i(\theta)\big) = 1, \tag{12}$$

where $\mu_i^o(\theta)$ could be 0 at multiple values of θ and some restrictions are imposed on $q_i(\theta)$.[10] In the working paper version of this paper, we showed that $\hat{T}_2$ is consistent under the sequence of (12), as long as $\delta_n \succ \log n/\sqrt[4]{nk}$. The rate condition of $\delta_n \succ \log n/\sqrt[4]{nk}$ is in fact reminiscent of Jun and Pinkse (2007). In that paper we show that under the null with point identification and sequential λ (i.e. $\mu_i^o = 0$, $\delta_n = \lambda$, $q_i = \mu_i^*$), θ_0 can be consistently estimated only when $\lambda \succ 1/\sqrt[4]{nk}$ (see Jun and Pinkse (2007)). If, in the case of point identification, θ_0 cannot be estimated consistently, then it is intuitive that departures from the null cannot be detected.

Assuming strong point identification under the null, singular (or high frequency) local alternatives have also been studied in the literature (e.g. Rosenblatt (1975), Fan and Li (2000), Guerre and Lavergne (2005)). In particular, Fan and Li (2000) showed that kernel–based tests have more power against such alternatives than integrated–moment–conditions (ICM) type tests. Theorem 7 illustrates that $\hat{T}_2$ is in fact comparable to the result of Fan and Li (2000) and hence it is more powerful than ICM type tests for singular local alternatives. Consider

$$H_{1sL} : P\big(\mu_i(\theta) = \mu_i^o(\theta) + \tilde{\delta}_n Q_{in}\big) = 1, \tag{13}$$

where $Q_{in} = q\big((z_i - v)/h_n\big)/h_n^{d_z}$ with h_n a bandwidth and $q(s)$ satisfies (i) $\Upsilon_1 = \int ||q(s)||ds < \infty$, (ii) $\Upsilon_2 = \int ||q(s)||^2 ds < \infty$, (iii) $||s||||q(s)|| \to 0$ as $||s|| \to \infty$, (iv) $\sup_s ||q(s)|| \leq M_q$ and (v) $||q(s) - q(t)|| \leq C_q||s - t||$. Conditions (i)–(iv) are those of usual kernel estimation. Condition (v) imposes a Lipschitz condition on $q(s)$. We then have the following theorem.

Theorem 7 *Suppose that assumptions A, B, D, G and I hold. Suppose that $||z_1 - z_2||^{d_z}$ has compact support with density f satisfying $c \leq \inf_z f(z) \leq \sup_z f(z) \leq C$. Suppose that $\tilde{\delta}_n^2 \sim h_n^{d_z}/\sqrt{nk}$ and that $k/n \prec h_n^{d_z} \prec 1$. Then, under (13),*

$$\hat{T}_2 - \sqrt{nk} h_n^{-d_z} \tilde{\delta}_n^2 \frac{f(v)\Upsilon_2}{\sqrt{k/n} D_2} \xrightarrow{d} N(0,1).$$

Theorem 7 shows that $\hat{T}_2$ has non–negligible power as long as $\tilde{\delta}_n$ goes to zero more slowly than $k^{1/4}/n^{3/4}$ and $k/(nh_n^{d_z})$ converges sufficiently slowly. Recall that T_1 requires that $n^{1/2+\alpha} \prec k \prec n$ for some $\alpha > 0$ whereas T_2 only needs $1 \prec k \prec n$. Therefore, T_2 can in fact detect singular alternatives that are arbitrarily close to $n^{-3/4}$. Note here that the rate that is arbitrarily close to $n^{-3/4}$ is also the best that kernel–based tests can detect (see e.g. Fan and Li (2000)).

5 Simulations

We now compare several specification tests in simulation experiments. The main focus is on the behavior of the test statistics under the null of correct specification. Throughout this section the

[10] For instance, functions $q_i(\theta)$ that are proportional to $\mu_i^o(\theta)$ should be contained in the null hypothesis.

null hypothesis is

$$H_0: \quad E(y_i - Y_i\theta_0|z_i) = 0 \quad a.s. \quad \text{for some } \theta_0 \in \Theta. \tag{14}$$

Note that the model is given by conditional moment conditions and that the first stage equation $E(Y_i|z_i)$ is not specified. We compare the performance of six different statistics: $\hat{T}_1(\hat{\theta}_{CUE1}), \hat{T}_2(\hat{\theta}_{CUE2})$, $\hat{T}_2(\hat{\theta}), \hat{T}_k(\hat{\theta}_{CUEk}), \hat{T}_k(\hat{\theta})$, and $\hat{T}_2(\theta_0)$, where $\hat{T}_1(\theta)$ and $\hat{T}_2(\theta)$ are defined in equations (5) and (6), and $\hat{T}_k(\theta)$ is a kernel–based statistic defined in Zheng (1996).[11] $\hat{T}_2(\theta_0)$ is infeasible but it is included as a benchmark. $\hat{\theta}_{CUE1}, \hat{\theta}_{CUE2}$ and $\hat{\theta}_{CUEk}$ denote (not necessarily unique) global minimizers of $\hat{T}_1(\theta), \hat{T}_2(\theta)$, and $\hat{T}_k(\theta)$, respectively. $\hat{T}_2(\hat{\theta})$ and $\hat{T}_k(\hat{\theta})$ represent two–step (or plugin) statistics, where $\hat{\theta}$ is an estimator that is $\sqrt{n}$–consistent for θ_0 under strong point identification when the null hypothesis is satisfied.

Size — Rejection frequencies for several significance levels

α	$\hat{T}_1(\hat{\theta}_{CUE1})$	$\hat{T}_2(\hat{\theta}_{CUE2})$	$\hat{T}_2(\hat{\theta}_{2SLS})$	$\hat{T}_k(\hat{\theta}_{CUEk})$	$\hat{T}_k(\hat{\theta}_{2SLS})$	$\hat{T}_2(\theta_0)$
0.010	0.004	0.012	0.511	0.012	0.480	0.035
0.025	0.007	0.016	0.533	0.022	0.509	0.050
0.050	0.015	0.025	0.551	0.030	0.531	0.070
0.100	0.027	0.049	0.584	0.045	0.559	0.101
0.200	0.053	0.078	0.626	0.075	0.602	0.171
	$\rho = 0.5, \lambda = 1, n = 100, k = 40, h = 0.4$, and $\tilde{g}(z_i) = z_i^2 - 1$.					

Table 1: Conditional moments with strong identification but unconditional moments with weak identification (sub–optimal choice of instruments)

We first investigate the size properties of the six statistics, for which we use the following design:

$$\begin{cases} y_i = Y_i + u_i \\ Y_i = \lambda\tilde{g}(z_i) + v_i, \end{cases}$$

where u_i and v_i are drawn from a mean zero multivariate normal distribution with variances equal to one and covariance equal to ρ, and z_i independently from a standard normal. All simulation results are based on 1,000 replications.

Size — Rejection frequencies for several significance levels

α	$\hat{T}_1(\hat{\theta}_{CUE1})$	$\hat{T}_2(\hat{\theta}_{CUE2})$	$\hat{T}_2(\hat{\theta}_{SP})$	$\hat{T}_k(\hat{\theta}_{CUEk})$	$\hat{T}_k(\hat{\theta}_{SP})$	$\hat{T}_2(\theta_0)$
0.010	0.018	0.018	0.341	0.024	0.362	0.045
0.025	0.027	0.027	0.360	0.030	0.395	0.061
0.050	0.036	0.036	0.382	0.046	0.419	0.077
0.100	0.052	0.056	0.422	0.067	0.455	0.122
0.200	0.077	0.093	0.487	0.106	0.519	0.181
	$\rho = -0.99, \lambda = 0.07, n = 200, k = 69, h = 0.35$, and $\tilde{g}(z_i) = z_i$.					

Table 2: Conditional moments with weak identification

Tables 1 and 2 contain examples that were constructed to demonstrate problems with the use of plugin statistics. In table 1, the first–step estimator used is the 2SLS estimator. Since the

[11] We did not include Guerre and Lavergne's (2005) test here since it does not apply to the designs used here.

2SLS estimator is inconsistent when $\tilde{g}_i$ is orthogonal to z_i — as is the case here — plugging in the 2SLS estimator results in invalid inferences.[12] While it is true that using additional powers of z_i as instruments would improve performance in this particular example, it is still possible that $\tilde{g}_i$ is (close to) orthogonal to all such instruments.[13]

Size — Rejection frequencies under the null for the nominal level 5%

$\dim(z_i)$	n	$\hat{T}_1(\hat{\theta}_{CUE1})$	$\hat{T}_2(\hat{\theta}_{CUE2})$	$\hat{T}_2(\hat{\theta}_{2SLS})$	$\hat{T}_k(\hat{\theta}_{CUEk})$	$\hat{T}_k(\hat{\theta}_{2SLS})$	$\hat{T}_2(\theta_0)$
1	100	0.030	0.035	0.041	0.038	0.046	0.071
	200	0.037	0.036	0.043	0.050	0.056	0.074
	400	0.036	0.032	0.040	0.045	0.050	0.070
	800	0.034	0.036	0.038	0.044	0.048	0.069
2	100	0.050	0.047	0.053	0.049	0.062	0.081
	200	0.039	0.031	0.034	0.034	0.037	0.053
	400	0.051	0.042	0.047	0.054	0.059	0.067
	800	0.046	0.043	0.044	0.044	0.053	0.069
4	100	0.027	0.041	0.050	0.033	0.041	0.058
	200	0.046	0.044	0.046	0.039	0.042	0.054
	400	0.054	0.051	0.055	0.040	0.046	0.060
	800	0.060	0.040	0.041	0.050	0.052	0.054
8	100	0.020	0.056	0.073	0.025	0.045	0.080
	200	0.026	0.048	0.052	0.032	0.046	0.063
	400	0.043	0.051	0.054	0.034	0.042	0.060
	800	0.060	0.060	0.062	0.030	0.034	0.074

Table 3: Regular cases with strong identification

Plugging in a semiparametric estimator based on knn estimation of g_i does not have this problem, but is not generally valid, either, especially when instruments are weak as the experiment represented in table 2 shows. The problem here is that the asymptotic behavior of the semiparametric estimator is nonstandard (and even inconsistent) under weak identification; see Jun and Pinkse (2007).

The main problem with the continuous updating statistics is that their true size can be much less than the nominal size under weak identification (not tabulated). But a conservative test is preferable to an invalid one. Note also that they cease to be conservative in more regular cases; see table 3.

Table 3 summarizes the behavior of the statistics in a more standard situation. Here we used $\lambda = 1$ and a linear $\tilde{g}$ and all statistics appear to have reasonable size properties. Note that the continuous updating versions by definition have lower rejection rates than the corresponding plugin ones. The differences are modest, which can be ascribed to the fact that all estimators are $\sqrt{n}$–consistent here.

The rejection rates under the alternative (tables 4 and 5) are by definition again lower for the continuous updating statistics than for the corresponding plugin versions. The differences can be substantial if $V[m_i(\theta)|z_i]$ is large for θ far from θ_0 because the plugin and continuous updating

[12] To see this point, note that $E\big(z_i(y_i - Y_i\theta)\big) = 0$ for any $\theta \in \mathbb{R}$.

[13] If $\tilde{g}_i$ is vector–valued, the nonorthogonality requirement becomes a maximum rank condition.

Power — Rejection frequencies under the alternative for the nominal level 5%

$\dim(z_i)$	n	$\hat{T}_1(\hat{\theta}_{CUE1})$	$\hat{T}_2(\hat{\theta}_{CUE2})$	$\hat{T}_2(\hat{\theta}_{2SLS})$	$\hat{T}_k(\hat{\theta}_{CUEk})$	$\hat{T}_k(\hat{\theta}_{2SLS})$	$\hat{T}_2(\theta_0)$
1	100	0.168	0.201	0.229	0.205	0.232	0.516
	200	0.350	0.410	0.442	0.404	0.436	0.843
	400	0.710	0.777	0.796	0.761	0.782	0.989
	800	0.953	0.969	0.971	0.963	0.966	1.000
2	100	0.152	0.169	0.183	0.169	0.200	0.425
	200	0.358	0.355	0.372	0.371	0.395	0.753
	400	0.607	0.638	0.654	0.607	0.643	0.971
	800	0.926	0.930	0.933	0.924	0.935	1.000
4	100	0.086	0.105	0.126	0.108	0.141	0.280
	200	0.231	0.249	0.271	0.253	0.280	0.549
	400	0.465	0.444	0.461	0.460	0.501	0.842
	800	0.813	0.770	0.777	0.782	0.803	0.993
8	100	0.020	0.092	0.115	0.054	0.099	0.189
	200	0.048	0.128	0.150	0.095	0.128	0.313
	400	0.090	0.242	0.259	0.224	0.261	0.564
	800	0.261	0.438	0.460	0.454	0.477	0.866

Table 4: Regular cases with strong identification — logarithmic alternative

Power — Rejection frequencies under the alternative for the nominal level 5%

n	$\hat{T}_1(\hat{\theta}_{CUE1})$	$\hat{T}_2(\hat{\theta}_{CUE2})$	$\hat{T}_2(\hat{\theta}_{2SLS})$	$\hat{T}_k(\hat{\theta}_{CUEk})$	$\hat{T}_k(\hat{\theta}_{2SLS})$	$\hat{T}_2(\theta_0)$
100	0.309	0.397	0.441	0.373	0.422	0.489
200	0.617	0.721	0.755	0.688	0.733	0.786
400	0.928	0.970	0.976	0.956	0.969	0.976
800	0.998	1.000	1.000	0.999	0.999	1.000

Table 5: Regular cases with strong identification — inverted normal alternative

estimators can have different probability limits under the alternative. So having a valid test can hurt power. In our experiments we use designs in which the variances under the null and alternative are not very different; we use $y_i = Y_i\theta_0 + 0.2\log(Y_i^2+1) + u_i$ for the experiments of table 4 and $y_i = Y_i\theta_0 + 0.1/\sqrt{\phi^*(z_i)} + u_i$ for those of table 5, where ϕ^* is the standard normal density function. Table 4 suggests that the power of plugin and continuous update versions are similar. Moreover, the performance of kernel and knn versions is similar, albeit that differences in power between kernel and knn versions are likely to arise if instruments have distributions very different from the normal, as in table 5. Although the power difference between kernel and knn estimators is less than one would expect on the basis of the discussion following theorem 6, one should bear in mind that (i) the Zheng test is consistent, (ii) $E(\Gamma_i q_i) \neq 0$ in this example, and (iii) the choice of smoothing parameters affects power. Finally, it is apparent that the leave–one–out version of our test ($\hat{T}_2$) does better than $\hat{T}_1$.

We also conducted a limited set of experiments under a singular local alternative (not tabulated). The intuitive conclusion of these experiments is that the way the local alternative is structured largely determines (local) power. For instance, if q in (13) is chosen the same as the kernel in the

Size — varying with k

k	n	$\hat{T}_1(\hat{\theta}_{CUE1})$	$\hat{T}_2(\hat{\theta}_{CUE2})$	$\hat{T}_2(\hat{\theta}_{2SLS})$	$\hat{T}_k(\hat{\theta}_{CUEk})$	$\hat{T}_k(\hat{\theta}_{2SLS})$	$\hat{T}_2(\theta_0)$
20	100	0.034	0.035	0.042	0.033	0.046	0.067
46	200	0.032	0.036	0.040	0.049	0.056	0.076
81	400	0.026	0.028	0.031	0.043	0.047	0.066
140	800	0.031	0.040	0.042	0.045	0.052	0.072
40	100	0.032	0.028	0.037	0.041	0.052	0.068
69	200	0.037	0.036	0.043	0.050	0.056	0.074
121	400	0.036	0.032	0.040	0.045	0.050	0.070
210	800	0.034	0.036	0.038	0.044	0.048	0.069
60	100	0.047	0.039	0.042	0.043	0.048	0.054
104	200	0.043	0.039	0.048	0.046	0.050	0.075
180	400	0.041	0.037	0.038	0.046	0.051	0.071
315	800	0.040	0.035	0.039	0.045	0.048	0.072

Table 6: Regular cases with strong identification

Zheng test, then the Zheng tests outperform the knn tests. See Guerre and Lavergne (2005) for a substantial simulation study of their statistic under such alternatives.

Finally, we studied the effect of the choice of smoothing parameter on performance. The results for size are in table 6 and those for power in table 7. The design is the same as that used to analyze the size and power properties of the various tests under strong identification. For the kernel–based statistics the bandwidth was chosen equal to $(k/n)^{1/d_z}$. It is evident that there is some variation in the size and power properties of all test statistics and that, while smoothing parameters should always be chosen with some care, all test statistics are fairly insensitive to the choice of smoothing parameter.

6 Conclusion

We proposed two (closely related) generally applicable nonparametric specification tests, which are robust to many (identification) problems unrelated to the hypothesis being tested. We find that the leave–one–out version of our test statistic ($\hat{T}_2$) performs better than the complete quadratic version ($\hat{T}_1$). Our tests are shown to be asymptotically valid regardless of the identification situation. Although two–step tests that use plugin estimators may have substantially more power than ours in some cases, they can be seriously size–distorted. If strong point identification is not in doubt (e.g. in linear regression models without multicollinearity), then two–step statistics are preferable because the power of continuous–updating tests is by definition no greater than that of two–step ones; the only reason to prefer our test to the plugin version is that the test statistic yields an efficient estimator of the parameter of interest under the null hypothesis. In all other cases, continuous updating is preferable to plugin. Our simulation study does not reveal substantial differences in performance between knn and kernel–based statistics. However, a classical local power analysis shows that knn–based tests are preferable because they are local power invariant.

Power — varying with k

k	n	$\hat{T}_1(\hat{\theta}_{CUE1})$	$\hat{T}_2(\hat{\theta}_{CUE2})$	$\hat{T}_2(\hat{\theta}_{2SLS})$	$\hat{T}_k(\hat{\theta}_{CUEk})$	$\hat{T}_k(\hat{\theta}_{2SLS})$	$\hat{T}_2(\theta_0)$
20	100	0.153	0.180	0.194	0.184	0.212	0.455
46	200	0.356	0.381	0.402	0.385	0.418	0.796
81	400	0.707	0.737	0.745	0.726	0.750	0.984
140	800	0.954	0.966	0.968	0.951	0.956	1.000
40	100	0.162	0.197	0.215	0.190	0.209	0.511
69	200	0.375	0.438	0.462	0.418	0.451	0.834
121	400	0.710	0.776	0.795	0.761	0.783	0.989
210	800	0.955	0.972	0.972	0.960	0.964	1.000
60	100	0.206	0.254	0.278	0.237	0.257	0.517
104	200	0.405	0.478	0.505	0.449	0.478	0.852
180	400	0.715	0.813	0.830	0.785	0.802	0.989
315	800	0.945	0.978	0.979	0.972	0.974	1.000

Table 7: Regular cases with strong identification

References

Bierens, Herman J. (1990), A consistent conditional coment test of functional Form, *Econometrica* 58, 1443–1458.

Chao, John C. and Norman R. Swanson (2005), Consistent estimation with a large number of weak instruments, *Econometrica* 73, 1673–1692.

Chernozhukov, Victor, Han Hong, and Elie Tamer (2007), Estimation and confidence regions for parameter sets in econometric models, *Econometrica* 75, 1243–1284.

Delgado, Miguel, Manuel Domínguez and Pascal Lavergne (2006), Consistent tests of conditional moment restrictions, *Annales d'Economie et de Statistique* 81, 33–67.

Eubank, Randall L. and C. H. Spiegelman (1990), Testing the goodness of fit of a linear model via nonparametric regression techniques, *Journal of American Statistical Association* 85, 387–392.

Fan, Yanqin and Qi Li (1996), Consistent model specification tests: Omitted variables and semiparametric functional forms, *Econometrica* 64, 865–890.

Fan, Yanqin and Qi Li (2000), Consistent model specification tests: Kernel–based tests versus Bierens' ICM tests, *Econometric Theory* 2000, 1016–1041.

Guerre, Emmanuel and Pascal Lavergne (2005), Data–driven rate–optimal specification testing in regression models, *Annals of Statistics* 33, 840–870.

Guggenberger, Patrik, Jinyong Hahn, and Kyooil Kim (2008), Specification testing under moment inequalities, *Economics Letters* 99, 375–378.

Härdle, Wolfgang and Enno Mammen (1993), Comparing nonparametric versus parametric regression fits, *Annals of Statistics* 21, 1926–1947.

Horowitz, Joel L. and Vladimir G. Spokoiny (2001), An adaptive, rate–optimal test of parametric mean–regression model against a nonparametric alternative, *Econometrica* 69, 599–631.

Jun, Sung Jae and Joris Pinkse (2007), Weak identification and conditional moment restrictions, *Working Paper*.

Jun, Sung Jae and Joris Pinkse (2008), Semiparametric tests of conditional moment restrictions under weak or partial identification, *Working Paper*.

Koul, Hira L. and Pingping Ni (2004), Minimum distance regression model checking, *Journal of Statistical Planning and Inference* 119, 109–141.

Linton, Oliver (1997), Asymptotic inefficiency of an estimator derived from a kernel–based test statistic, *Econometric Theory* 13, 306–307.

Linton, Oliver (2000), Asymptotic inefficiency of an estimator derived from a kernel–based test statistic, *Econometric Theory* 14, 153–154.

Robinson, Peter M. (1987), Asymptotically efficient estimation in the presence of heteroskedasticity of unknown form, *Econometrica* 55, 875–891.

Robinson, Peter M. (1988), Root–N–consistent semiparametric regression, *Econometrica* 56, 931–954.

Robinson, Peter M. (1991), Consistent nonparametric entropy based testing, *Review of Economic Studies* 58, 437–453.

Rosenblatt, Murray (1975), A quadratic measure of deviation of two–dimensional density estimates and a test of independence, *Annals of Statistics* 3, 1–14.

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.

Stone, Charles J. (1977), Consistent nonparametric regression, *Annals of Statistic* 5, 595–620.

Wolak, Frank A. (1991), The local nature of hypothesis tests involving inequality constraints in nonlinear models, *Econometrica* 59, 981–995.

Wooldridge, Jeffrey M. (1992), A test for functional form against nonparametric alternatives, *Econometric Theory* 8, 452–475.

Zheng, John Xu (1996), A consistent test of functional form via nonparametric estimation techniques, *Journal of Econometrics* 75, 263–289.

A Proof of Theorem 1

Under H_0, the numerators of $T_s(\theta_0)$ are given by

$$N_s(\theta_0) = \sum_i h_{ii}\big(\sum_{t=1}^{d}(m_{it}(\theta_0)^2 - 1)\big) + \sum_i \sum_{j<i}(h_{ij} + h_{ji})\big(\sum_{t=1}^{d} m_{it}(\theta_0)m_{jt}(\theta_0)\big), \tag{15}$$

where $h_{ij} = a_{ij} = \sum_t w_{ti}w_{tj}$ for $s = 1$ and $h_{ij} = w_{ij}$ for $s = 2$; we used the fact that $\mathrm{Var}(m_i(\theta_0)|\mathcal{Z}_n) = I_d$, where $\mathcal{Z}_n$ denotes $z_1, \cdots, z_n$. Suppressing θ_0, let $A_i = \sum_{t=1}^{d} h_{ii}(m_{it}^2 - 1)$ and $B_i = \sum_{t=1}^{d}\sum_{j<i}(h_{ij} + h_{ji})m_{it}m_{jt}$ and note that

$$\xi_i = \frac{A_i + B_i}{\sqrt{\sum_i E(B_i^2|\mathcal{Z}_n)}}$$

forms a martingale difference sequence. Since $T_s = \sum_i \xi_i$, the limiting normal distribution of T_s can be obtained by verifying conditions (i), (ii), and (iii) of lemma D3 in Jun and Pinkse (2007). See also theorem 24.3 in Davidson (1994). We first show that $\sum_i E(B_i^2|Z)$ is bounded away from 0 so that xi_i is well-defined. Let $D_s^2 = \sum_i E(B_i^2|\mathcal{Z}_n) = d\sum_i \sum_{j<i}(h_{ij} + h_{ji})^2$. $C > 0$ will denote a generic constant throughout.

Lemma A1 *$D_1^2 = 2d\sum_{i\neq j} a_{ij}^2 \geq 2d + o(1)$, and $D_2^2 = d\sum_{ij} w_{ij}(w_{ij}+w_{ji}) \geq Cn/k$.*

Proof: First, note that

$$D_1^2 = 4d\sum_i\sum_{j<i} a_{ij}^2 = 2d\sum_i\sum_{j\neq i} a_{ij}^2 \overset{\text{Jensen}}{\geq} \frac{2d}{n(n-1)}\big(\sum_i\sum_{j\neq i} a_{ij}\big)^2.$$

Now,

$$\sum_i\sum_{j\neq i} a_{ij} = \sum_t\sum_i(w_{ti}\sum_{j\neq i} w_{tj}) = \sum_t\sum_i(w_{ti}(1-w_{ti})) \geq n(1-\frac{C}{k}).$$

Therefore,

$$D_1^2 \geq 2d\frac{n}{n-1}(1-\frac{C}{k})^2 \to 2d.$$

For D_2^2, note that $\sum_{ij} w_{ij}(w_{ij}+w_{ji}) = \sum_i\sum_{j<i}(w_{ij}+w_{ji})^2$, where $(w_{ij}+w_{ji})^2 \geq \frac{C}{k}(w_{ij}+w_{ji})$. Therefore,

$$D_2^2 \geq \frac{C}{k}\sum_i\sum_{j<i}(w_{ij}+w_{ji}) = \frac{C}{k}\sum_{ij} w_{ij} = C\frac{n}{k}. \quad \square$$

Lemma A2 *Suppose that*

$$E\Big(\frac{\sum_i h_{ii}^2}{\sum_i\sum_{j<i}(h_{ij}+h_{ji})^2}\Big) \to 0 \quad \textit{and} \quad n^{p/2-1}E\Big(\frac{\sum_i\sum_{j<i}|h_{ij}+h_{ji}|^p}{\big(\sum_i\sum_{j<i}(h_{ij}+h_{ji})^2\big)^{p/2}}\Big) \to 0 \tag{16}$$

for some $2 < p \leq 4$. Then, ξ_i satisfies conditions (ii) and (iii) of lemma D3 in Jun and Pinkse (2007): i.e.

$$\textit{(a)}\ \max_i|\xi_i| \xrightarrow{p} 0 \quad \textit{and} \quad \textit{(b)}\ \sum_i \xi_i^2 \xrightarrow{p} 1. \tag{17}$$

Proof: Let

$$\tilde{A}_i = \frac{A_i}{\sqrt{\sum_i E(B_i^2|\mathcal{Z}_n)}} \quad \text{and} \quad \tilde{B}_i = \frac{B_i}{\sqrt{\sum_i E(B_i^2|\mathcal{Z}_n)}}.$$

Noting that

$$E(\tilde{A}_i^2) \leq CE\Big(\frac{h_{ii}^2}{\sum_i\sum_{j<i}(h_{ij}+h_{ji})^2}\Big) \quad \text{and} \quad E(|\tilde{B}_i|^p) \leq Cn^{p/2-1}E\Big(\frac{\sum_{j<i}|h_{ij}+h_{ji}|^p}{\big(\sum_i\sum_{j<i}(h_{ij}+h_{ji})^2\big)^{p/2}}\Big)$$

due to assumption B, the Burkholder and the C_r inequalities, we will show that

$$\sum_i E(\tilde{A}_i^2) \to 0 \quad \text{and} \quad \sum_i E(|\tilde{B}_i|^p) \to 0 \tag{18}$$

imply conditions (17). Since $\max_i|\xi_i| \leq \max_i|\tilde{A}_i| + \max_i|\tilde{B}_i|$, it follows from the Bonferroni and Markov inequalities that

$$P(\max_i|\xi_i| > 2\epsilon) \leq \sum_i P(|\tilde{A}_i| > \epsilon) + \sum_i P(|\tilde{B}_i| > \epsilon) \leq \frac{1}{\epsilon^2}\sum_i E(\tilde{A}_i^2) + \frac{1}{\epsilon^p}\sum_i E(|\tilde{B}_i|^p).$$

Therefore, condition (a) follows from (18). For condition (b), note that

$$P(|\sum_i \tilde{B}_i^{\,2} - 1| > \epsilon) \le \frac{1}{\epsilon^{p/2}} E\Big(\Big|\sum_i (\tilde{B}_i^{\,2} - E(\tilde{B}_i^2|\mathcal{Z}_n))\Big|^{p/2}\Big)$$
$$\le \frac{C}{\epsilon^{p/2}} E\Big((\sum_i \Big|\tilde{B}_i^2 - E(\tilde{B}_i^2|\mathcal{Z}_n)\Big|^2)^{p/4}\Big) \le \frac{C}{\epsilon^{p/2}} E\big((\sum_i \tilde{B}_i^4)^{p/4}\big) \le \frac{C}{\epsilon^{p/2}} \sum_i E(|\tilde{B}_i|^p) \to 0, \quad (19)$$

using the Burkholder inequality, the C_r inequality and condition (18). Now, condition (b) follows from

$$|\sum_i \xi_i^2 - 1| \overset{\text{Schwarz}}{\le} |\sum_i \tilde{A}_i^2| + |\sum_i \tilde{B}_i^2 - 1| + 2\sqrt{\sum_i \tilde{A}_i^2}\sqrt{\sum_i \tilde{B}_i^2}. \qquad \square$$

Lemma A3 *ξ_i satisfies condition (i) of lemma D3 in Jun and Pinkse (2007): i.e.* $\sup_n E(\max_{i\le n} \xi_i^2) < \infty$.

Proof: Since $\xi_i^2 \le 2(\tilde{A}_i^2 + \tilde{B}_i^2)$, we know that

$$E(\max_{i\le n} \xi_i^2|\mathcal{Z}_n) \le 2\sum_i E(\tilde{A}_i^2|\mathcal{Z}_n) + 1. \quad (20)$$

For $h_{ij} = w_{ij}$, the lemma is trivially true, because $\tilde{A}_i = 0$. For $h_{ij} = a_{ij}$, note that

$$\sum_i E(\tilde{A}_i^2|\mathcal{Z}_n) = \frac{\sum_{t,s=1}^d \sum_i a_{ii}^2 \big(E(m_{it}^2 m_{is}^2|\mathcal{Z}_n) - 1\big)}{2d^2 \sum_i \sum_{j\ne i} a_{ij}^2} \overset{\text{B, Schwarz}}{\le} C\frac{\sum_i a_{ii}^2}{\sum_i \sum_{j\ne i} a_{ij}^2}.$$

Note here that $\sum_i a_{ii}^2 = \sum_{its} w_{ti}^2 w_{si}^2 \le Cn^2/k^3 \to 0$ because k is chosen such that $n^3 \prec k^4$ when $h_{ij} = a_{ij}$. Therefore, using lemma A1, we know that there is a non-stochastic sequence γ_n converging to 0 such that $\sum_i E(\tilde{A}_i^2|\mathcal{Z}_n) \le \gamma_n \quad a.s.$. It then follows that for any constant $C > 0$, there exists $N \in \mathbb{N}$ such that $\sum_i E(\tilde{A}_i^2|\mathcal{Z}_n) \le \max\{\gamma_1, \gamma_2, \cdots, \gamma_N, C\} \quad a.s.$. Combining this with (20) completes the proof. $\square$

Lemma A4 $\sup_a E(w_{21}w_{31}w_{41}w_{51}|z_1 = a) = O(n^{-4})$.

Proof: The proof is similar to that of lemma B3 of Jun and Pinkse (2007). Here, we only consider the case that $P(z_1 = a) = 0$. As in Jun and Pinkse (2007), define $\tau(a,b) = P(||z-b|| \le ||a-b||)$ and $A(a) = \{b : \tau(a,b) \le 2k/n\}$. Let $\mathcal{N}_i$ denote the set of neighbors of z_i. Since for any a,

$$E(w_{21}w_{31}w_{41}w_{51}|z_1 = a) \le \frac{C}{k^4} E(I(z_1 \in \mathcal{N}_2)I(z_1 \in \mathcal{N}_3)I(z_1 \in \mathcal{N}_4)I(z_1 \in \mathcal{N}_5)|z_1 = a),$$

it suffices to show that $\sup_a E\big(I(z_1 \in \mathcal{N}_2)I(z_1 \in \mathcal{N}_3)I(z_1 \in \mathcal{N}_4)I(z_1 \in \mathcal{N}_5)|z_1 = a\big) = O(k^4/n^4)$. Let $S_j(a,b) = \sum_{i\ne j,1} I(||z_i - b|| \le ||a-b||)$. Note then that

$$I(z_1 \in \mathcal{N}_2) \le I(z_1 \in \mathcal{N}_2)I(\tau(z_1,z_2) > 2k/n) + I(\tau(z_1,z_2) \le 2k/n)$$
$$\le I\big(S_2(z_1,z_2) < k\big)I\big(\tau(z_1,z_2) > 2k/n\big) + I\big(\tau(z_1,z_2) \le 2k/n\big)$$
$$\le I\big(|S_2(z_1,z_2) - (n-2)\tau(z_1,z_2)| > k\big) + I\big(\tau(z_1,z_2) \le 2k/n\big).$$

It then follows that

$$
\begin{aligned}
&E\big(I(z_1 \in \mathcal{N}_2)I(z_1 \in \mathcal{N}_3)I(z_1 \in \mathcal{N}_4)I(z_1 \in \mathcal{N}_5)|z_1 = a\big) \\
&\quad \le E\big(I(\tau(a,z_2) \le 2k/n)I(\tau(a,z_3) \le 2k/n)I(\tau(a,z_4) \le 2k/n)I(\tau(a,z_5) \le 2k/n)|z_1 = a\big) \\
&\qquad + C_1 P\big(|S_2(a,z_2) - (n-2)\tau(a,z_2)| > k\big) + C_2 P\big(|S_3(a,z_3) - (n-2)\tau(a,z_3)| > k\big) \\
&\qquad\quad + C_3 P\big(|S_4(a,z_4) - (n-2)\tau(a,z_4)| > k\big) + C_4 P\big(|S_5(a,z_5) - (n-2)\tau(a,z_5)| > k\big).
\end{aligned}
$$

The RHS2–RHS5 are all the same, and they are bounded by $\exp(-2k^2/(n-2))$ by the Hoeffding inequlity. The RHS1 is bounded by $\sup_a P(z_2 \in A(a))^4 = O(k^4/n^4)$ by lemma B2 of Jun and Pinkse (2007). □

Lemma A5 *Under H_0, $T_1(\theta_0) \xrightarrow{d} N(0,1)$ and $T_2(\theta_0) \xrightarrow{d} N(0,1)$.*

Proof: In view of lemma A3, it suffices to show that the two conditions of (16) are satisfied. First, consider $T_2(\theta_0)$. The first condition of (16) is trivially satisfied, because $w_{ii} = 0$. For the second condition of (16), use lemma A1 to obtain

$$
n^{p/2-1} E\Big(\frac{\sum_i \sum_{j<i} |w_{ij} + w_{ji}|^p}{\big(\sum_i \sum_{j<i}(w_{ij} + w_{ji})^2\big)^{p/2}}\Big) \le C n^{-1} k^{-p/2+1} E(\sum_{ij} w_{ij}) = C k^{-p/2+1} \to 0.
$$

Now $T_1(\theta_0)$. Since $\sum_i \sum_{j<i}(a_{ij}+a_{ji})^2$ is bounded away from 0 by lemma A1, it suffices to show that $E(\sum_i a_{ii}^2) = o(1)$ and $n^{\frac{p}{2}-1} E(\sum_i \sum_{j<i} a_{ij}^p) = o(1)$. Note first that $\sum_i E(a_{ii}^2) = \sum_{its} E(w_{ti}^2 w_{si}^2) = O(n/k^2) = o(1)$ because k is chosen such that $n^3 \prec k^4$ for $\hat{T}_1$. Further,

$$
E(a_{ij}^4) = \sum_{t_1 t_2 t_3 t_4} E(w_{t_1 i} w_{t_1 j} w_{t_2 i} w_{t_2 j} w_{t_3 i} w_{t_3 j} w_{t_4 i} w_{t_4 j}) \le C k^{-4} \sum_{t_1 t_2 t_3 t_4} E(w_{t_1 i} w_{t_2 i} w_{t_3 i} w_{t_4 i}) \overset{\text{A4}}{=} O(k^{-4}),
$$

which implies that

$$
n^{\frac{p}{2}-1} E(\sum_i \sum_{j<i} a_{ij}^p) \overset{\text{Jensen}}{\le} n^{\frac{p}{2}-1} \sum_i \sum_{j<i} E(a_{ij}^4)^{\frac{p}{4}} = O(n^{\frac{p}{2}+1}/k^p).
$$

Since $2 < p \le 4$ is arbitrary, using $p = 4$ completes the proof. □

Lemma A6 *$||\hat{V}(\theta_0)^{-1/2} - V(\theta_0)^{-1/2}||$ is either $o_p(k/n)$ or $o(1)$, depending on $n^{3/4} \prec k \prec n$ or $1 \prec k \prec n$.*

Proof: We only consider the case of $n^{3/4} \prec k \prec n$. Suppressing θ_0, let $\hat{\varsigma}_{ts}$ and ς_{ts} be the t–s elements of $\hat{V}$ and V, respectively. Note that

$$
n(\hat{\varsigma}_{ts} - \varsigma_{ts}) = \underbrace{\sum_i (\tilde{m}_{it}\tilde{m}_{is} - \varsigma_{ts})}_{=O_p(\sqrt{n})=o_p(k)} - \underbrace{\sum_{ij} w_{ij}\tilde{m}_{it}\tilde{m}_{js}}_{=O_p(\sqrt{n/k})=o_p(k)} - \underbrace{\sum_{ij} w_{ij}\tilde{m}_{is}\tilde{m}_{jt}}_{=O_p(\sqrt{n/k})=o_p(k)} + \underbrace{\sum_{ijr} w_{ij}w_{ir}\tilde{m}_{jt}\tilde{m}_{rs}}_{=O_p(n/k)=o_p(k)}.
$$

Since d is finite, it follows that $||\hat{V} - V|| = o_p(k/n)$. Now, note that

$$
||\hat{V}^{-1} - V^{-1}|| \le ||V^{-1}||||\hat{V}^{-1}||||V - \hat{V}|| \le ||V^{-1}||\Big(||\hat{V}^{-1} - V^{-1}|| + ||V^{-1}||\Big)||V - \hat{V}||,
$$

which implies that

$$\Big(1 - o_p(k/n)\Big)||\hat{V}^{-1} - V^{-1}|| \leq o_p(k/n),$$

because the smallest eigenvalue of V is bounded away from 0. Then, the lemma follows from the delta method. □

Proof of Theorem 1: Since $P(\hat{T}_s > q_\alpha) \leq P(\hat{T}_s(\theta_0) > q_\alpha)$ for $s = 1, 2$, it suffices to show that $\hat{T}_s(\theta_0) = T_s(\theta_0) + o_p(1)$ due to lemma A5. Suppressing θ_0, let $\hat{\varsigma}^{t_1t_2}$ and $\varsigma^{t_1t_2}$ denote the t_1–t_2 elements of $\hat{V}^{-1/2}$ and $V^{-1/2}$, respectively. Note that

$$\begin{aligned}|\hat{T}_s - T_s| &= \Big| \sum_{t_1,t_2,t_3=1}^{d} \big(\hat{\varsigma}^{t_1t_2}\hat{\varsigma}^{t_1t_3} - \varsigma^{t_1t_2}\varsigma^{t_1t_3}\big)\big(\sum_{ij} h_{ij}\tilde{m}_{it_2}m_{jt_3}\big)\Big|/D_s \\ &\leq d^3 \max_{t_1,t_2,t_3} \Big|\hat{\varsigma}^{t_1t_2}\hat{\varsigma}^{t_1t_3} - \varsigma^{t_1t_3}\varsigma^{t_1t_3}\Big| \max_{t_2,t_3} \Big|\sum_{ij} h_{ij}\tilde{m}_{it_2}\tilde{m}_{jt_3}\Big|/D_s. \quad (21)\end{aligned}$$

Note here that

$$\Big|\sum_{ij} h_{ij}\tilde{m}_{it_2}\tilde{m}_{jt_3}\Big| \leq \Big|\sum_{ij} h_{ij}\tilde{m}_{it_2}\tilde{m}_{jt_3} - \sum_i h_{ii}E\big(\tilde{m}_{it_2}\tilde{m}_{it_3}|z_i\big)\Big| + \Big|\sum_i h_{ii}E\big(\tilde{m}_{it_2}\tilde{m}_{it_3}|z_i\big)\Big|,$$

where the RHS1 divided by D_s is $O_p(1)$, because

$$\sum_i \sum_{j<i} \Big(h_{ij}\tilde{m}_{it_2}\tilde{m}_{jt_3} + h_{ji}\tilde{m}_{jt_2}\tilde{m}_{it_3}\Big) + \sum_i h_{ii}\Big(\tilde{m}_{it_2}\tilde{m}_{it_3} - E\big(\tilde{m}_{it_2}\tilde{m}_{it_3}|z_i\big)\Big)$$

is the sum of martingale difference sequences and theorem 24.3 of Davidson (1994) applies. Therefore, in view of lemma A6, it suffices to show that $\Big|\sum_i h_{ii}E\big(\tilde{m}_{it_1}\tilde{m}_{it_2}|z_i\big)\Big|/D_s = O_p(n/k)$. This is trivially true for $s = 2$, because $h_{ii} = w_{ii} = 0$. When $s = 1$, recall that D_1 is bounded below by a nonzero constant by lemma A1 and the conclusion follows from

$$\sum_i E\Big(a_{ii}E\big(\tilde{m}_{it_1}\tilde{m}_{it_2}|z_i\big)\Big) \leq C_w k^{-1} \sum_{ij} E\Big(w_{ji}E\big(\tilde{m}_{it_1}\tilde{m}_{it_2}|z_i\big)\Big) = O_p(n/k). \qquad \square$$

B Proof of theorem 2

Following Robinson (1987), let $\tilde{\sigma}_j^2 = \sum_t w_{jt}\sigma_t^2$.

Lemma B1 $\max_i |\hat{\sigma}_i^2 - \tilde{\sigma}_i^2| = O_p(n^{1/p^*}k^{-1/2})$.

Proof: Follows from Robinson (1987), lemma 9. □

Lemma B2 $\lim_{n\to\infty} P[\min_i \hat{\sigma}_i^2 < C_s/2] = 0$.

Proof: Follows from Robinson (1987), lemma 10. □

Lemma B3 $\sum_{ij} w_{ij}\tilde{m}_i\tilde{m}_j\big(1/(\hat{\sigma}_i\hat{\sigma}_j) - 1/(\tilde{\sigma}_i\tilde{\sigma}_j)\big) = o_p(\sqrt{n/k})$.

Proof: First,

$$\sum_{ij} w_{ij}\tilde{m}_i\tilde{m}_j\big(\frac{1}{\hat{\sigma}_i\hat{\sigma}_j}-\frac{1}{\tilde{\sigma}_i\tilde{\sigma}_j}\big)=\sum_{ij} w_{ij}\tilde{m}_i\tilde{m}_j\big(\frac{1}{\hat{\sigma}_i}-\frac{1}{\tilde{\sigma}_i}\big)\big(\frac{1}{\hat{\sigma}_j}-\frac{1}{\tilde{\sigma}_j}\big)$$
$$+\sum_{ij} w_{ij}\tilde{m}_i\tilde{m}_j\big(\frac{1}{\hat{\sigma}_i}-\frac{1}{\tilde{\sigma}_i}\big)\frac{1}{\tilde{\sigma}_j}+\sum_{ij} w_{ij}\tilde{m}_i\tilde{m}_j\big(\frac{1}{\hat{\sigma}_j}-\frac{1}{\tilde{\sigma}_j}\big)\frac{1}{\tilde{\sigma}_i}. \quad (22)$$

RHS3 is harder than RHS1 and similar to RHS2, so we only show that RHS3 is $o_p(\sqrt{n/k})$. Note that $1/\sigma=1/\sqrt{\sigma^2}$, such that by the mean value theorem,

$$\Big|\frac{1}{\hat{\sigma}_j}-\frac{1}{\tilde{\sigma}_j}+\frac{1}{2\tilde{\sigma}_j^3}(\hat{\sigma}_j^2-\tilde{\sigma}_j^2)-\frac{3}{8\tilde{\sigma}_j^5}(\hat{\sigma}_j^2-\tilde{\sigma}_j^2)^2\Big|\leq\frac{5}{16\min(\tilde{\sigma}_j^7,\hat{\sigma}_j^7)}|\hat{\sigma}_j^2-\tilde{\sigma}_j^2|^3, \quad (23)$$

Now, by lemmas B1 and B2, $\max_j|\hat{\sigma}_j^2-\tilde{\sigma}_j^2|^3/\min(\tilde{\sigma}_j^7,\hat{\sigma}_j^7)=O_p(n^{3/p^*}k^{-3/2})$ and $E\big|\tilde{m}_j\sum_i w_{ij}\tilde{m}_i/\tilde{\sigma}_i\big|\leq\sqrt{nE\tilde{m}_j^2E(w_{ij}^2\tilde{m}_i^2/\tilde{\sigma}_i^2)}=O(1/\sqrt{k})$, such that

$$\sum_j\Big|\frac{\tilde{m}_j(\hat{\sigma}_j^2-\tilde{\sigma}_j^2)^3}{\min(\tilde{\sigma}_j^7,\hat{\sigma}_j^7)}\sum_i w_{ij}\frac{\tilde{m}_i}{\tilde{\sigma}_i}\Big|\leq\max_j\frac{|\hat{\sigma}_j^2-\tilde{\sigma}_j^2|^3}{\min(\tilde{\sigma}_j^7,\hat{\sigma}_j^7)}\sum_j\Big|\tilde{m}_j\sum_i w_{ij}\frac{\tilde{m}_i}{\tilde{\sigma}_i}\Big|=O_p(n^{3/p^*+1}k^{-2})=o_p(\sqrt{n/k}).$$

Further,

$$E\left(\sum_{ij} w_{ij}\frac{\tilde{m}_i\tilde{m}_j(\hat{\sigma}_j^2-\tilde{\sigma}_j^2)}{2\tilde{\sigma}_i\tilde{\sigma}_j^3}\right)^2=E\left(\sum_{ijt} w_{ij}w_{jt}\frac{\tilde{m}_i\tilde{m}_j(\tilde{m}_t^2-\sigma_t^2)}{2\tilde{\sigma}_i\tilde{\sigma}_j^3}\right)^2$$
$$=\sum_{ijt}E\left(w_{ij}^2w_{jt}^2\Big(\frac{\tilde{m}_i\tilde{m}_j(\tilde{m}_t^2-\sigma_t^2)}{2\tilde{\sigma}_i\tilde{\sigma}_j^3}\Big)^2\right)+\text{ similar terms }=O(n^3/k^4)=o(n/k). \quad (24)$$

Finally, $\sum_{ij} w_{ij}\tilde{m}_i\tilde{m}_j(\hat{\sigma}_j^2-\tilde{\sigma}_j^2)^2/(\tilde{\sigma}_i\tilde{\sigma}_j^5)$ can be dealt with an argument similar to the one used in (24). □

Lemma B4 $\sum_{ij} w_{ij}\tilde{m}_i\tilde{m}_j\big(1/(\tilde{\sigma}_i\tilde{\sigma}_j)-1/(\sigma_i\sigma_j)\big)=o_p(\sqrt{n/k})$.

Proof: Square and take expectation to obtain

$$\sum_{ij}E\big((w_{ij}^2+w_{ij}w_{ji})\tilde{m}_i^2\tilde{m}_j^2(\tilde{\sigma}_i^{-1}\tilde{\sigma}_j^{-1}-\sigma_i^{-1}\sigma_j^{-1})^2\big)\leq(C/k)\sum_{ij}E\big(w_{ij}\tilde{m}_i^2\tilde{m}_j^2(\tilde{\sigma}_i\tilde{\sigma}_j-\sigma_i\sigma_j)^2\big),$$

where we use the fact that both $\tilde{\sigma}_i$ and σ_i are bounded away from zero. Now, using assumption B, $(\tilde{\sigma}_i\tilde{\sigma}_j-\sigma_i\sigma_j)^2\leq|\tilde{\sigma}_i^2\tilde{\sigma}_j^2-\sigma_i^2\sigma_j^2|\leq C\big(|\tilde{\sigma}_i^2-\sigma_i^2|+|\tilde{\sigma}_j^2-\sigma_j^2|\big)$. Now, by the Schwarz inequality,

$$k^{-1}\sum_{ij}E\big(w_{ij}\tilde{m}_i^2\tilde{m}_j^2|\tilde{\sigma}_i^2-\sigma_i^2|\big)\leq k^{-1}\sqrt{\sum_i E\big(\sum_j w_{ij}\tilde{m}_j^2\tilde{m}_i^2\big)^2\sum_i E|\tilde{\sigma}_i^2-\sigma_i^2|}=o(n/k). \quad \square$$

Proof of Theorem 2: From lemmas B3 and B4 it follows that

$$\sum_{ij} w_{ij}\frac{\tilde{m}_i}{\tilde{\sigma}_i}\frac{\tilde{m}_j}{\tilde{\sigma}_j}-\sum_{ij} w_{ij}\frac{\tilde{m}_i}{\sigma_i}\frac{\tilde{m}_j}{\sigma_j}=o_p(\sqrt{n/k}), \quad (25)$$

which is sufficient for the stated result. □

C Proof of Theorem 3

Let $\mu_i(\theta) = E\big(m_i(\theta)|z_i\big) = \lambda\mu_i^*(\theta)$. Our statistic at θ is given by $T_s(\theta) = N_s(\theta)/D_s$, where

$$N_s(\theta) = \sum_{t=1}^{d}\sum_{ij} h_{ij}m_{it}(\theta)m_{jt}(\theta) - d\sum_i h_{ii}.$$

We will write $u_i(\theta)$ for $m_i(\theta) - \mu_i(\theta)$ in the following discussion.

Lemma C1 *Let $f(y_i,\theta) \in \mathcal{F}$. Then, $\sup_{\theta\in\Theta}\|n^{-1}\sum_i f(y_i,\theta) - E\big(f(y_i,\theta)\big)\| = o_p(1)$.*

Proof: Without loss of generality, we will assume that $E(f_i(\theta)) = 0$, where $f_i(\theta) = f(y_i,\theta)$. Choose an arbitrary $\eta > 0$, and let $\delta > 0$ satisfy (8). Since Θ is compact, it can be divided up into δ-balls $\Theta_1,\cdots,\Theta_{K_\delta}$ for $K_\delta < \infty$. Letting θ_κ be the center of Θ_κ, we have

$$\begin{aligned}\sup_{\theta\in\Theta}\|n^{-1}\sum_i f_i(\theta)\| &\le \max_{\kappa=1,\cdots,K_\delta} n^{-1}\sum_i \sup_{\theta\in\Theta_\kappa}\|f_i(\theta) - f_i(\theta_\kappa)\| + \max_{\kappa=1,\cdots,K_\delta}\|n^{-1}\sum_i f_i(\theta_\kappa)\| \\ &\le \eta + \sum_{\kappa=1}^{K_\delta}\|n^{-1}\sum_i f_i(\theta_\kappa)\|\end{aligned}$$

with probability greater than $1-\eta$. Since K_δ is finite, and $\eta > 0$ is arbitrariy, applying the law of large numbers completes the proof. □

Lemma C2 *Let $f(z_i,\theta)$ be a function of z_i and θ such that $f \in \mathcal{F}$ and $E(\sup_{\theta\in\Theta}|f(z_i,\theta)|^{p_f}) < \infty$ for some $0 < p_f < \infty$. Then, for any $q \ge 1$, $\sup_\theta \sum_{ij} w_{ij}\|f(z_i,\theta) - f(z_j,\theta)\|^q = o_p(n)$.*

Proof: Without loss of generality, we assume that f is real-valued. Choose an arbitrary $\eta > 0$, and let $\delta > 0$ satisfy (8). Since Θ is compact, it can be divided up into δ-balls $\Theta_1,\cdots,\Theta_{K_\delta}$ for $K_\delta < \infty$. Now, let $\Delta_{j\kappa}(\theta) = |\sup_{\theta\in\Theta_\kappa} f(z_j,\theta) - \inf_{\theta\in\Theta_\kappa} f(z_j,\theta)|^q$. Letting $f_j(\theta) = f(z_j,\theta)$ and suppressing θ, we have

$$\begin{aligned}&\sup_{\theta\in\Theta} n^{-1}\sum_{ij} w_{ij}|f_i - f_j|^q \le \max_{\kappa=1,2,\cdots,K_\delta} n^{-1}\sum_{ij} w_{ij}\sup_{\theta\in\Theta_\kappa}|f_i - f_j|^q \\ &\le C\big(\max_{\kappa=1,2,\cdots,K_\delta} n^{-1}\sum_i \Delta_{i\kappa} + \max_{\kappa=1,2,\cdots,K_\delta} n^{-1}\sum_{ij} w_{ij}\Delta_{j\kappa} + \max_{\kappa=1,2,\cdots,K_\delta} n^{-1}\sum_{ij} w_{ij}|\inf_{\theta\in\Theta_\kappa} f_i - \inf_{\theta\in\Theta_\kappa} f_j|^q\big).\end{aligned}$$

Here, the RHS3 is $o_p(1)$ by Stone's lemma. Further, the RHS2 is bounded by

$$\max_{\kappa=1,2,\cdots,K_\delta} n^{-1}\sum_i \Delta_{i\kappa} + \sum_{\kappa=1}^{K_\delta} n^{-1}\sum_{ij} w_{ij}|\Delta_{j\kappa} - \Delta_{i\kappa}| = \max_{\kappa=1,2,\cdots,K_\delta} n^{-1}\sum_i \Delta_{i\kappa} + o_p(1)$$

again by Stone. Therefore, it suffices to show that the RHS1 is negligible. Note that

$$\max_{\kappa=1,2,\cdots,K_\delta} n^{-1}\sum_i \Delta_{i\kappa} \le \max_{\kappa=1,2,\cdots,K_\delta} E(\Delta_{i\kappa}) + \sum_{\kappa=1}^{K_\delta}|n^{-1}\sum_i \Delta_{i\kappa} - E(\Delta_{i\kappa})| = \max_{\kappa=1,2,\cdots,K_\delta} E(\Delta_{i\kappa}) + o_p(1),$$

where

$$\max_{\kappa=1,2,\cdots,K_\delta} E(\Delta_{i\kappa}) \le \eta + \max_{\kappa=1,2,\cdots,K_\delta} E(\Delta_{i\kappa} I(\Delta_{i\kappa} > \eta))$$
$$\le \eta + \max_{\kappa=1,2,\cdots,K_\delta} E(\Delta_{i\kappa}^{p_f})^{1/p_f} \max_{\kappa=1,2,\cdots,K_\delta} P(\Delta_{i\kappa} > \eta)^{(p_f-1)/p_f} \le \eta + \big(2E(\sup_{\theta\in\Theta} |f_i|^{p_f})\big)^{1/p_f} (\eta)^{(p_f-1)/(p_f q)}.$$

Since η is arbitrary, the lemma is proved. □

Lemma C3 *(i)* $\sup_{\theta\in\Theta} \sum_{ij} w_{ij}u_{it}(\theta)u_{jt}(\theta) = o_p(n)$ *and (ii)* $\sup_{\theta\in\Theta} \sum_i \big(\sum_j w_{ij}u_{jt}(\theta)\big)^2 = o_p(n)$ *for each* $t = 1, 2, \cdots, d$.

Proof: Since $u_i \in \mathcal{F}$, for any $\eta > 0$, we can choose $\delta > 0$ as (8). Divide Θ up into δ-balls $\Theta_1, \cdots, \Theta_{K_\delta}$ for $K_\delta < \infty$, and let θ_κ be the center of Θ_κ. Note then

$$\sup_{\theta\in\Theta_\kappa} |u_{jt}(\theta) - u_{jt}(\theta_\kappa)| \le \eta \quad \text{and} \quad \sup_{\theta\in\Theta_\kappa} |u_{jt}(\theta) - u_{jt}(\theta_\kappa)| \sup_{\theta\in\Theta_\kappa} |u_{it}(\theta) - u_{it}(\theta_\kappa)| \le \eta^2$$

with probability greater than $1-\eta$. Therefore,

$$|\sup_{\theta\in\Theta} n^{-1} \sum_{ij} w_{ij}u_{it}(\theta)u_{jt}(\theta)| \le \eta^2 + \eta \max_{\kappa=1,\cdots,K_\delta} n^{-1} \sum_{ij} w_{ij}|u_{jt}(\theta_\kappa)|$$
$$+ \eta \max_{\kappa=1,\cdots,K_\delta} n^{-1} \sum_i |u_{it}(\theta_\kappa)| + \max_{\kappa=1,\cdots,K_\delta} |n^{-1} \sum_{ij} w_{ij}u_{it}(\theta_\kappa)u_{jt}(\theta_\kappa)|$$

with probability greater than $1-\eta$. Here, the RHS2 is $O_p(\eta)$, because

$$\max_{\kappa=1,\cdots,K_\delta} n^{-1} \sum_{ij} w_{ij}|u_{jt}(\theta_\kappa)| \le n^{-1} \sum_{ij} w_{ij} (\sum_{\kappa=1}^{K_\delta} |u_{jt}(\theta_\kappa)|)$$
$$= n^{-1} \sum_i E(\sum_{\kappa=1}^{K_\delta} |u_{it}(\theta_\kappa)||z_i) + o_p(1) = E(\sum_{\kappa=1}^{K_\delta} |u_{it}(\theta_\kappa)|) + o_p(1) = O(1).$$

The RHS3 is also $O_p(\eta)$ by the law of large numbers. Lastly, squaring and taking expectation of the RHS4 shows that it is $O_p(n^{-1/2}k^{-1/2}) = o_p(1)$. Taking $\eta \to 0$ proves the first statement of the lemma. The second statement of the lemma follows similarly and it will be omitted. □

Lemma C4 $\sup_{\theta\in\Theta} \sum_i (\hat{\mu}_{it}(\theta) - \mu_{it}(\theta))^2 = o_p(n)$ *for each* $t = 1, 2, \cdots, d$.

Proof: Since $(\hat{\mu}_{it}(\theta) - \mu_{it}(\theta))^2 \le 2\Big(\big(\sum_j w_{ij}u_{jt}(\theta)\big)^2 + \big(\sum_j w_{ij}(\mu_{jt}(\theta) - \mu_{it}(\theta))\big)^2\Big)$, we know that

$$\sup_{\theta\in\Theta} \sum_i (\hat{\mu}_{it}(\theta) - \mu_{it}(\theta))^2 \le 2 \sup_{\theta\in\Theta} \sum_i \Big(\sum_j w_{ij}u_{jt}(\theta)\Big)^2 + 2 \sup_{\theta\in\Theta} \sum_i \Big(\sum_j w_{ij}\big(\mu_{jt}(\theta) - \mu_{it}(\theta)\big)\Big)^2$$
$$\overset{\text{Jensen}}{\le} \underbrace{2 \sup_{\theta\in\Theta} \sum_i \Big(\sum_j w_{ij}u_{jt}(\theta)\Big)^2}_{\overset{\text{C3}}{=} o_p(n)} + \underbrace{2 \sup_{\theta\in\Theta} \sum_{ij} w_{ij}\big(\mu_{jt}(\theta) - \mu_{it}(\theta)\big)^2}_{\overset{\text{C2}}{=} o_p(n\lambda^2)}. \quad \square$$

Lemma C5 *(i)* $\sup_{\theta\in\Theta}|\sum_i u_{it}(\theta)(\hat{\mu}_{it}(\theta)-\mu_{it}(\theta))| = o_p(n)$ *and*
(ii) $\sup_{\theta\in\Theta}|\sum_i \mu_{it}(\theta)(\hat{\mu}_{it}(\theta)-\mu_{it}(\theta))| = o_p(n\lambda)$ *for each* $t = 1, 2, \cdots, d$.

Proof: Note that

$$\sup_{\theta\in\Theta}|\sum_i u_{it}(\theta)(\hat{\mu}_{it}(\theta)-\mu_{it}(\theta))| \overset{\text{Schwarz}}{\leq} \big(\sup_{\theta\in\Theta}\sum_i u_{it}(\theta)^2\big)^{1/2}\big(\sup_{\theta\in\Theta}\sum_i(\hat{\mu}_{it}(\theta)-\mu_{it}(\theta))^2\big)^{1/2}$$
$$\overset{\text{C1, C4}}{=} \sqrt{O_p(n)(o_p(n)+o_p(n\lambda^2))}.$$

The second statement is similar and it will be omitted. □

Lemma C6 $0 < C_1k^2/n^2 \leq D_1^{-2} \leq O(1)$ *and* $0 < C_2k/n \leq D_2^{-2} \leq C_3k/n$.

Proof: Note that

$$D_1^2 = 2d\sum_{i\neq j} a_{ij}^2 \leq 2d\sum_{ijts} w_{ti}w_{tj}w_{si}w_{sj} \leq Ck^{-1}\sum_{its} w_{ti}w_{si} \leq Cn^2/k^2,$$
$$D_2^2 = d\sum_{ij} w_{ij}(w_{ij}+w_{ji}) \leq Ck^{-1}\sum_{ij} w_{ij} = Cn/k.$$

Therefore, the conclusion follows from lemma A1. □

Lemma C7 $\sup_\theta ||\hat{V}(\theta)^{-1} - V(\theta)^{-1}|| = o_p(1)$.

Proof: We will prove the uniform convergence for each element. For $t, s \in \{1, 2, \cdots, d\}$, consider the t–s element $\hat{\varsigma}_{ts}(\theta)$ of $\hat{V}(\theta)$; suppressing the argument θ,

$$n\hat{\varsigma}_{ts} = \sum_i \tilde{u}_{it}\tilde{u}_{is} + \sum_i \tilde{u}_{it}(\tilde{\mu}_{is}-\hat{\tilde{\mu}}_{is}) + \sum_i \tilde{u}_{is}(\tilde{\mu}_{it}-\hat{\tilde{\mu}}_{it}) + \sum_i(\tilde{\mu}_{it}-\hat{\tilde{\mu}}_{it})(\tilde{\mu}_{is}-\hat{\tilde{\mu}}_{is}), \tag{26}$$

where $\hat{\tilde{\mu}}_{it} = \sum_j w_{ij}\tilde{m}_{jt}$ and $\tilde{u}_{it} = \tilde{m}_{it} - \tilde{\mu}_{it}$. Applying the Schwarz inequality and lemmas C1 and C4 shows that the RHS2–RHS4 are $o_p(n)$ uniformly in θ. Since $\big|\sum_i \tilde{u}_{it}\tilde{u}_{is} - nE\big(\tilde{u}_{it}\tilde{u}_{is}\big)\big| = o_p(n)$ uniformly in θ by lemma C1, equation (26) shows that $\big|\hat{\varsigma}_{ts} - E\big(\tilde{u}_{it}\tilde{u}_{is}\big)\big| = o_p(1)$ uniformly in θ and hence $\sup_\theta ||\hat{V}(\theta) - V(\theta)|| = o_p(1)$. Since $P^{-1} - Q^{-1} = P^{-1}(Q-P)Q^{-1}$ for matrices P and Q, we have an inequality $||P^{-1} - Q^{-1}|| \leq \big(||P^{-1} - Q^{-1}|| + ||Q^{-1}||\big)||Q^{-1}||||Q - P||$. Using this inequality and assumption D shows that

$$\sup_\theta ||\hat{V}(\theta)^{-1} - V(\theta)^{-1}|| \leq \big(\sup_\theta ||\hat{V}(\theta)^{-1} - V(\theta)^{-1}|| + O(1)\big)O(1)o_p(1). \qquad \square$$

Proof of Theorem 3: Let

$$N_s^{t_1t_2}(\theta) = \sum_{ij} h_{ij}m_{it_1}(\theta)m_{jt_2}(\theta) - \sum_i h_{ii}, \quad \tilde{N}_s^{t_1t_2}(\theta) = \sum_{ij} h_{ij}\tilde{m}_{it_1}(\theta)\tilde{m}_{jt_2}(\theta) - \sum_i h_{ii}$$

and let $\hat{N}_s(\theta)$ be similarly defined; note that $\hat{T}_s(\theta) = \sum_{t=1}^d \hat{N}_s^{tt}(\theta)/D_s$. Since $s = 1$ and $s = 2$ are similar, we only consider $s = 1$ (i.e. $h_{ij} = a_{ij}$) here. Note first that

$$\tilde{N}_1^{t_1t_2}(\theta) = \sum_{ij} a_{ij}\tilde{m}_{it_1}(\theta)m_{it_2}(\theta) - \sum_i a_{ii} = \sum_i \hat{\tilde{\mu}}_{it_1}(\theta)\hat{\tilde{\mu}}_{it_2}(\theta) - \sum_{ij} w_{ij}^2$$
$$\overset{\text{Schwarz,C1,C4}}{=} \sum_i \tilde{\mu}_{it_1}(\theta)\tilde{\mu}_{it_2}(\theta) + o_p(n) - \sum_{ij} w_{ij}^2 \overset{\text{C1}}{=} n\lambda^2 E\big(\tilde{\mu}^*_{it_1}(\theta)\tilde{\mu}^*_{it_2}(\theta)\big) + o_p(n), \quad (27)$$

uniformly in θ. By the same reasoning, we also have

$$N_1^{t_1t_2}(\theta) = n\lambda^2 E\big(\mu^*_{it_1}(\theta)\mu^*_{it_2}(\theta)\big) + o_p(n) \quad (28)$$

uniformly in θ. Now, letting $\hat{\varsigma}^{t_1t_2}(\theta)$ and $\varsigma^{t_1t_2}(\theta)$ be the t_1–t_2 elements of $\hat{V}^{-1/2}(\theta)$ and $V^{-1/2}(\theta)$, respectively and suppressing the argument θ, we note that

$$\sup_\theta |\hat{N}_1^{tt} - N_1^{tt}| \le d^2 \max_{s_1,s_2=1,\cdots,d} \sup_\theta \big|\hat{\varsigma}^{ts_1}\hat{\varsigma}^{ts_2} - \varsigma^{ts_1}\varsigma^{ts_2}\big|\Big(\max_{s_1,s_2=1,\cdots,d} \sup_\theta \big|\tilde{N}_1^{s_1s_2}\big| + \sum_{ij} w_{ij}^2\Big)$$
$$\overset{\text{C7, (27)}}{=} o_p(1)(O_p(n) + O_p(n/k)) = o_p(n). \quad (29)$$

It then follows that

$$\hat{T}_1 \overset{(28),\,(29)}{=} \inf_\theta \big(n\lambda^2 \sum_{t=1}^d E\big(\mu^{*2}_{it}(\theta)\big) + o_p(n)\big)/D_1 \overset{\text{C6}}{\ge} k\lambda^2 \sum_{t=1}^d CE\big(\inf_\theta \mu^{*2}_{it}(\theta)\big) + o_p(k) \to \infty$$

at the rate of k, because λ is fixed and $E\big(\inf_\theta \tilde{\mu}^{*^2}_{it}(\theta)\big) > 0$ under H_1. In view of lemma C6, we also note that $\hat{T}_2$ diverges at the rate of $\sqrt{nk}$ when λ is fixed. □

D Proof of Theorem 4

Lemma D1 $\sup_\theta \max_i |\hat{\sigma}_i^2(\theta) - \tilde{\sigma}_i^2(\theta)| = o_p(1)$.

Proof: Choose an arbitrary $\eta > 0$ and let $\delta > 0$ satisfy (8). Since Θ is compact, it can be divided up into δ-balls $\Theta_1, \cdots, \Theta_{K_\delta}$ for $K_\delta < \infty$. Let θ_κ be the center of Θ_κ. Then

$$\sup_\theta \max_i |\hat{\sigma}_i^2(\theta) - \tilde{\sigma}_i^2(\theta)| \le \eta + \max_{\kappa=1,2,\cdots,K_\delta} \max_i |\hat{\sigma}_i^2(\theta_\kappa) - \tilde{\sigma}_i^2(\theta_\kappa)|, \quad (30)$$

with probability greater than $1 - \eta$. By lemma B1 RHS2 in (30) is $o_p(1)$. Now let $\eta \downarrow 0$ to make RHS1 disappear, also. □

Lemma D2 $\sup_\theta \big(\min_i \hat{\sigma}_i^2(\theta)\big)^{-1} = O_p(1)$.

Proof: Note that $0 < C \le \min_i \tilde{\sigma}_i^2(\theta) \le \sup_\theta \max_i |\tilde{\sigma}_i^2(\theta) - \hat{\sigma}_i^2(\theta)| + \min_i \hat{\sigma}_i^2(\theta)$. Therefore, the result follows from lemma D1. □

Lemma D3 $\sup_\theta \big|\sum_{ij} w_{ij}\tilde{m}_i(\theta)\tilde{m}_j(\theta)\big(\hat{\sigma}_i^{-1}(\theta)\hat{\sigma}_j^{-1}(\theta) - \tilde{\sigma}_i^{-1}(\theta)\tilde{\sigma}_j^{-1}(\theta)\big)\big| = o_p(n)$.

Proof: Omitting the θ–argument, note that by repeated use of the Schwarz inequality,

$$\sup_\theta\Big|\sum_{ij} w_{ij}\tilde m_i\tilde m_j\Big(\frac{1}{\hat\sigma_i\hat\sigma_j}-\frac{1}{\tilde\sigma_i\tilde\sigma_j}\Big)\Big| \le \max_{i,j}\sup_\theta\Big|\frac{1}{\hat\sigma_i\hat\sigma_j}-\frac{1}{\tilde\sigma_i\tilde\sigma_j}\Big|\sqrt{\sup_\theta\sum_{ij} w_{ij}\tilde m_j^2}\sqrt{\sup_\theta\sum_i \tilde m_i^2}. \tag{31}$$

The first RHS factor in (31) is $o_p(1)$ by lemmas D1 and D2. The remaining two RHS factors are $O_p(\sqrt{n})$ by lemmas C2 and C1, respectively, using the assumption that $\tilde m_i^2\in\mathcal{F}$. □

Lemma D4 $\sup_\theta\big|\sum_{ij} w_{ij}\tilde m_i(\theta)\tilde m_j(\theta)\big(\tilde\sigma_i^{-1}(\theta)\tilde\sigma_j^{-1}(\theta)-\sigma_i^{-1}(\theta)\sigma_j^{-1}(\theta)\big)\big| = o_p(n)$.

Proof: Noting that $\tilde\sigma_i,\sigma_i$ are uniformly bounded and uniformly bounded away from zero and that $|\tilde\sigma_i-\sigma_i|^2\le|\tilde\sigma_i^2-\sigma_i^2|$, it follows that (omitting the θ–argument)

$$\sup_\theta\Big|\sum_{ij} w_{ij}\tilde m_i\tilde m_j\Big(\frac{1}{\tilde\sigma_i\tilde\sigma_j}-\frac{1}{\sigma_i\sigma_j}\Big)\Big| \le C\sup_\theta\sum_{ij} w_{ij}|\tilde m_i\tilde m_j|\big(\sqrt{|\tilde\sigma_i^2-\sigma_i^2|}+\sqrt{|\tilde\sigma_j^2-\sigma_j^2|}\big). \tag{32}$$

Now, by repeated application of the Schwarz inequality,

$$\sup_\theta\sum_{ij} w_{ij}|\tilde m_i\tilde m_j|\sqrt{|\tilde\sigma_i^2-\sigma_i^2|} \le \sqrt[4]{\sup_\theta\sum_i\tilde m_i^4}\sqrt[4]{\sup_\theta\sum_i|\tilde\sigma_i^2-\sigma_i^2|^2}\sqrt{\sup_\theta\sum_{ij} w_{ij}\tilde m_j^2}. \tag{33}$$

The first and third RHS factors in (33) are $O_p(n^{1/4})$ by lemmas C1 and C2, respectively. The middle RHS factor in (33) is $o_p(\sqrt{n})$ since $\sup_\theta\sum_i|\tilde\sigma_i^2-\sigma_i^2|^2\le\sup_\theta\sum_{it} w_{it}|\sigma_t^2-\sigma_i^2|^2$, such that lemma C2 again applies. The argument for the $\sup_\theta\sum_{ij} w_{ij}|\tilde m_i\tilde m_j|\sqrt{|\tilde\sigma_j^2-\sigma_j^2|}$ portion of (32) is similar to that of (33). □

Proof of Theorem 4: Since $N_2^H(\theta)=\sum_{ij} w_{ij}\tilde m_i(\theta)\tilde m_j(\theta)\frac{1}{\sigma_i(\theta)}\frac{1}{\sigma_j(\theta)} = n\lambda^2E\big(\mu_i^{*2}(\theta)\big)+o_p(n)$ uniformly in θ,

$$\hat T_2^H \overset{D4}{=} \inf_\theta\big(n\lambda^2E\big(\mu_i(\theta)^2\big)+o_p(n)\big)/D_2 \ge \sqrt{nk}\lambda^2CE\big(\inf_\theta\tilde\mu_i^{*2}(\theta)\big)+o_p(\sqrt{nk})\to\infty. \qquad \square$$

E Proof of Theorem 5

Let $u_i(\theta)=m_i(\theta)-\mu_i(\theta)=m_i(\theta)-\mu_i^o(\theta)-\delta_n q_i$. In this subsection, we use the following expansion:

$$N_s(\theta)=N_s(\theta_0)+N_{\theta s}(\theta_0)(\theta-\theta_0)+\frac{1}{2}(\theta-\theta_0)'N_{\theta\theta s}(\theta_0)(\theta-\theta_0)+O_p(n||\theta-\theta_0||^3), \tag{34}$$

where

$$N_s(\theta_0)=\sum_{t=1}^d\sum_{ij} h_{ij}u_{it}(\theta_0)u_{jt}(\theta_0)-d\sum_i h_{ii}+\delta_n^2\sum_{t=1}^d\sum_{ij} h_{ij}q_{it}q_{jt}+O_p(\sqrt{n}\delta_n) \tag{35}$$

$$N_{\theta s}(\theta_0)=\delta_n\sum_{t=1}^d\sum_{ij} h_{ij}(m_{\theta it}(\theta_0)'q_{jt}+q_{it}m_{\theta jt}(\theta_0)')+O_p(\sqrt{n}) \tag{36}$$

$$N_{\theta\theta s}(\theta_0)=\sum_{t=1}^d\sum_{ij} h_{ij}(m_{\theta it}(\theta_0)m_{\theta jt}(\theta_0)'+m_{\theta jt}(\theta_0)m_{\theta it}(\theta_0)')+O_p(n\delta_n). \tag{37}$$

We start from showing that θ_n converges to θ_0 under the hypothesis (9).

Lemma E1 *Under (9), we have* $\theta_n - \theta_0 = -\delta_n E(\Gamma_i'\Gamma_i)^{-1}E(\Gamma_i' q_i) + o(\delta_n)$.

Proof: First, we show that $\theta_n \to \theta_0$. For this, it suffices to show that $E(\mu_i(\theta)'\mu_i(\theta))$ uniformly converges to $E(\mu_i^o(\theta)'\mu_i^o(\theta))$ under (9). Note that

$$\sup_\theta |E(\mu_i(\theta)'\mu_i(\theta) - \mu_i^o(\theta)'\mu_i^o(\theta))| \le 2\delta_n \sup_\theta E(||q_i||||\mu_i^o(\theta)||) + \delta_n^2 E(q_i'q_i)$$
$$\le 2\delta_n E(q_i'q_i)^{1/2} E\big(\sup_\theta ||\mu_i^o(\theta)||^2\big)^{1/2} + \delta_n^2 E(q_i'q_i) \to 0.$$

Since we have established the convergence of θ_n to θ_0, it now suffices to consider neighborhood of θ_0. Since θ_0 is in the interior of Θ, θ_n will satisfy the first order condition for sufficiently large n. Letting $\Gamma_i(\theta) = \frac{\mu_i(\theta)}{\theta'} = \frac{\mu_i^o(\theta)}{\theta'}$,

$$0 = E(\Gamma_i(\theta_n)'\mu_i(\theta_n)) = E\big(\Gamma_i(\theta_n)'\Gamma_i\big)(\theta_n - \theta_0) + \delta_n E(\Gamma_i(\theta_n)'q_i) + o(||\theta_n - \theta_0||),$$

which implies that $\theta_n - \theta_0 = -\delta_n\big(E(\Gamma_i'\Gamma_i) + o(1)\big)^{-1}\big(E(\Gamma_i'q_i) + o(1)\big) + o(||\theta_n - \theta_0||)$. □

Lemma E2 $N_{\theta s}(\theta_0) = 2n\delta_n E(q_i'\Gamma_i) + o_p(n\delta_n)$ *under the hypothesis (9).*

Proof: Since $s = 1$ and $s = 2$ are similar, we only consider $s = 1$. Suppressing θ_0, note first that $\sum_{ij} a_{ij}(m_{\theta it}q_{jt} + m_{\theta jt}q_{it}) = 2\sum_{ijr} w_{ri}w_{rj}m_{\theta it}q_{jt} = 2nE\big(m_{\theta rt}q_{rt}\big) + o_p(n)$, because

$$|\sum_{ijr} w_{ri}w_{rj}m_{\theta it}q_{jt} - \sum_r m_{\theta rt}q_{rt}|$$
$$\le \sum_{ijr} w_{ri}w_{rj}|m_{\theta it} - m_{\theta rt}||q_{jt} - q_{rt}| + \sum_{ir} w_{ri}|m_{\theta it} - m_{\theta rt}||q_{rt}| + \sum_{jr} w_{rj}|q_{jt} - q_{rt}||m_{\theta rt}|$$
$$\overset{\text{Schwarz, Stone}}{=} \sqrt{o_p(n)o_p(n)} + o_p(n) + o_p(n) = o_p(n).$$

It then follows that $N_{\theta 1}(\theta_0) = 2n\delta_n E(\sum_{t=1}^d \Gamma_{rt}'q_{rt}) + o_p(n\delta_n) = 2n\delta_n E(q_r'\Gamma_r) + o_p(n\delta_n)$. □

Lemma E3 $N_{\theta\theta s}(\theta_0) = 2nE(\Gamma_i'\Gamma_i) + o_p(n)$ *under the hypothesis (9).*

Proof: It is similar to the proof of lemma E2, and it will be omitted. □

Lemma E4 *For any* $C > 0$, $\sup_{||\theta-\theta_0||\le C\delta_n} ||\,\mathrm{d}\,\hat{V}(\theta)^{-1} - \mathrm{d}\,V(\theta)^{-1}|| = o_p(\delta_n)$.

Proof: Similarly to the proof of lemma C7, it is easy to show that the derivative of each element of $\hat{V}(\theta)$ uniformly converges to the derivative of each element of $V(\theta)$. Therefore, it easily follows that $\sup_{||\theta-\theta_0||\le C\delta_n} ||\,\mathrm{d}\,\hat{V}(\theta) - \mathrm{d}\,V(\theta)|| = o_p(\delta_n)$. We will show that the same order obtains for the differentials of the inverses. Suppressing the argument θ, note that

$$\mathrm{d}\,\hat{V}^{-1} - \mathrm{d}\,V^{-1} = \hat{V}^{-1}\,\mathrm{d}\,\hat{V}\hat{V}^{-1} - V^{-1}\,\mathrm{d}\,VV^{-1}$$
$$= \hat{V}^{-1}\big(\mathrm{d}\,\hat{V} - \mathrm{d}\,V\big)\hat{V}^{-1} - \big(\hat{V}^{-1}\otimes\hat{V}^{-1} - V^{-1}\otimes V^{-1}\big)\,\mathrm{vec}(\mathrm{d}\,V).$$

Therefore,

$$||\,\mathrm{d}\,\hat{V}^{-1} - \mathrm{d}\,V^{-1}|| \le \big(||\hat{V}^{-1} - V^{-1}|| + ||V^{-1}||\big)^2||\,\mathrm{d}\,\hat{V} - \mathrm{d}\,V|| + ||\,\mathrm{d}\,V||||\hat{V}^{-1}\otimes\hat{V}^{-1} - V^{-1}\otimes V^{-1}||.$$

Then, use the uniform convergence of $\hat{V}^{-1}$ and $\mathrm{d}\,\hat{V}$ together with the fact that $\sup_{||\theta-\theta_0||\le C\delta_n} ||\,\mathrm{d}\,V(\theta)|| = O(\delta_n)$. □

Lemma E5 *For any constant $C > 0$, (i) $\sup_{||\theta-\theta_0||\leq C\delta_n} |\hat{N}_s(\theta) - N_s(\theta)| = o_p(n\delta_n^2)$, (ii) $\sup_{||\theta-\theta_0||\leq C\delta_n} ||\hat{N}_{\theta s}(\theta) - N_{\theta s}(\theta)|| = o_p(n\delta_n)$ and (iii) $\sup_{||\theta-\theta_0||\leq C\delta_n} ||\hat{N}_{\theta\theta s}(\theta) - N_{\theta\theta s}(\theta)|| = o_p(n)$ under the hypothesis (9).*

Proof: Note first that under (9), we have $\tilde{\mu}_i(\theta) = \tilde{\mu}_i^o(\theta) + \delta_n V^{1/2}(\theta) q_i$, where $\tilde{\mu}_i^o(\theta_0) = 0$. For $r_1, r_2 = 1, 2, \cdots, d$, consider

$$\tilde{N}_s^{r_1 r_2}(\theta) = \sum_{ij} h_{ij} \tilde{m}_{ir_1}(\theta) \tilde{m}_{jr_2}(\theta) - \sum_i h_{ii}.$$

Let $\tilde{q}_i(\theta_0) = V(\theta_0)^{1/2} q_i$ and $\tilde{u}_i(\theta_0) = \tilde{m}_i(\theta_0) - \tilde{\mu}_i(\theta_0)$. Suppressing θ_0, we note that

$$\tilde{N}_s^{r_1 r_2} = \underbrace{\sum_{ij} h_{ij} \tilde{u}_{ir_1} \tilde{u}_{jr_2} - \sum_i h_{ii}}_{=O_p(\sqrt{n/k}) = O_p(n\delta_n^2)} + \underbrace{\sum_{ij} h_{ij} \tilde{u}_{ir_1} \tilde{q}_{jr_2} \delta_n}_{=O_p(\sqrt{n}\delta_n) = o_p(n\delta_n^2)} + \underbrace{\sum_{ij} h_{ij} \tilde{u}_{jr_1} \tilde{q}_{ir_2} \delta_n}_{=o_p(n^{3/4}\delta_n/k^{1/4}) = o_p(n\delta_n^2)} + \underbrace{\sum_{ij} h_{ij} \tilde{q}_{ir_1} \tilde{q}_{jr_2} \delta_n^2}_{=O_p(n\delta_n^2)}, \tag{38}$$

$$\tilde{N}_{s\theta}^{r_1 r_2} = \underbrace{\sum_{ij} h_{ij} \tilde{m}_{\theta ir_1} \tilde{u}_{jr_2} + \sum_{ij} h_{ij} \tilde{u}_{ir_1} \tilde{m}_{\theta jr_2}}_{=o_p(n^{3/4}/k^{1/4}) = o_p(n\delta_n)} + \underbrace{\sum_{ij} h_{ij} \tilde{m}_{\theta ir_1} \tilde{q}_{jr_2} \delta_n + \sum_{ij} h_{ij} \tilde{q}_{ir_1} \tilde{m}_{\theta jr_2} \delta_n}_{=O_p(n\delta_n)}, \tag{39}$$

using lemma B4 of Jun and Pinkse (2007), and

$$\begin{aligned} \sup_\theta ||\tilde{N}_{s\theta\theta}^{r_1 r_2}(\theta)|| \leq \sup_\theta || \sum_{ij} h_{ij} \tilde{m}_{\theta\theta ir_1}(\theta) \tilde{m}_{jr_2}(\theta)|| + \sup_\theta || \sum_{ij} h_{ij} \tilde{m}_{ir_1}(\theta) \tilde{m}_{\theta\theta jr_2}(\theta)|| \\ + \sup_\theta || \sum_{ij} h_{ij} (\tilde{m}_{\theta ir_1}(\theta) m_{\theta jr_2}(\theta)' + \tilde{m}_{\theta jr_2}(\theta) \tilde{m}_{\theta ir_1}(\theta)')|| \overset{\text{C1,C4}}{=} O_p(n). \end{aligned} \tag{40}$$

Therefore,

$$\sup_{||\theta-\theta_0||\leq C\delta_n} |\tilde{N}_s^{r_1 r_2}(\theta)| \leq |\tilde{N}_s^{r_1 r_2}(\theta_0)| + C||\tilde{N}_{s\theta}^{r_1 r_2}(\theta_0)||\delta_n + \sup_\theta ||\tilde{N}_{s\theta\theta}^{r_1 r_2}(\theta)||\delta_n^2 C^2/2 = O_p(n\delta_n^2), \tag{41}$$

$$\sup_{||\theta-\theta_0||\leq C\delta_n} ||\tilde{N}_{\theta s}^{r_1 r_2}(\theta)|| \leq ||\tilde{N}_{\theta s}^{r_1 r_2}(\theta_0)|| + C \sup_\theta ||\tilde{N}_{s\theta\theta}^{r_1 r_2}(\theta)||\delta_n = O_p(n\delta_n). \tag{42}$$

Now, let $\hat{\varsigma}^{t_1 t_2}(\theta)$ and $\varsigma^{t_1 t_2}(\theta)$ be the t_1–t_2 element of $\hat{V}^{-1/2}(\theta)$ and $V^{-1/2}(\theta)$, respectively. Then,

$$\begin{aligned} &\sup_{||\theta-\theta_0||\leq C\delta_n} |\hat{N}_s(\theta) - N_s(\theta)| \\ &\leq d^3 \max_{t_1,t_2,t_3} \sup_{||\theta-\theta_0||\leq C\delta_n} |\hat{\varsigma}^{t_1 t_2}(\theta)\hat{\varsigma}^{t_1 t_3}(\theta) - \varsigma^{t_1 t_2}(\theta)\varsigma^{t_1 t_3}(\theta)| \max_{t_2,t_3=1} \Big(\sup_{||\theta-\theta_0||\leq C\delta_n} |\tilde{N}_s^{t_2 t_3}(\theta)| + \sum_i h_{ii} \Big) \\ &\overset{\text{A6,C7,E4,(41)}}{=} \big(o_p(k/n) + o_p(\delta_n)\big)\big(O_p(n\delta_n^2) + O_p(n/k)\big) = o_p(n\delta_n^2). \end{aligned}$$

The other cases of (ii) and (iii) similarly follow from equations (40), (41), (42), and uniform convergence of $\hat{V}(\theta)$ and their derivatives. □

Lemma E6 *Let $\hat{\theta}_s$ be the minimizer of $\hat{T}_s(\theta)$. Then, $\hat{\theta}_s - \theta_n = o_p(\delta_n)$ under the hypothesis (9). In particular, $\hat{\theta}_s - \theta_0 = -\delta_n E(\Gamma_i'\Gamma_i)^{-1}E(\Gamma_i' q_i) + o_p(\delta_n)$.*

Proof: Since $\hat{N}_s(\theta) = N_s(\theta) + o_p(n)$ and $N_s(\theta) = nE\big(||\mu_i(\theta)||^2\big) + o_p(n)$ uniformly in θ, it is clear that $\hat{\theta}_s - \theta_n = o_p(1)$. Therefore, we only focus on the rate of $\hat{\theta}_s - \theta_n$. Since θ_0 is in the interior of Θ, the first order condition $\hat{N}_{\theta s}(\hat{\theta}_s) = 0$ is available for sufficiently large n. Expanding the first order condition yields

$$\hat{N}_{\theta s}(\theta_n) + \hat{N}_{\theta\theta s}(\bar{\theta}_s)(\hat{\theta}_s - \theta_n) = 0,$$

where $\bar{\theta}_s$ is between $\hat{\theta}_s$ and θ_n. Here, it can be easily shown that $\hat{N}_{\theta\theta s}(\theta) = N_{\theta\theta s}(\theta) + o_p(n)$ uniformly in θ. Also, following the proof of lemma C4, we note that $N_{\theta\theta s}(\theta) = 2n\sum_{t=1}^d E\big(\mu_{\theta\theta it}(\theta)\mu_{it}(\theta) + \mu_{\theta it}(\theta)'\mu_{\theta it}(\theta)\big) + o_p(n)$. Therefore, $\hat{N}_{\theta\theta s}(\bar{\theta}_s) = 2n\sum_{t=1}^d E\big(\mu_{\theta it}(\theta_n)'\mu_{\theta it}(\theta_n)\big) + o_p(n)$, and hence it suffices to show that $\hat{N}_{\theta s}(\theta_n) = o_p(n\delta_n)$ for the result. Suppressing θ_n, note that

$$\hat{N}_{\theta s} \overset{\text{E1,E5}}{=} N_{\theta s} + o_p(n\delta_n) = \sum_{t=1}^d \sum_{ij} h_{ij} m_{\theta it} m_{jt} + \sum_{t=1}^d \sum_{ij} h_{ij} m_{\theta jt} m_{it} + o_p(n\delta_n).$$

Since the RHS1 and the RHS2 are similar, we will only consider the RHS1. Letting $u_j = m_j - \mu_j$, write

$$\sum_{t=1}^d \sum_{ij} h_{ij} m_{\theta it} m_{jt} = \sum_{t=1}^d \sum_{ij} h_{ij} m_{\theta it} u_{jt} + \sum_{t=1}^d \sum_{ij} h_{ij} m_{\theta it}(\mu_{jt} - \mu_{it}) + \sum_{t=1}^d \sum_{ij} h_{ij} m_{\theta it}\mu_{it}. \tag{43}$$

The RHS3 of (43) can be easily shown to be $o_p(n\delta_n)$, because $\sum_{t=1}^d m_{\theta it}(\theta_n)\mu_{it}(\theta_n)$ is an independent mean zero array. For the RHS2 of (43), expand $\mu_{it}(\theta_n) - \mu_{jt}(\theta_n)$ around θ_0, and we obtain

$$\begin{aligned}|\sum_{ij} h_{ij} m_{\theta it}\big(\mu_{jt} - \mu_{it}\big)| \le \sum_{ij} h_{ij}|m_{\theta it}||q_{jt} - q_{it}|\delta_n + \sum_{ij} h_{ij}|m_{\theta it}|||\mu_{\theta jt}(\theta_0) - \mu_{\theta it}(\theta_0)||||\theta_n - \theta_0|| \\ + \sum_{ij} h_{ij}|m_{\theta it}|\big(\sup_\theta ||\mu_{\theta\theta jt}(\theta)|| + \sup_\theta ||\mu_{\theta\theta it}(\theta)||\big)||\theta_n - \theta_0||^2,\end{aligned}$$

where the last term is $O_p(n\delta_n^2) = o_p(n\delta_n)$ by lemma E1. The first two terms are similar and we only consider the first one. Also, when $h_{ij} = w_{ij}$, it is clearly $o_p(n)$ by Stone's lemma and we only consider $h_{ij} = a_{ij}$. In this case,

$$\sum_{ij} a_{ij}|m_{\theta it}||q_{jt} - q_{it}| \le \sum_{jr} w_{rj}\big(\sum_i w_{ri}|m_{\theta it}|\big)|q_{jt} - q_{rt}| + \sum_{ir} w_{ri}|m_{\theta it}||q_{rt} - q_{it}|,$$

which is $o_p(n)$ by lemma B1 of Jun and Pinkse (2007). Lastly, consider the RHS1 of (43). Note that $E\Big(\big(\sum_{ij} h_{ij} m_{\theta it} u_{jt}\big)^2\Big) = E\Big(\sum_{ijr} h_{ij}h_{rj} m_{\theta it} m_{\theta rt} u_{jt}^2\Big)$. When $h_{ij} = w_{ij}$, it follows from lemma B4 of Jun and Pinkse (2007) that $E\Big(\sum_{ijr} w_{ij}w_{rj} m_{\theta it} m_{\theta rt} u_{jt}^2\Big) = o(n^{3/2}k^{-1/2})$, which implies that the RHS1 of (43) is $o_p(n^{3/4}k^{-1/4}) = o_p(n\delta_n)$. When $h_{ij} = a_{ij}$, note that

$$E\big(\sum_{ijr} a_{ij}a_{rj} m_{\theta it} m_{\theta rt} u_{jt}^2\big) = \sum_{jr_2r_3} E\Big(w_{r_2j}w_{r_3j} u_{it}^2\big(\sum_i w_{r_2i} m_{\theta it}\big)\big(\sum_{r_1} w_{r_3r_1} m_{\theta r_1 t}\big)\Big) = o(n^{3/2}k^{-1/2}),$$

where the last equality is due to lemma B4 of Jun and Pinkse (2007). Applying lemma E1 completes the proof. □

Lemma E7 *Suppressing θ_0,*

$$\hat{N}_s(\hat{\theta}_s) = \sum_{ij} h_{ij} u_i' u_j - d \sum_i h_{ii} + n\delta_n^2 \big(E(q_i' q_i) - E(q_i' \Gamma_i) E(\Gamma_i' \Gamma_i)^{-1} E(\Gamma_i' q_i) \big) + o_p(n\delta_n^2)$$

under the hypothesis (9).

Proof: From equation (34), lemmas E2, E3, and E5, it follows that

$$\hat{N}_s(\hat{\theta}_s) = N_s(\hat{\theta}_s) + o_p(n\delta_n^2) = N_s(\theta_0) - n\delta_n^2 E(q_i' \Gamma_i) E(\Gamma_i' \Gamma_i)^{-1} E(\Gamma_i' q_i) + o_p(n\delta_n^2).$$

Therefore, in view of equation (35), we only need to show that $\delta_n^2 \sum_{ij} h_{ij} q_i' q_j = n\delta_n^2 E(q_i' q_i) + o_p(n\delta_n^2)$. Since the case of $s = 1$ (i.e. $h_{ij} = a_{ij}$) is similar, we only consider $s = 2$ (i.e. $h_{ij} = w_{ij}$); see also the proof of lemma E2. Noting that $|\sum_{ij} w_{ij} q_{it} q_{jt} - \sum_i q_{it}^2| \leq \sum_{ij} w_{ij} |q_{it}||q_{jt} - q_{it}| = o_p(n)$, we know that

$$\delta_n^2 \sum_{t=1}^d \sum_{ij} w_{ij} q_{it} q_{jt} = \delta_n^2 \sum_{t=1}^d \sum_i q_{it}^2 + o_p(n\delta_n^2) = n\delta_n^2 E(\sum_{t=1}^d q_{it}^2) + o_p(n\delta_n^2). \quad \square$$

Proof of Theorem 5: Let $D_1^2 = 2d \sum_{j \neq i} a_{ij}^2$ and $D_2^2 = d \sum_{ij} w_{ij}(w_{ij} + w_{ji})$. Then, they are bounded away from 0, and they are $O_p(n/k)$; see lemmas A1 and C6. Since $\sqrt{nk}\delta_n^2 = O(1)$ by the setup, lemma E7 shows that

$$\hat{T}_s - \frac{\sqrt{nk}\delta_n^2 \Big(E(q_i' q_i) - E(q_i' \Gamma_i) E(\Gamma_i' \Gamma_i)^{-1} E(\Gamma_i' q_i) \Big)}{\sqrt{k/n} D_s} = \frac{\sum_{t=1}^d \sum_{ij} h_{ij} u_{it} u_{jt} - d \sum_i h_{ii}}{D_s} + o_p(1),$$

where we suppressed θ_0. Therefore, we only need to show that $\big(\sum_{t=1}^d \sum_{ij} h_{ij} u_{it} u_{jt} - d \sum_i h_{ii}\big)/D_s$ has a normal distribution under (9). But, it follows from the same proofs of section A. $\square$

F Proof of Theorem 6

Lemma F1 $P(S_n \leq k/n,\ \tilde{S}_n > k/n) = o(k/n)$.

Proof: Follows from lemmas C7 and C11 of Jun and Pinkse (2008). $\square$

Lemma F2 $P(z_1 \in \mathcal{N}_2, z_2 \in \mathcal{N}_1, z_3 \in \mathcal{N}_4, z_4 \in \mathcal{N}_3) = P(z_1 \in \mathcal{N}_2, z_2 \in \mathcal{N}_1) P(z_3 \in \mathcal{N}_4, z_4 \in \mathcal{N}_3) + o(k^2/n^2)$.

Proof: Follows from lemma C12 of Jun and Pinkse (2008). $\square$

Proof of Theorem 6: It suffices to prove that $(k/n) D_2^2 = (k/n) \sum_{ij} w_{ij}(w_{ij} + w_{ji}) = 1 + (k/n) \sum_{ij} w_{ij} w_{ji} \xrightarrow{p} 2$. Let z and $\tilde{z}$ be two independent copies of z_i. Let $S_n = S_n(z, \tilde{z}) = n^{-1} \sum_{i=1}^n I\big(||z_i - z|| \leq ||\tilde{z} - z||\big)$ and $\tilde{S}_n = S_n(\tilde{z}, z)$. Note that by lemma F1,

$$\begin{aligned} E(w_{12} w_{21}) &= k^{-2} P(S_n \leq k/n, \tilde{S}_n \leq k/n) \\ &= k^{-2} P(S_n \leq k/n) - k^{-2} P(S_n \leq k/n, \tilde{S}_n > k/n) = 1/nk - o(1/nk). \end{aligned} \quad (44)$$

It thus suffices to show that $\mathrm{Var}\big(\sum_{ij} w_{ij}w_{ji}\big) = o(n^2/k^2)$. Finally, since by lemma F2, $\mathrm{Cov}(w_{12}w_{21}, w_{34}w_{43}) = o(1/n^2k^2)$, it follows that

$$\begin{aligned}\mathrm{Var}\Big(\sum_{ij} w_{ij}w_{ji}\Big) &= \sum_{ijts} \mathrm{Cov}(w_{ij}w_{ji}, w_{ts}w_{st}) \\ &= n^2(n-2)^2\,\mathrm{Cov}(w_{12}w_{21}, w_{34}w_{43}) + 4n^2(n-2)\,\mathrm{Cov}(w_{12}, w_{21}, w_{13}w_{31}) + 2n^2\,\mathrm{Var}(w_{12}w_{21}) \\ &= o(n^2/k^2) + O(n/k^2) + O(n/k) = o(n^2/k^2). \quad \square\end{aligned}$$

G Proof of Theorem 7

Let $\aleph_j = ||z - z_j||^{d_z}$ and let $\aleph_{(j)}$ be the j^{th} (smallest) order statistic. Let $\mathbb{D}_n = \aleph_{(k)}$ and consider a sequence τ_n such that $k/n \prec \tau_n^{d_z} \prec h_n^{d_z}$.

Lemma G1 *$P(\mathbb{D}_n > \tau_n^{d_z}) = o(1)$.*

By the mean value theorem, $F(s) \geq cs$ for $c = \inf_s f(s) > 0$. Since $\mathbb{D}_n$ is the k^{th} order statistic, we can write by the Markov inequality,

$$\begin{aligned}\tau^{d_z} P(\mathbb{D}_n > \tau_n^{d_z}) \leq E(\mathbb{D}_n) &= \int \frac{n!}{(k-1)!(n-k)!} F(s)^{k-1}\big(1-F(s)\big)^{n-k} f(s) s ds \\ &\leq c^{-1} \int \frac{n!}{(k-1)!(n-k)!} F(s)^{k}\big(1-F(s)\big)^{n-k} f(s) ds \\ = c^{-1}\frac{n!}{(k-1)!(n-k)!}\int_0^1 y^k(1-y)^{n-k}dy &\overset{\text{Beta}}{=} c^{-1}\frac{n!}{(k-1)!(n-k)!}\frac{k!(n-k)!}{(n+1)!} = c^{-1}\frac{k}{n+1},\end{aligned}$$

which completes the proof. $\square$

Lemma G2 *For any sequence $\{\alpha_{in}\}$ with $E(||\alpha_{in}||) = O(1)$, $\sum_{ij} w_{ij}\|\alpha_{in}(Q_{in} - Q_{jn})\| = o_p(n/h_n^{d_z})$.*

Proof: We have

$$\begin{aligned}\sum_{ij} w_{ij}\|\alpha_{in}(Q_{in} - Q_{jn})\| &\leq \sum_{ij} w_{ij}\|\alpha_{in}(Q_{in} - Q_{jn})\|\big(I(\|z_i - z_j\| > \tau_n) + I(\|z_i - z_j\| \leq \tau_n)\big) \\ &\leq (2C_w M_q/k)\sum_{ij}\|\alpha_{in}\| I\big(\tau_n^{d_z} < \|z_i - z_j\|^{d_z} \leq \mathbb{D}_n\big) + C_q(\tau_n/h_n^{d_z+1})\sum_i \|\alpha_{in}\| \\ &\leq \Big((2C_w n M_q/k) I(\tau_n^{d_z} < \mathbb{D}_n) + C_q(\tau_n/h_n^{d_z+1})\Big)\sum_i \|\alpha_{in}\| = o_p(n/h_n^{d_z}),\end{aligned}$$

by lemma G1. $\square$

Proof of Theorem 7: Because (i) $E(||Q_{in}||) = O(1)$ and by lemma G2, (ii) $\sum_{ij} w_{ij}(Q_{in} - Q_{jn})'Q_{in} = o_p(n/h_n^{d_z})$, and (iii) $\sum_{ij} w_{ij}\|Q_{in} - Q_{jn}\|\,\|m_{\theta i}\| = o_p(n/h_n^{d_z})$, we follow the proof of

lemma E7 to obtain the expansion

$$\begin{aligned}\hat{N}_2(\hat{\theta}_2) &= N_2(\theta_0) - n\tilde{\delta}_n^2 E(Q'_{in}\Gamma_i)E(\Gamma'_i\Gamma_i)^{-1}E(\Gamma'_i Q_{in}) + o_p(n\tilde{\delta}_n^2/h_n^{d_z}) \\ &= \sum_{ij} w_{ij}u'_i(\theta_0)u_j(\theta_0) + \tilde{\delta}_n^2 \sum_{ij} w_{ij}Q'_{i,n}Q_{j,n} - n\tilde{\delta}_n^2 E(Q'_{in}\Gamma_i)E(\Gamma'_i\Gamma_i)^{-1}E(\Gamma'_i Q_{in}) + o_p(n\tilde{\delta}_n^2/h_n^{d_z}) \\ &= \sum_{ij} w_{ij}u'_i(\theta_0)u_j(\theta_0) + \tilde{\delta}_n^2 \sum_{ij} w_{ij}Q'_{i,n}Q_{j,n} + o_p(n\tilde{\delta}_n^2/h_n^{d_z}).\end{aligned}$$

Now, by lemma G2, $\left|\sum_{ij} w_{ij}Q'_{in}Q_{jn} - \sum_i \|Q_{in}\|^2\right| \leq \sum_{ij} w_{ij}\|Q_{in}\|\,\|Q_{in} - Q_{jn}\| = o_p(n/h_n^{d_z})$. Since $\sum_{i=1}^n \|Q_{in}\|^2 \simeq nh_n^{-d_z}f(v)\Upsilon_2$,

$$\hat{T}_2 = \frac{\sum_{ij} w_{ij}u'_i(\theta_0)u_j(\theta_0)}{D_2} + \frac{\sqrt{nk}(\tilde{\delta}_n^2/h_n^{d_z})f(v)\Upsilon_2}{\sqrt{k/n}D_2} + o_p(\sqrt{nk}\tilde{\delta}_n^2/h_n^{d_z}). \qquad \square$$