Testing under Weak Identification with Conditional Moment Restrictions*

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

September 2011

We propose a semiparametric test for the value of coefficients in models with conditional moment restrictions which has correct size regardless of identification strength. The test is in essence an Anderson–Rubin (AR) test using nonparametrically estimated instruments to which we apply a standard error correction. We show that the test is (1) always size-correct, (2) consistent when identification is not too weak, and (3) asymptotically equivalent to an infeasible AR test when identification is sufficiently strong. We moreover prove that under homoskedasticity and strong identification our test has a limiting noncentral chi–square distribution under a sequence of local alternatives, where the noncentrality parameter is given by a quadratic form of the inverse of the semiparametric efficiency bound.

Key Words: Conditional Moment Restrictions, Weak Identification, Optimal Instruments, k–Nearest Neighbors
JEL Classification Codes: C12, C14, C21, C31

*This paper is in part based on research supported by NSF grant SES-0922127. We thank seminar participants at Syracuse University, Columbia University, Arizona State University, the University of Arizona, The Pennsylvania State University, McGill University, and The Festschrift Conference for Tony Lancaster at Brown University, Yuichi Kiramura, Whitney Newey and four anonymous referees for their comments. We thank the Human Capital Foundation (www.hcfoundation.ru), and especially Andrey P. Vavilov, for their support of the Department of Economics, the Center for the Study of Auctions, Procurements, and Competition Policy (CAPCP, http://capcp.psu.edu/), and the Center for Research in International Financial and Energy Security (CRIFES, http://crifes.psu.edu/) at Penn State University. Sung Jae Jun is the Strumpf early career professor of economics. Joris Pinkse is an extramural fellow of the Center for Economic Research at Tilburg University.

1. Introduction

It is now well understood that standard asymptotic approximations can be misleading when instruments are *weak* (e.g. Phillips (1989); Staiger and Stock (1997); Stock, Wright, and Yogo (2002)). There is a substantial amount of work establishing limit distributions with alternative asymptotic variances of estimators in which the number of weak instruments is allowed to increase with the sample size, e.g. Newey (2004); Chao and Swanson (2005); Stock and Yogo (2005); Han and Phillips (2006); Newey and Windmeijer (2009). However, even t–tests based on such alternative asymptotic variances do not allow for correct inferences in the sense of Dufour (1997); see also Newey and Windmeijer (2009). There are testing–based inference procedures that are robust to weak identification but all of them are designed in a parametric setup (e.g. Anderson and Rubin, 1949; Kleibergen, 2002; Moreira, 2003; Kleibergen, 2005). In this paper we propose a semiparametric test for the value of coefficients in models with conditional moment restrictions that always has correct size. We show that the proposed test is consistent and has optimal local power under homoskedasticity when identification is sufficiently strong; i.e. under a sequence of local alternatives the test statistic then has a noncentral chi–square distribution whose noncentrality parameter is equal to a quadratic form of the inverse of the semiparametric efficiency bound.[1]

We begin our discussion with a simple example in which there is a single linear structural equation, a scalar parameter of interest θ_0 and no nuisance parameters, i.e. $y_i = Y_i\theta_0 + u_i$. In this example we have a vector of instruments z_i for which $\mathbb{E}[u_i|z_i] = 0$ and $\mathrm{Var}[u_i|z_i] = \sigma^2$. We use this simple example because it highlights the issues and differences across testing procedures with the least number of caveats and exceptions. The optimal instruments in this model are given by $g_i = g(z_i) = \mathbb{E}[Y_i|z_i]$, which is the derivative of $M_i(\theta) = \mathbb{E}[Y_i\theta - y_i|z_i] = \mathbb{E}[Y_i|z_i](\theta - \theta_0)$. Note that g is an unknown, but nonparametrically estimable, function of z_i.

We let $g_i = \lambda\tilde{g}_i$ with $0 < \mathbb{E}\tilde{g}_i^2 < \infty$, where λ may depend on the sample size n. This construction is a technical but common tool to study and improve the behavior of statistics in finite samples, which was first introduced by Staiger and Stock (1997) who assumed that $g(z) = z^\intercal\pi_0 = z^\intercal\tilde{\pi}_0\lambda$. Staiger and Stock showed that when $\lambda \preceq 1/\sqrt{n}$, the two stage least squares (2SLS) estimator is inconsistent for θ_0 and its limiting distribution is nonnormal, where $a \preceq b$ means that $a/b \to C$ for some constant C; when $C = 0$, we write $a \prec b$.[2] Given this fact, it is not surprising that standard semiparametric estimation theory also breaks down when λ decreases too fast.

Indeed, consider the semiparametric estimator $\hat{\theta}$ based on k nearest neighbor (knn) estimates of g_i, which was also considered by Newey (1990); k is a smoothing parameter chosen by the researcher (see also Stone, 1977; Robinson, 1987; Newey, 1993). Although standard theory says that $\hat{\theta}$ is asymptotically normal with asymptotic variance equal to the semiparametric efficiency bound, this occurs only when identification is sufficiently strong. For instance, if $\lambda \preceq 1/\sqrt[4]{nk}$, the

[1] We also propose a general statistic in the presence of heteroskedasticity of unknown form, and we show that the dominant term of the test statistic under the alternative has an expansion around the truth that is equal to a quadratic form of the inverse of the semiparametric efficiency bound.

[2] We will similarly use $\succ, \succeq$ in this paper.

estimator $\hat{\theta}$ ceases to be consistent for θ_0. In fact, the limiting distribution implied by standard theory requires that $\lambda \succ 1/\sqrt{k}$. When $1/\sqrt[4]{nk} \prec \lambda \preceq 1/\sqrt{k}$, $\hat{\theta}$ is consistent and asymptotically normal but its asymptotic variance exceeds the one implied by standard theory; see example I in section 2.[3] This phenomenon is not specific to the knn method and similar conclusions obtain if other nonparametric estimators of g_i are used.[4]

The focus of this paper is on testing rather than estimation. As noted before, our objective is to obtain correct inferences in the sense of Dufour (1997) while maintaining performance in more favorable circumstances. More specifically, we propose a semiparametric test of the hypothesis $H_0 : \theta_0 = \theta_H$, which is always size–correct, is consistent when $\lambda \succ 1/\sqrt[4]{nk}$, and is equivalent in terms of local power to a t–test based on the efficient estimator $\hat{\theta}$ when $\lambda \succ 1/\sqrt{k}$.

One way of dealing with a conditional moment restrictions model is to make a choice of instruments to obtain a set of unconditional moment conditions. When the conditional moment condition $\mathbb{E}[Y_i\theta_0 - y_i|z_i] = 0$ a.s. is replaced with an unconditional one with a naïve choice of instruments, e.g. $\mathbb{E}\left[z_i(Y_i\theta_0 - y_i)\right] = \mathbb{E}[z_i g_i](\theta_0 - \theta_0) = 0$, two things can go wrong: (i) g_i is so small that the model is intrinsically difficult to estimate and/or (ii) the instruments z_i are chosen poorly in the sense that $\mathbb{E}[z_i g_i]$ is close to zero. In both cases generalized methods of moments (GMM) estimators are known to behave poorly and t–tests based on them suffer from serious size distortions (e.g. Dufour, 1997; Stock and Wright, 2000). Problem (ii) can be addressed by using the optimal instrument g_i instead of z_i, albeit that even unconditional moment conditions using optimal instruments are not generally equivalent to the original conditional moment restrictions (Domínguez and Lobato, 2004, example 2).

There are several methods that address weak identification (caused by either (i) or (ii)) in the context of unconditional moment conditions models such as the Anderson–Rubin (1949, AR) procedure, the Kleibergen (2002, 2005; K) test, and the Moreira (2003, M) test.[5] Andrews, Moreira, and Stock (2004) found in a simulation study that the M–test approximates optimal average power in the above model with linear g and normally distributed errors which are independent of the instruments.

Although these testing procedures are robust to potential identification failure, all of them are fully parametric and designed for the case of a given choice of instruments. Therefore, when poor identification is due to a poor choice of instruments, inference based on these procedures can be improved. In particular, we will show that a robust test providing improved inference can be constructed by using nonparametrically estimated g_i.

---

[3] Note that as k has a rate that is arbitrarily close to n, both $1/\sqrt[4]{nk}$ and $1/\sqrt{k}$ are arbitrarily close to the corresponding parametric cut–off rate $1/\sqrt{n}$.

[4] In the case of series estimation similar conclusions can be deduced from the many–instrument literature; e.g. LIML estimators are known to be consistent and asymptotically normal with a nonstandard asymptotic variance when the number of instruments is allowed to increase slowly so that the concentration parameter increases sufficiently fast. See Bekker (1994); Chao and Swanson (2005); Hausman, Newey, Woutersen, and Chao (2007); Newey and Windmeijer (2009) for related work.

[5] See also Staiger and Stock (1997) for the AR–test and Andrews, Moreira, and Stock (2004, 2006) for the M–test.

The meaning of improved inference is twofold. Since the optimal instrument g_i is used, the problem due to a poor choice of instruments does not arise. Therefore, the proposed test may have nontrivial power even when other tests, not using optimal instruments, have no power. When identification is sufficiently strong, a further benefit of using optimal instruments is local power optimality. Indeed, the local power of the proposed test is equal to that of a t–test based on the efficient estimator $\hat{\theta}$, where the square of the mean of the local power distribution equals the inverse of the semiparametric efficiency bound (e.g. Newey (1993)) times the square of the local drift parameter.

The above discussion can be extended in several directions. When the parameter of interest θ_0 is vector–valued, a semiparametric chi–square statistic can be constructed without much difficulty. In this case, our test becomes comparable with a Wald test based on a semiparametric efficient estimator.[6] Further, nuisance parameters arising from the presence of additional regressors (endogenous or exogenous) whose coefficients are not of interest, can be partialled out without changing our results, provided that such nuisance parameters are identified if θ_0 is known (see also Kleibergen, 2002, 2005).

When the parameter of interest is scalar–valued (with or without nuisance parameters), the asymptotic null distribution of our test statistic is standard normal. Consequently, testing one–sided alternatives is straightforward. Andrews, Moreira, and Stock (2004) pointed out that no one–sided version of the K–test is consistent and they proposed a one–sided modification of the M–test. They also showed that the power function of the one–sided version of the M–test is close to the Gaussian power envelope when errors are normal and g_i is indeed linear in z_i. However, the M–test has no such property if errors are nonnormal or g_i is nonlinear. Indeed, our simulation experiments show that the one–sided version of our test can have better power than the corresponding M–test when g_i is nonlinear. A further advantage of an asymptotic normal or chi–square distribution is that the usual critical values can be used, whereas the critical values for the M–test must be simulated.

The presence of heteroskedasticity of unknown form does not affect the asymptotic validity (i.e. correct size) or consistency of (the simple form of) our test but it causes it to lose local power optimality because g_i is then no longer the optimal instrument. This feature is comparable to the 2SLS estimator in a parametric setup, which is efficient only under homoskedasticity but (provided that identification is sufficiently strong) is consistent under heteroskedasticity, also; the same applies to the M test.[7] We hence generalize our results to the case of multiple equations with heteroskedasticity of unknown form in section 3.

As mentioned before, the number k is an input parameter. Having a sample–size–dependent input parameter is common for nonparametric procedures. Although we state rate conditions on k, these are of limited practical use in a single sample of finite size. In estimation problems, an

[6] In this paper we will refer to both the t–test and the Wald test based on an efficient estimator as the N–test.

[7] Andrews, Moreira, and Stock (2004) proposed a modified M–test that is designed to be robust to heteroskedasticity. However, our simulation experiments show that the heteroskedasticity–robust version of the M–test tends to have size distortions in small samples when identification is weak and many instruments are used.

'optimal' choice of k can be motivated by minimizing a criterion like the (integrated) mean square error. With hypothesis testing, it is unclear what optimality criterion one would use. Indeed, the choice of k that minimizes estimation error of the nonparametric nuisance function is unlikely to equal the value optimizing test performance, however defined. Fortunately, we have found the performance of the test to be relatively insensitive to the choice of k in extensive Monte Carlo experiments.[8]

Finally, knn is not the only nonparametric method available, but it is attractive in this context because knn estimators have smaller variances in the sparse density area of z_i than e.g. kernel estimators, thereby obviating the need for an unnatural trimming procedure and (in settings like ours) knn estimators require minimal conditions.

The paper is organized as follows. In section 2 we propose a simple statistic for the single equation case that does not require estimation of conditional variances. The size, power, and local power properties of the test are carefully discussed. Section 3 extends the discussion to a more general case with multiple equations, where we use optimal instruments under heteroskedasticity. In section 4 we report the results of our Monte Carlo experiments. Section 5 concludes.

2. Inference with a Single Equation and Weak Identification

2.1. Setup and Motivation.

We consider the following model:

$$\mathbb{E}\left[m(\omega_i, \theta_0)|z_i\right] = \mathbb{E}\left[m_i(\theta_0)|z_i\right] = 0 \quad a.s., \tag{1}$$

where $m : \mathcal{X} \times \Theta \to \mathbb{R}$ is twice continuously differentiable in its second argument with $\mathcal{X}$ the support of ω_i and $\Theta \subset \mathbb{R}^{d_\theta}$ the parameter space; so θ_0 is a d_θ–dimensional parameter vector of interest.

Identification strength will be measured by λ, which is a nonincreasing (with the sample size n) sequence of numbers such that for any $\theta \in \Theta \in \mathbb{R}^{d_\theta}$,

$$M(z_i, \theta) = M_i(\theta) = \mathbb{E}\left[m_i(\theta)|z_i\right] = \lambda \tilde{M}(z_i, \theta) = \lambda \tilde{M}_i(\theta) \text{ a.s.}, \tag{2}$$

where $\tilde{M} : \mathcal{Z} \times \Theta \to \mathbb{R}$ is a function that does not depend on n, where $\mathcal{Z}$ is the support of z_i. By letting $\mathbb{E}\left[m_i(\theta)|z_i\right]$ vary with λ (and hence with n), we follow the setup of e.g. Staiger and Stock (1997), albeit that Staiger and Stock's focus was on the case $\lambda \sim n^{-1/2}$. Note that the above construction is artificial in that it is not intended to describe the true data generating process; moment conditions do not generally vary with the sample size in real applications. The sequence λ is introduced as a tool to study and improve on the properties of statistical procedures in finite samples by analyzing the asymptotic effect of allowing identification strength to deteriorate as the sample size increases.

The following example shows that semiparametric estimators are not consistent when λ decreases too fast, i.e. identification is too weak. An implication is that standard hypothesis tests

[8] Not all experiments are reported here.

based on such estimators are invalid. Although this is a simple example, we note that the problem applies generally.

Example I. *Recall the simple model from the introduction, where* $m_i(\theta) = y_i - Y_i\theta$, $M_i(\theta) = \mathbb{E}(Y_i|z_i)(\theta - \theta_0)$, *and* $g_i = \mathbb{E}(Y_i|z_i) = \lambda\tilde{g}_i$.[9] *Let* $\hat{\theta}_I = \sum_i g_i y_i / \sum_i g_i Y_i$ *and* $\hat{\theta} = \sum_i \hat{g}_i y_i / \sum_i \hat{g}_i Y_i$, *where* $\hat{g}_i$ *is a knn estimator with* $k \prec n$.[10] *So,* $\hat{\theta}_I$ *and* $\hat{\theta}$ *are the infeasible and feasible semiparametric estimators, respectively, which are known to be efficient under homoskedasticity and strong identification (see Newey (1990, 1993)). However, as* λ *is allowed to vary with* n, *we have*

$$\begin{cases} \hat{\theta}_I - \theta_0 & \xrightarrow{d} \quad \Psi_{\tilde{g}u}/(\tilde{C}\mathbb{E}\tilde{g}_i^2 + \Psi_{\tilde{g}v}), & \text{if } 0 < \sqrt{n}\lambda \to \tilde{C} < \infty, \\ \sqrt{n}\lambda(\hat{\theta}_I - \theta_0) & \xrightarrow{d} \quad \Psi_{\tilde{g}u}/\mathbb{E}\tilde{g}_i^2, & \text{if } \sqrt{n}\lambda \to \infty, \end{cases} \tag{3}$$

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

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

An informal justification for the above results is in appendix G.[11] □

The behavior of the infeasible estimator as a function of λ is effectively the same as that of a parametric estimator such as 2SLS. The behavior of the feasible estimator as a function of λ, however, is highly unusual and the asymptotic distribution of the two estimators coincides only when $\lambda \succ 1/\sqrt{k}$. When $1/\sqrt[4]{nk} \prec \lambda \preceq 1/\sqrt{k}$, the feasible estimator is consistent and asymptotically normal, but its variance is nonstandard. The feasible estimator loses consistency when $\lambda \preceq 1/\sqrt[4]{nk}$, compared with $\lambda \preceq 1/\sqrt{n}$ for the infeasible version. The results for the feasible estimator above are similar to those obtained in Chao and Swanson (2005) for the many weak instrument case.

A GMM estimator based on 'linear' instruments z_i will be inconsistent both when $\lambda \preceq 1/\sqrt{n}$ and at all levels of λ when g_i is orthogonal to z_i; to see this point, note that $\mathbb{E}[z_i m_i(\theta)] = \mathbb{E}[z_i g_i](\theta - \theta_0)$. Since no prior information is available for g_i there is no way of knowing ex–ante if one's instruments are sufficiently strong for consistency to obtain. Note also that the asymptotic distribution of $\hat{\theta}$ in example I becomes nonstandard for larger λ than with an estimator based on a naïve choice of instruments such as 2SLS; i.e. semiparametrics requires stronger identification. Therefore, finding a testing method which is robust to weak identification is more valuable when the first stage equation is left unspecified.

We now describe our approach here. Note first that the optimal instruments are given by $p^*(z_i) = g_i(\theta_0)/\sigma_i^2(\theta_0)$, where $g_i(\theta_0) = \mathbb{E}[\partial m_i(\theta_0)/\partial\theta|z_i]$ and $\sigma_i^2(\theta_0) = \text{Var}[m_i(\theta_0)|z_i]$. Since the

[9] In the parametric analysis of Staiger and Stock (1997), $g_i = z_i^\intercal \pi_0 = z_i^\intercal \tilde{\pi}_0/\sqrt{n}$ so that $\lambda = 1/\sqrt{n}$ and $\tilde{g}_i = z_i^\intercal \tilde{\pi}_0$.

[10] For a formal definition of a knn estimator, please see e.g. (6)

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

conditional variance can be ignored under homoskedasticity,[12] we start with the moment condition

$$\mathbb{E}\big[g_i(\theta_0)m_i(\theta_0)\big] = 0. \tag{5}$$

Our semiparametric procedure tests the null hypothesis $H_0 : \theta_0 = \theta_H$ based on a nonparametric estimator of the expectation in (5) evaluated at θ_H. We establish that it has correct size irrespective of λ and discuss its power properties as a function of λ.

In the present section we ignore the conditional variance and focus on (5) because it leads to a simpler (yet optimal under homoskedasticity) procedure, postponing the vector–valued case with estimated variance matrix to section 3.

2.2. **The Test Statistic.** We assume i.i.d. data throughout. We begin by describing our knn method. We define the *nearest neighbor* weights w_{ij}'s as follows. Let k be such that $n^{3/5+c_k} \prec k \prec n$ for some $c_k > 0$. For each observation i, observation j is assigned a nearest neighbor weight w_{ij}, which is positive only if z_j belongs to the k nearest neighbors of z_i, where $j \neq i$ and the distance between z_i and z_j is measured by the Euclidean distance.[13] In the case of ties positive weights are assigned randomly among the ties. Further, the weights $\{w_{ij}\}$ are chosen such that $w_{ii} = 0$, $\sum_{j=1}^n w_{ij} = 1$, and for some fixed C_w^-, C_w, either $w_{ij} = 0$ or $0 < C_w^- \leq k w_{ij} \leq C_w$.

One can for instance use uniform weights $\{w_{ij}\}$, where $w_{ij} = 1/k$ for any observation $j(\neq i)$ whose z_j belongs to the k closest observations to z_i in terms of the Euclidean norm. The condition $w_{ii} = 0$ removes the bias arising from the correlation between the moment function and its derivative. Since the weights are chosen by the researcher and there are always weights that satisfy the above requirements, the conditions imposed on the weights are innocuous. For further discussion on nearest neighbor weights, see e.g. Stone (1977); Robinson (1987); Jun and Pinkse (2009).

To define our test statistic based on (5), we first estimate $g_i(\theta)$ by the knn method; i.e.

$$\hat{g}_i(\theta) = \sum_{j=1}^n w_{ij} m_{\theta i}^\intercal(\theta), \tag{6}$$

where $m_{\theta i}(\theta) = \partial m_i(\theta)/\partial\theta^\intercal$. Our statistic for testing $H_0 : \theta_0 = \theta_H$ is then given by $S(\theta_H) = \|\boldsymbol{t}(\theta_H)\|^2$,[14] where

$$\boldsymbol{t}(\theta) = \hat{\hat{D}}_s^{-1}(\theta)\hat{\hat{N}}_s(\theta), \tag{7}$$

with (omitting the θ–argument)

$$\begin{cases} \hat{\hat{D}}_s^2 &= \sum_i \hat{g}_i \hat{g}_i^\intercal m_i^2 - n^{-1}\big(\sum_i \hat{g}_i m_i\big)\big(\sum_i \hat{g}_i^\intercal m_i\big) + \sum_{ij} w_{ij} w_{ji} m_{\theta j}^\intercal m_{\theta i} m_i m_j, \\ \hat{\hat{N}}_s &= \sum_{i=1}^n \hat{g}_i m_i. \end{cases}$$

[12] I.e., $\mathrm{Var}[m_i(\theta_0)|z_i]$ does not depend on z_i.

[13] Alternatively, the Mahalonobis distance $(z_i - z_j)^\intercal \big(\sum_{s=1}^n z_s z_s^\intercal\big)^{-1}(z_i - z_j)$ can be used to address scaling issues. Note however that taking the Mahalanobis transformation before applying the Euclidean distance is asymptotically irrelevant, albeit that we do not specifically address the issue of 'estimated' transformations.

[14] When $d_\theta = 1$, $\boldsymbol{t}(\theta_H)$ can be used as a statistic instead of $S(\theta_H)$, which can be useful for one–sided testing.

Note that $n^{-1}\sum_i \hat{g}_i(\theta_H)m_i(\theta_H)$ is an estimator of $\mathbb{E}[g_1(\theta_H)m_1(\theta_H)]$, which equals zero under the null hypothesis. So, $S(\theta_H)$ is a semiparametric version of the Anderson–Rubin statistic (e.g. Anderson and Rubin, 1949; Stock and Wright, 2000). However, the statistic has an unusual feature; $\hat{\hat{D}}_s^2(\theta)$ has an additional correction term, which is asymptotically irrelevant if identification is strong.

2.3. **Size, Power, and Local Power Optimality.** We now discuss the asymptotic properties of the statistic $S(\theta_H)$. Note that

$$\sum_{i=1}^{n}\hat{g}_i(\theta_H)m_i(\theta_H) = \sum_{i=1}^{n}\big(\hat{g}_i(\theta_H) - g_i(\theta_H)\big)m_i(\theta_H) + \lambda\sum_{i=1}^{n}\tilde{g}_i(\theta_H)m_i(\theta_H), \tag{8}$$

where $\tilde{g}_i(\theta) = \partial\tilde{M}_i(\theta)/\partial\theta$. If $\lambda \neq 0$ is fixed, the first term on the right hand side of (8) can be shown to be asymptotically negligible relative to the second term. However, if λ is very small, the estimation error in $\hat{g}_i$ can dominate the second term, in which case the nonparametric estimation error must be taken into account. In what follows we study the behavior of $\boldsymbol{t}(\theta_H)$ as a function of λ, allowing for such estimation error.

Let $u_i(\theta) = m_i(\theta) - M_i(\theta)$ and $v_i(\theta) = m_{\theta i}^{\intercal}(\theta) - g_i(\theta)$. We make the following assumptions.

Assumption A. $\mathbb{E}\big[\mathbb{E}\big(m_1^2(\theta_H)|z_1\big)\mathbb{E}\big(v_1^2(\theta_H)|z_1\big) - \big\{\mathbb{E}\big(m_1(\theta_H)v_1(\theta_H)|z_1\big)\big\}^2\big] > 0$.[15]

If $\mathrm{Var}[m_1(\theta_H)|z_1 = z], \mathrm{Var}[v_1(\theta_H)|z_1 = z]$ are bounded away from zero on the support of z_1, as is often assumed (see e.g. Robinson, 1987), then assumption A is implied by

$$\mathbb{P}\Big[\big|\mathrm{Corr}(m_1(\theta_H), v_1(\theta_H)|z_1)\big| = 1\Big] < 1,$$

which is unlikely ever to be violated in empirical work.

Assumption B. $\mathbb{E}\big[|\tilde{M}_1(\theta_H)|^4\big] < \infty, \mathbb{E}\big[\|\tilde{g}_1(\theta_H)\|^4\big] < \infty$, *and* $\mathbb{E}\big[\big\{\mathbb{E}(\|u_1(\theta_H)\|^4|z_1)\big\}^4\big] < \infty$. *Further,* $\mathbb{E}\big[\big\{\mathbb{E}(|v_1(\theta_H)|^4|z_1)\big\}^4\big] < \infty$ *and* $\mathrm{Var}[\tilde{g}_1(\theta_H)m_1(\theta_H)] > 0$.

Assumption B excludes fat–tailed error distributions, which is a common regularity assumption in cross section analysis.

Theorem 1. *Suppose that assumptions A and B hold. Under the null of* $H_0 : \theta_0 = \theta_H$, $\boldsymbol{t}(\theta_H) \xrightarrow{d} N(0, I_{d_\theta})$ *and hence* $S(\theta_H) \xrightarrow{d} \chi^2(d_\theta)$ *regardless of the (nonincreasing)* λ*–sequence.*

Please note that we have not assumed homoskedasticity for theorem 1; it holds even under heteroskedasticity of unknown form.

Theorem 1 establishes our first and most important desideratum, i.e. asymptotically correct size regardless of the quality of identification. Given that our test has correct size, we now proceed with an analysis of its power.

[15] We implicitly assume here that only the conditional means vary with λ (and hence n). One can allow the conditional (co)variances depend on λ by taking infimums over n.

We begin with a discussion of consistency against a fixed alternative $H_1 : \theta_0 \neq \theta_H$. For this we need to discuss the issue of identification since consistency cannot be obtained without (sufficiently strong) identification.

Assumption C. $\mathbb{E}\left[\tilde{g}_1(\theta_H)\tilde{M}_1(\theta_H)\right] \neq 0 \Leftrightarrow \theta_H \neq \theta_0$.

Assumption C means that λ provides a complete description of identification quality; without assumption C, consistency against $H_1 : \theta_0 \neq \theta_H$ does not obtain even for fixed $\lambda \neq 0$. But consistency moreover requires sufficient identification quality, i.e. λ not going to zero too fast. For parametric weak identification–robust tests, consistency requires that $\lambda \succ 1/\sqrt{n}$. In the semiparametric case, $\lambda \succ 1/\sqrt[4]{nk}$ is the best achievable rate. To see why this is so, consider

$$\sum_{i=1}^{n} \hat{g}_i(\theta_H)m_i(\theta_H) = \sum_{i=1}^{n} \hat{g}_i(\theta_H)u_i(\theta_H) + \lambda \sum_{i=1}^{n} \big(\hat{g}_i(\theta_H) - g_i(\theta_H)\big)\tilde{M}_i(\theta_H) + \lambda^2 \sum_{i=1}^{n} \tilde{g}_i(\theta_H)\tilde{M}_i(\theta_H). \quad (9)$$

In the proof of theorem 1 we establish that the first term on the right hand side of (9) is $O_p(\rho)$ with $\rho = \sqrt{n}(\lambda + 1/\sqrt{k})$. The middle term on the right hand side of (9) can be shown to be negligible relative to the other terms and the last term is $O_p(n\lambda^2)$. Because the first term on the right hand side of (9) has a limiting normal distribution (after renorming), the last right hand side term must dominate the first for consistency to obtain. Such dominance only occurs when $n\lambda^2 \succ \sqrt{n/k}$, i.e. when $\lambda \succ 1/\sqrt[4]{nk}$. This rate is exactly the cutoff rate for consistency of the semiparametric estimator discussed in example I, which is no coincidence; if a parameter cannot be estimated consistently it cannot be tested consistently.

In the parametric case with $g_i = z_i^\intercal \tilde{\pi}_0 \lambda$ and $\hat{g}_i = z_i^\intercal \hat{\pi}$ with $\hat{\pi} = \lambda\tilde{\pi}_0 + O_p(1/\sqrt{n})$, the first term on the right hand side of (9) becomes $\hat{\pi}^\intercal \sum_i z_i u_i(\theta_H) = O_p(\sqrt{n}\lambda + 1)$ such that consistency requires that $n\lambda^2 \succ \sqrt{n}\lambda + 1$ or $\lambda \succ 1/\sqrt{n}$. So the cause for the discrepancy is the fact that the nonparametric estimator of g_i converges more slowly than does the parametric one. It should be noted that the rate at which k increases can be chosen to be close to n, in which case the difference is minor.[16]

Theorem 2. *Let assumptions A to C hold and let $\lambda \succ 1/\sqrt[4]{nk}$. Under $H_1 : \theta_0 \neq \theta_H$,*
(i) $\forall C < \infty : \lim_{n\to\infty} \mathbb{P}\left[\boldsymbol{S}(\theta) > C\right] = 1$, (ii) if $\lambda \succ 1/\sqrt{k}$, $\boldsymbol{S}(\theta_H) = \|\mathscr{B}_n(\theta_H)\|^2 + o_p(n\lambda^2)$, where $\mathscr{B}_n(\theta_H) = \sqrt{n}\lambda\big(\mathrm{Var}\left[\tilde{g}_1(\theta_H)m_1(\theta_H)\right]\big)^{-1/2}\mathbb{E}\left[\tilde{g}_1(\theta_H)\tilde{M}_1(\theta_H)\right]$.

Although not stated in theorem 2 it should be clear that $\boldsymbol{t}(\theta_H)$ can be used for one–sided testing when $d_\theta = 1$. For instance, if $\mathbb{E}[\tilde{g}_1(\theta)\tilde{M}_1(\theta)]$ is continuous in θ and zero at θ_0 then, in view of assumption C, it is positive when $\theta_H > \theta_0$ and negative when $\theta_H < \theta_0$.

Part (ii) of theorem 2 implies that if $\lambda \succ 1/\sqrt{k}$ then $\boldsymbol{S}(\theta_H)$ is asymptotically equivalent to an infeasible AR–test based on (5). Recall that the power of a traditional Wald test based on a semiparametric estimator (i.e. N–test) comes from $\mathscr{B}_n^o(\theta_H) = \sqrt{n}\lambda(\theta_0 - \theta_H)\mathscr{B}$ where $\mathscr{B} = \left\{\mathrm{Var}\left[\tilde{g}_1(\theta_0)m_1(\theta_0)\right]\right\}^{-1/2}\mathbb{E}\left[\tilde{g}_1(\theta_0)\tilde{g}_1^\intercal(\theta_0)\right]$, which is comparable but not generally equal to $\mathscr{B}_n(\theta_H)$.

[16] However, using k too large causes size distortions in practice although it helps power under strong identification.

Comparing $\mathscr{B}_n(\theta_H)$ to $\mathscr{B}_n^o(\theta_H)$ allows us to draw some conclusions about the local power of our test. For the sequence $\theta_H = \theta_n = \theta_0 + \Delta/(\sqrt{n}\lambda)$, $\mathscr{B}_n(\theta_n) = \mathscr{B}\Delta + o(1)$ while $\mathscr{B}_n^o(\theta_n) = -\mathscr{B}\Delta$. Therefore, when identification is sufficiently strong (i.e. $\lambda \succ 1/\sqrt{k}$), our test has the same local power as the usual Wald test based on a semiparametric estimator using $\hat{g}_i$.[17] Theorem 3 formalizes this idea.

Theorem 3. *Consider $H_{1L}: \theta_H = \theta_n = \theta_0 + \Delta/(\sqrt{n}\lambda)$ and let assumptions A to C hold. Let further $\lambda \succ 1/\sqrt{k}$ and $k \prec n/(\log n)^4$. Further, suppose that $\mathbb{E}\left[\sup_{\mathscr{N}^*} |u_i(\theta)|^{p_u}\right], \mathbb{E}\left[\sup_{\mathscr{N}^*} |v_i(\theta)|^{p_v}\right] < \infty$ for some $1/p_u + 1/p_v < 1/2$ in an open neighborhood $\mathscr{N}^*$ of θ_0. Then under H_{1L} we have*

$$\boldsymbol{t}(\theta_n) \xrightarrow{d} N(\mathscr{B}\Delta, I_{d_\theta}). \tag{10}$$

Consequently, $\boldsymbol{S}(\theta_n)$ converges to a noncentral χ^2–distribution with d_θ degrees of freedom and noncentrality parameter $\mathscr{B}\Delta$; note that $\lambda\mathscr{B}^\intercal\mathscr{B}$ is the inverse of the semiparametric efficiency bound under homoskedasticity with fixed $\lambda \neq 0$.

The rate at which the local alternative approximates the null is $1/(\sqrt{n}\lambda)$, not the usual $1/\sqrt{n}$–rate. The $1/(\sqrt{n}\lambda)$–rate is not specific to our test but applies to all tests under weak identification; see e.g. Andrews, Moreira, and Stock (2004, 2006).

2.4. **Nuisance Parameters.** Up to now, we have only considered testing the entire parameter vector of the model. It is, however, more realistic to consider a model that contains both the parameter(s) of interest and a vector of nuisance parameters. We will write θ_0 and β_0 for the subvector of interest and the nuisance parameter vector, respectively. We could for instance have $m_i(\theta_0, \beta_0) = y_i - Y_i\theta_0 - x_i^\intercal\beta_0$. Under additional conditions discussed below, all results derived earlier go through in the presence of nuisance parameters provided that the nuisance parameters can be consistently estimated under the null hypothesis.

We define $\beta(\theta)$ as

$$\beta(\theta) = \operatorname*{argmin}_\beta \mathbb{E}\left[\mathbb{E}\left\{m_1(\theta, \beta)|z_1\right\}\right]^2, \tag{11}$$

which is assumed to be unique. Other definitions of $\beta(\theta)$ are conceivable and it is possible that $\beta(\theta)$ is not uniquely determined by (11). In linear models this would occur when the number of good instruments is less than the total number of regressors minus d_θ. If there are only d_θ endogenous regressors with coefficient vector θ then uniquenss of $\beta(\theta)$ requires the absence of multicollinearity of the exogenous regressors. With more than d_θ endogenous regressors the assumption of the availability of good instruments can be unrealistic and goes against the spirit of the weak identification literature, but there are sufficiently many applications in which it is warranted to allow for identification of nuisance parameters once the parameters of interest are known; see e.g. Kleibergen (2002, 2005). For nonlinear models conditions under which $\beta(\theta)$ is uniquely defined are discussed at length in the GMM literature.

[17] The power of our test against a fixed alternative can be greater or less than that of the Wald test depending on the alternative.

Let $h_i(\theta,\beta) = E\left[m_{\beta i}^{\intercal}(\theta,\beta)|z_i\right]$ and let $h_i(\theta) = h_i\big(\theta,\beta(\theta)\big)$. All other symbols that depend on both θ and β are similarly defined. In particular, $M_i(\theta) = M_i\big(\theta,\beta(\theta)\big) = \lambda\tilde{M}_i(\theta) = \lambda\tilde{M}_i\big(\theta,\beta(\theta)\big)$. Note that $\beta(\theta)$ satisfies the first order condition

$$\mathbb{E}\left[h_1\big(\theta,\beta(\theta)\big)m_1\big(\theta,\beta(\theta)\big)\right] = 0, \tag{12}$$

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

We propose to estimate $\beta(\theta)$ on the basis of the moment condition (12), which requires its identification.

Assumption D. *For all $\theta \in \Theta$ there is a unique $\beta = \beta(\theta) \in \mathcal{B}$, with $\mathcal{B}$ compact, such that (12) is satisfied. Further, at this unique $\beta = \beta(\theta)$ the Jacobian $Q = Q(\theta) = \mathbb{E}\left[h_1\big(\theta\big)h_1^{\intercal}(\theta) + M_1(\theta)m_{\beta\beta 1}(\theta)\right]$ is invertible.*

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

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

$$\sum_i \hat{h}_i\big(\theta,\hat{\beta}(\theta)\big)m_i\big(\theta,\hat{\beta}(\theta)\big) = 0, \tag{13}$$

where $\hat{h}$ is a knn estimator of h. Under conditions to be outlined below, we will show that $\hat{\beta}(\theta)$ is a consistent estimator of $\beta(\theta)$. In fact, under the null hypothesis $\hat{\beta}(\theta) = \hat{\beta}(\theta_0)$ is a $\sqrt{n}$–consistent estimator achieving the semiparametric efficiency bound for the estimation of $\beta(\theta_0)$ under homoskedasticity. Under the alternative hypothesis, however, $\hat{\beta}(\theta)$ generally converges at a rate slower than $\sqrt{n}$. This is innocuous since it converges fast enough to ensure that the consistency properties of our test statistic are not affected provided that $\lambda \succ 1/\sqrt[4]{nk}$, as was assumed in theorem 2. The discrepancy in convergence rates under the null and alternative hypotheses is due to the fact that $M_1(\theta_0) = \mathbb{E}\left[m_1\big(\theta_0,\beta(\theta_0)\big)|z_1\right] = 0$ a.s. but $M_1(\theta)$ is not generally zero for other values of θ.[18] Since we maintain the setup of (2) here, the only exception arises when $\lambda = 0$.

The moment condition we tested previously, i.e. (5), can continue to be used here, albeit that now $g_i(\theta) = g_i\big(\theta,\beta(\theta)\big)$ and $m_i(\theta) = m_i\big(\theta,\beta(\theta)\big)$. However, it is more convenient to replace (5) with an equivalent one from which the effect of the estimation of the nuisance parameters has been isolated, i.e. letting $q_i(\theta) = q_i\big(\theta,\beta(\theta)\big) = g_i(\theta) - \kappa^{\intercal}(\theta)h_i(\theta)$,

$$\mathbb{E}\left[q_1(\theta_0)m_1(\theta_0)\right] = 0, \tag{14}$$

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

where $\kappa(\theta) = \kappa\big(\theta, \beta(\theta)\big) = \big(\mathbb{E}\big[h_1(\theta)h_1^\intercal(\theta)\big]\big)^{-1}\mathbb{E}\big[h_1(\theta)g_1^\intercal(\theta)\big]$. Let $\hat{q}_i(\theta) = \hat{g}_i(\theta) - \kappa^\intercal(\theta)\hat{h}_i(\theta)$ and further $\hat{\hat{q}}_i(\theta) = \hat{\hat{g}}_i(\theta) - \hat{\kappa}(\theta)^\intercal\hat{\hat{h}}_i(\theta)$, where $\hat{\hat{g}}_i(\theta) = \hat{g}_i(\theta, \hat{\beta}(\theta))$, $\hat{\hat{h}}_i(\theta) = \hat{g}_i(\theta, \hat{\beta}(\theta))$, and $\hat{\kappa}(\theta) = \big(\sum_{i=1}^n \hat{\hat{h}}_i(\theta)\hat{\hat{h}}_i^\intercal(\theta)\big)^{-1}\big(\sum_{i=1}^n \hat{\hat{h}}_i(\theta)\hat{\hat{g}}_i^\intercal(\theta)\big)$. Condition (14) has the appealing feature that for $\hat{m}_i(\theta) = m_i\big(\theta, \hat{\beta}(\theta)\big)$ the difference between $\sum_i \hat{\hat{q}}_i(\theta)\hat{m}_i(\theta)$ and $\sum_i \hat{q}_i(\theta)m_i(\theta)$ is asymptotically negligble. The transition from g to q is similar to 'partialing out' other regressors in a linear regression model.

Our test statistic then becomes

$$\hat{\boldsymbol{t}}(\theta) = \hat{\hat{\mathscr{D}}}_N^{-1}(\theta)\sum_i \hat{\hat{q}}_i(\theta)\hat{m}_i(\theta), \tag{15}$$

where $\hat{m}_{\theta i}(\theta) = m_{\theta i}\big(\theta, \hat{\beta}(\theta)\big)$ and (omitting the θ argument)

$$\hat{\hat{\mathscr{D}}}_N^2 = \sum_i \hat{\hat{q}}_i\hat{\hat{q}}_i^\intercal\hat{m}_i^2 - n^{-1}\Big(\sum_i \hat{\hat{q}}_i\hat{m}_i\Big)\Big(\sum_i \hat{\hat{q}}_i^\intercal\hat{m}_i\Big) + \sum_{ij} w_{ij}w_{ji}\hat{m}_j\hat{m}_{\theta j}^\intercal\hat{m}_i\hat{m}_{\theta i}.$$

The only difference between $\boldsymbol{t}$ and $\hat{\boldsymbol{t}}$ is that we now use q instead of g and that the right hand side quantities in (15) depend on $\hat{\beta}$. All assumptions made earlier will hence now be applied to q instead of g.

Previously g, being M's derivative, varied proportionally to λ. Now the entire q–function varies with λ. To see this, note that $M_{\theta i}(\theta) = g_i^\intercal(\theta) + h_i^\intercal(\theta)\beta_\theta(\theta)$, and that by the implicit function theorem it follows from (12) that for $Q(\theta) = E\big[h_1(\theta)h_1^\intercal(\theta) + M_1(\theta)m_{\beta\beta 1}(\theta)\big]$,

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

Hence since $M_{\theta i}(\theta) = q_i^\intercal(\theta) + h_i^\intercal(\theta)\big[\beta_\theta(\theta) + \kappa(\theta)\big]$, q_i varies proportionally to λ up to terms of order λ^2.

We are now in a position to state our assumptions and formulate our nuisance parameter theorem. Most assumptions made previously carry over albeit that they are now made with respect to q instead of g. The main implication of this shift is that assumption C requires (i) that there is a unique solution $\beta(\theta)$ to (12), which was assumed in assumption D, and (ii) that there is a unique combination (θ_0, β_0) that zeroes both $\mathbb{E}\big[\tilde{g}_1(\theta, \beta)m_1(\theta, \beta)\big]$ and $\mathbb{E}\big[h_1(\theta, \beta)m_1(\theta, \beta)\big]$. Therefore, (ii) requires that (θ_0, β_0) are identified for fixed λ if g_i, h_i are used as instruments. The only additional assumptions we make here relate to the smoothness of m.

Assumption E. *For $f = m, m_\theta, m_\beta, m_{\theta\beta}$ we have $\mathbb{E}\big[\mathbb{E}\big(\sup_{\beta\in\mathcal{B}}||f_1(\theta, \beta)||^2|z_1\big)\big]^2 < \infty$. Further, $\mathbb{E}\big[\mathbb{E}\big(\sup_{\beta\in\mathcal{B}}||m_{\beta\beta 1}(\theta, \beta)||\,|z_1\big)\big]^2 < \infty$.*

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

Theorem 4. *If assumptions A to C hold when g is replaced with q and moreover assumptions D and E are satisfied then theorems 1 and 2 hold when g is replaced with q and $\boldsymbol{t}$ with $\hat{\boldsymbol{t}}$.*

3. The General Case

In this section we provide a generalization of our test statistic in two respects: we allow for a vector–valued moment function m_i, as one would have in a demand/supply system, and we consider optimal instruments allowing for heteroskedasticity of unknown form using nonparametric estimation of the conditional variance function.

Consider the model in (1), where m_i now takes values in $\mathbb{R}^{d_m}$ for $d_m \geq 1$. Then, the unconditional moment equations based on optimal instruments are given by

$$\mathbb{E}\big[G_i^\intercal(\theta_0)\mathscr{V}_i^{-1}(\theta_0)m_i(\theta_0)\big] = 0, \tag{16}$$

where $G_i(\theta) = E\big[m_{\theta i}(\theta)|z_i\big]$ and $\mathscr{V}_i(\theta) = \mathrm{Var}\big[m_i(\theta)|z_i\big]$.[19] A semiparametric version of the AR statistic can be constructed based on (16).

Let $\hat{G}_i$ and $\hat{\mathscr{V}}_i$ be the (knn) nonparametric estimators for G_i and $\mathscr{V}_i$, respectively; i.e.

$$\hat{G}_i(\theta) = \sum_{j=1}^n w_{ij}m_{\theta j}(\theta) \quad \text{and} \quad \hat{\mathscr{V}}_i(\theta) = \sum_{j=1}^n w_{ij}\big(m_j(\theta) - \hat{\mu}_j(\theta)\big)\big(m_j(\theta) - \hat{\mu}_j(\theta)\big)^\intercal,$$

where $\hat{\mu}_i(\theta) = \sum_{j=1}^n w_{ij}m_j(\theta)$. We then define a statistic $\boldsymbol{T}(\theta_H)$ for testing $H_0 : \theta_0 = \theta_H$, where

$$\boldsymbol{T}(\theta) = \Big(\sum_i \hat{G}_i^\intercal(\theta)\hat{\mathscr{V}}_i^{-1}(\theta)m_i(\theta)\Big)^\intercal \hat{D}^{-2}(\theta)\Big(\sum_i \hat{G}_i^\intercal(\theta)\hat{\mathscr{V}}_i^{-1}(\theta)m_i(\theta)\Big), \tag{17}$$

where (omitting the θ–argument)

$$\hat{D}^2 = \sum_i \hat{G}_i^\intercal\hat{\mathscr{V}}_i^{-1}m_i m_i^\intercal\hat{\mathscr{V}}_i^{-1}\hat{G}_i - n^{-1}\Big(\sum_i \hat{G}_i^\intercal\hat{\mathscr{V}}_i^{-1}m_i\Big)\Big(\sum_i \hat{G}_i^\intercal\hat{\mathscr{V}}_i^{-1}m_i\Big)^\intercal + \sum_{ij} w_{ij}w_{ji}m_{\theta j}^\intercal\hat{\mathscr{V}}_i^{-1}m_i m_j^\intercal\hat{\mathscr{V}}_j^{-1}m_{\theta i}.$$

$\boldsymbol{T}(\theta_H)$ is similar to, but more complicated than, $\boldsymbol{S}(\theta_H)$ in that $\boldsymbol{T}(\theta_H)$ involves nonparametric estimation of conditional variances.

The asymptotic properties of $\boldsymbol{T}(\theta_H)$ are similar to those of $\boldsymbol{S}(\theta_H)$, but the proofs are more complicated. Let $u_i(\theta) = m_i(\theta) - M_i(\theta)$ as before and let $V_i(\theta) = m_{\theta i}(\theta) - G_i(\theta)$. Further, let $\underline{\mathrm{ev}}, \overline{\mathrm{ev}}$ denote the smallest and largest eigenvalues of their arguments, respectively. We will use tr to denote the trace of a matrix. For any matrix A, let $\|A\| = \sqrt{\overline{\mathrm{ev}}(A^\intercal A)}$. The following assumptions correspond to assumptions A to C. The comments on assumptions A to C also apply to the following assumptions.

Assumption AG. $\inf_z \underline{\mathrm{ev}}\big(\mathrm{Var}[u_1(\theta_H)|z_1 = z]\big) > 0$ *and for all* $c \in \mathbb{R}^{d_\theta}$ *with* $\|c\| = 1$,
$\mathbb{E}\,\mathrm{tr}\big(\mathbb{E}\big\{\mathrm{Var}\big[u_1(\theta_H)|z_1\big]\mathrm{Var}\big[V_1(\theta_H)c|z_1\big] - \mathrm{Cov}\big[u_1(\theta_H), V_1(\theta_H)c|z_1\big]\,\mathrm{Cov}\big[u_1(\theta_H), V_1(\theta_H)c|z_1\big]^\intercal\big\}\big) > 0$.

[19] Although $G_i^\intercal = g_i$ when m_i is scalar–valued, we use an upper–case letter to emphasize that the Jacobian is potentially matrix–valued in this section.

Assumption BG. $\mathbb{E}\big[\|\tilde{M}_1(\theta_H)\|^4\big] < \infty$, $\mathbb{E}\big[\|\tilde{G}_1(\theta_H)\|^4\big] < \infty$, *and* $\mathbb{E}\big[\big\{\mathbb{E}\big(\|u_1(\theta_H)\|^4|z_1\big)\big\}^4\big] < \infty$. *Further,* $\mathbb{E}\big[\big\{\mathbb{E}\big(\|V_1(\theta_H)\|^4|z_1\big)\big\}^4\big] < \infty$, *and* $\mathbb{E}\big[\tilde{G}_1^\intercal(\theta_H)\mathscr{V}_1^{-1}(\theta_H)\tilde{G}_1(\theta_H)\big] > 0$.

Assumption CG. $\mathbb{E}\big[\tilde{G}_1^\intercal(\theta_H)\mathscr{V}_1^{-1}(\theta_H)M_1(\theta_H)\big] \neq 0 \Leftrightarrow \theta_H \neq \theta_0$.

Theorem 5. *Suppose that assumptions AG and BG hold. Under* $H_0 : \theta_0 = \theta_H$, $\boldsymbol{T}(\theta_H) \xrightarrow{d} \chi^2(d_\theta)$ *regardless of the (nonincreasing)* λ*–sequence.*

Theorem 6. *Let assumptions AG to CG hold and let* $\lambda \succ 1/\sqrt[4]{nk}$. *Under* $H_1 : \theta_0 \neq \theta_H$,
(i) $\forall C < \infty : \lim_{n\to\infty} \mathbb{P}\big[\boldsymbol{T}(\theta_H) > C\big] = 1$, *(ii) if* $\lambda \succ 1/\sqrt{k}$, $\boldsymbol{T}(\theta_H) = \|\mathscr{B}_n^*(\theta_H)\|^2 + o_p(n\lambda^2)$, *where* $\mathscr{B}_n^*(\theta) = \sqrt{n}\lambda\big\{\mathrm{Var}\big[\tilde{G}_1^\intercal(\theta_H)\mathscr{V}_1^{-1}(\theta_H)m_1(\theta_H)\big]\big\}^{-1/2}\mathbb{E}\big[\tilde{G}_1^\intercal(\theta_H)\mathscr{V}_1^{-1}(\theta_H)\tilde{M}_1(\theta_H)\big]$.

Theorem 5 shows that $\boldsymbol{T}(\theta_H)$ is robust to weak identification. Although it does not formally derive the limiting distribution of $\boldsymbol{T}(\theta_H)$ under local alternatives, part (ii) of theorem 6 shows that the dominant term of $\boldsymbol{T}(\theta_H)$ under the alternative is closely related with the semiparametric efficiency bound when identification is sufficiently strong. In particular, when $\theta_H = \theta_n = \theta_0 + \Delta/(\sqrt{n}\lambda)$, the Taylor expansion of $\mathscr{B}_n^*(\theta_H)$ shows that $\mathscr{B}_n^*(\theta_H) = \mathscr{B}^*\Delta + o(1)$, where

$$\mathscr{B}^* = \big\{\mathrm{Var}\big[\tilde{G}_1^\intercal(\theta_0)\mathscr{V}_1^{-1}(\theta_0)m_1(\theta_0)\big]\big\}^{-1/2}\mathbb{E}\big[\tilde{G}_1^\intercal(\theta_0)\mathscr{V}_1^{-1}(\theta_0)\tilde{G}_1(\theta_0)\big].$$

Note that $\lambda\mathscr{B}^{*\intercal}\mathscr{B}^*$ is the inverse of the semiparametric efficiency bound of a conditional moment restrictions model with fixed $\lambda \neq 0$. We do not provide a full generalization of theorem 3 here.

4. Simulations

We now compare several tests proposed in the literature with ours using simulation experiments. Because not all tests can be used when the structural equation is nonlinear, we use the simple model from the introduction, i.e.

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

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

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

where $u_i = 5(\epsilon_i - \frac{1}{2}) + \eta_i$ with ϵ_i and η_i drawn from standard uniform and normal distributions, respectively. [20] Note that the correlation between u_i and ϵ_i is about 0.8, and hence Y_i is highly endogenous. Y_i is a binary variable taking values $\pm 1/2$ with conditional mean (given z_i)

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

which are the optimal instruments. For sufficiently small λ (i.e. less than $1/2$), we note that

$$g(z) = \Phi(z^\intercal\iota)\lambda \to 0 \text{ as } \lambda \to 0,$$

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

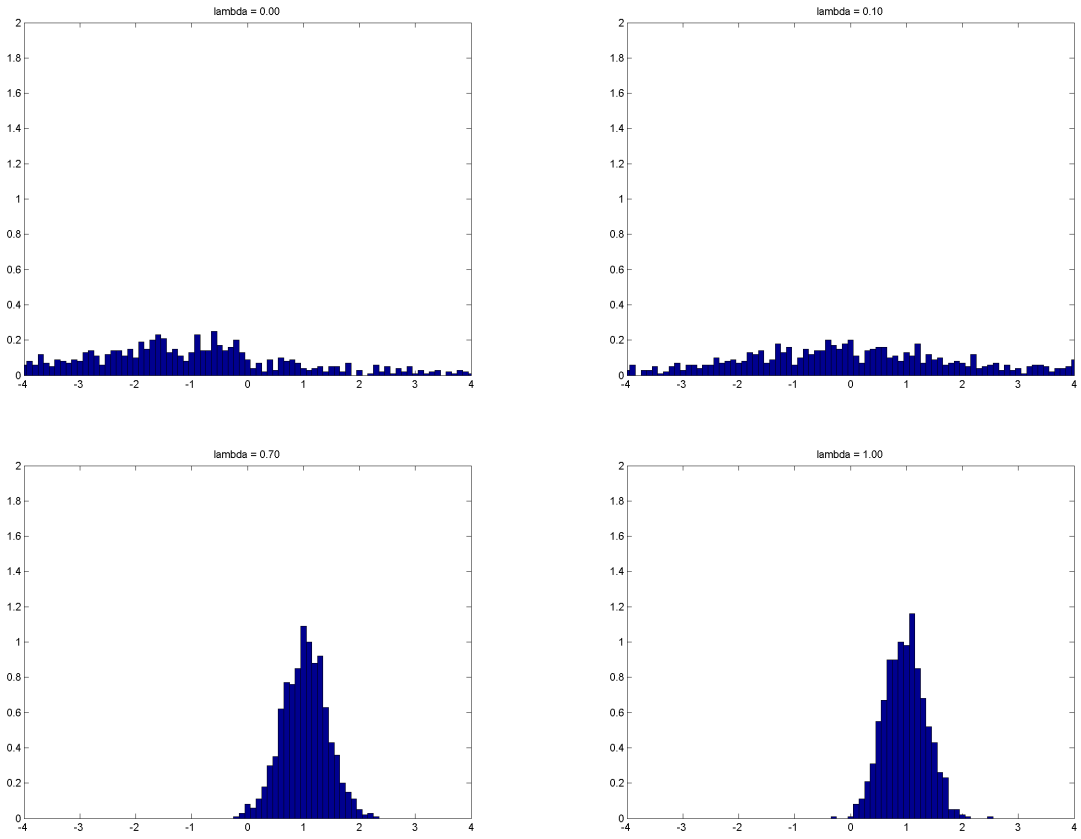

FIGURE 1. Histograms of the semiparametric (knn) estimator $\hat{\theta}$ under DGP1

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

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

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

Recall that the parameter λ indicates the strength of identification; we considered the following values for λ:0,0.05,0.1,0.3,0.5,0.7, 1 and 2. In all cases we test $H_0 : \theta_0 = \theta_H$ versus $H_1 : \theta_0 \neq \theta_H$ (or $H_0 : \theta_0 = \theta_H$ versus $H_1 : \theta_0 < \theta_H$ for one–sided alternatives). The primary goal is for the tests to have correct size properties irrespective of identification strength (as measured by λ), but we also want them to have good power if identification is sufficiently strong. We used heteroskedasticity–robust versions of all statistics (except for the N–test) in the experiments; our test is still valid

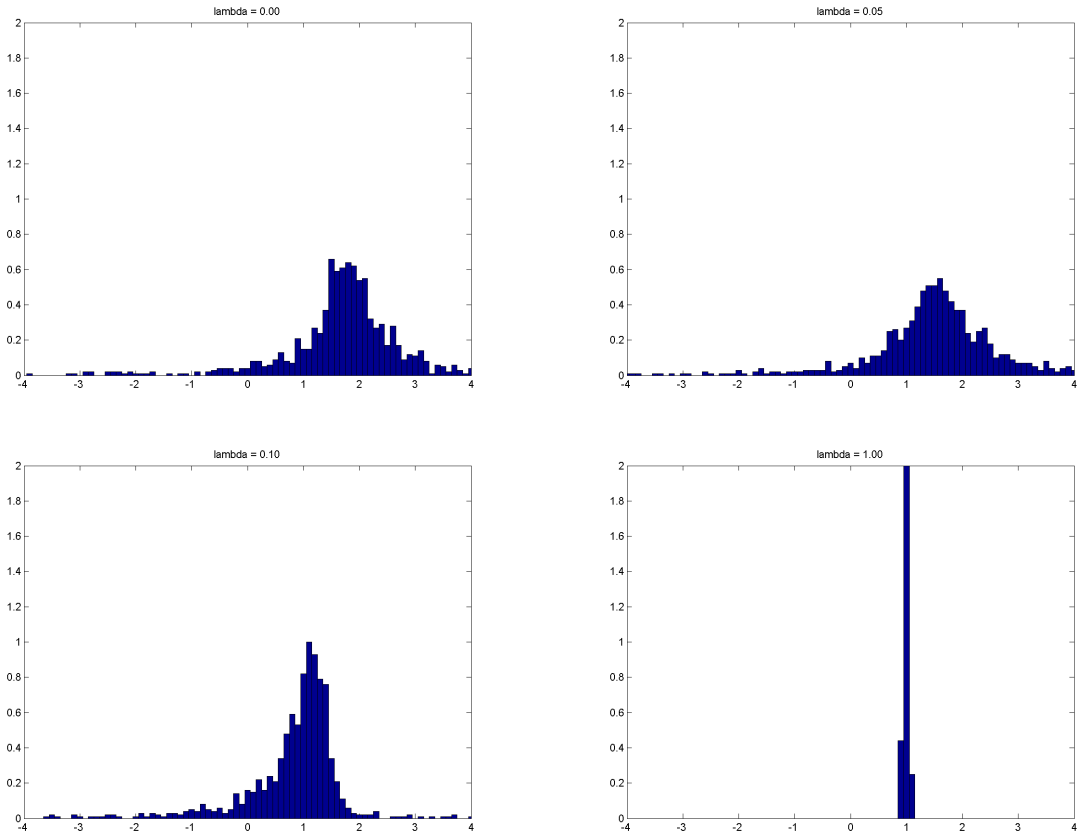

FIGURE 2. Histograms of the semiparametric (knn) estimator $\hat{\theta}$ under DGP2

in the presence of heteroskedasticity.[21] In all experiments, we used $n = 200, k = 70$, and $1,000$ replications.

We now illustrate our earlier finding that the semiparametric estimator $\hat{\theta}$ does not behave as strong identification asymptotics would indicate when λ is small; the results are depicted in figures 1 and 2, which contain histrograms of its distribution. For both DGP's, the histograms are not nicely concentrated around the true $\theta_0 = 1$, when the value of λ is small. This is not surprising in view of the inconsistency of $\hat{\theta}$ when $\lambda \preceq 1/\sqrt[4]{nk}$. We repeated this experiment for the HLIM estimator[22] of Hausman, Newey, Woutersen, and Chao (2007); the graphs (not shown) are similar to those in figures 1 and 2. See also figure 5.

Figures 3 and 4 show the rejection rates of a number of test statistics as a function of θ_H for various values of λ; figure 3 for DGP1 and figure 4 for DGP2. In all cases $\theta_0 = 1$ and in all cases except for the N–test we used heteroskedasticity–robust versions of the test statistics. The nominal

[21] For the heteroskedasticity–robust version of the M–test, see Andrews, Moreira, and Stock (2004).

[22] Heteroskedasticity robust Limited Information Maximum Likelihood estimators. These are LIML type estimators but designed to be robust to heteroskedasticity by jackknife methods. They are consistent under "many–weak" instruments asymptotics that allows the concentration parameter to increase more slowly than n.

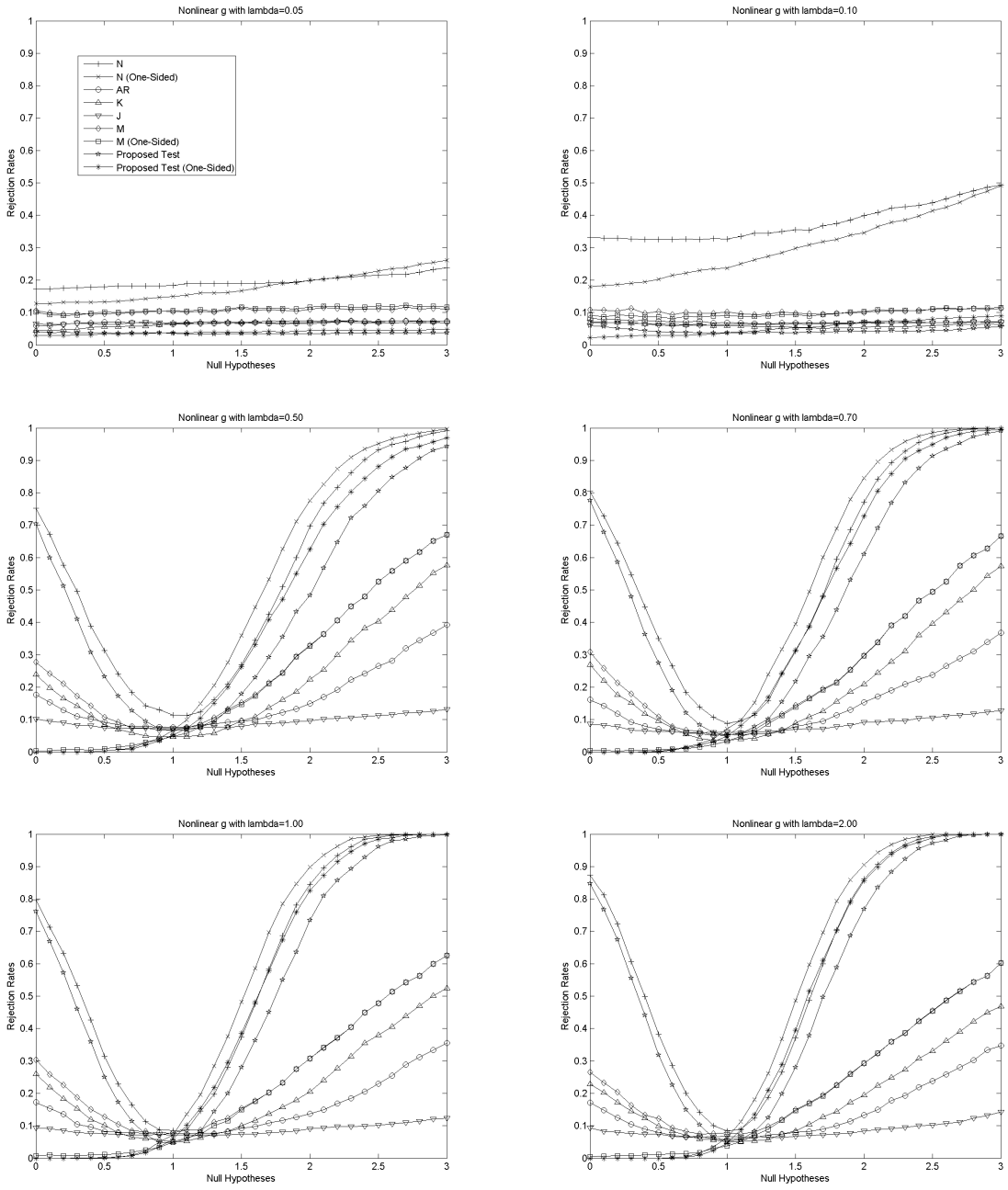

FIGURE 3. Monte Carlo rejection rates for nonlinear g (DGP1)

size of all tests is 0.05. As expected, the t–test based on an efficient point estimator suffers from serious size distortions, when the values of λ are small. But note that the power curves of the proposed tests are close to those of the t–tests based on the efficient estimator when λ is relatively large.

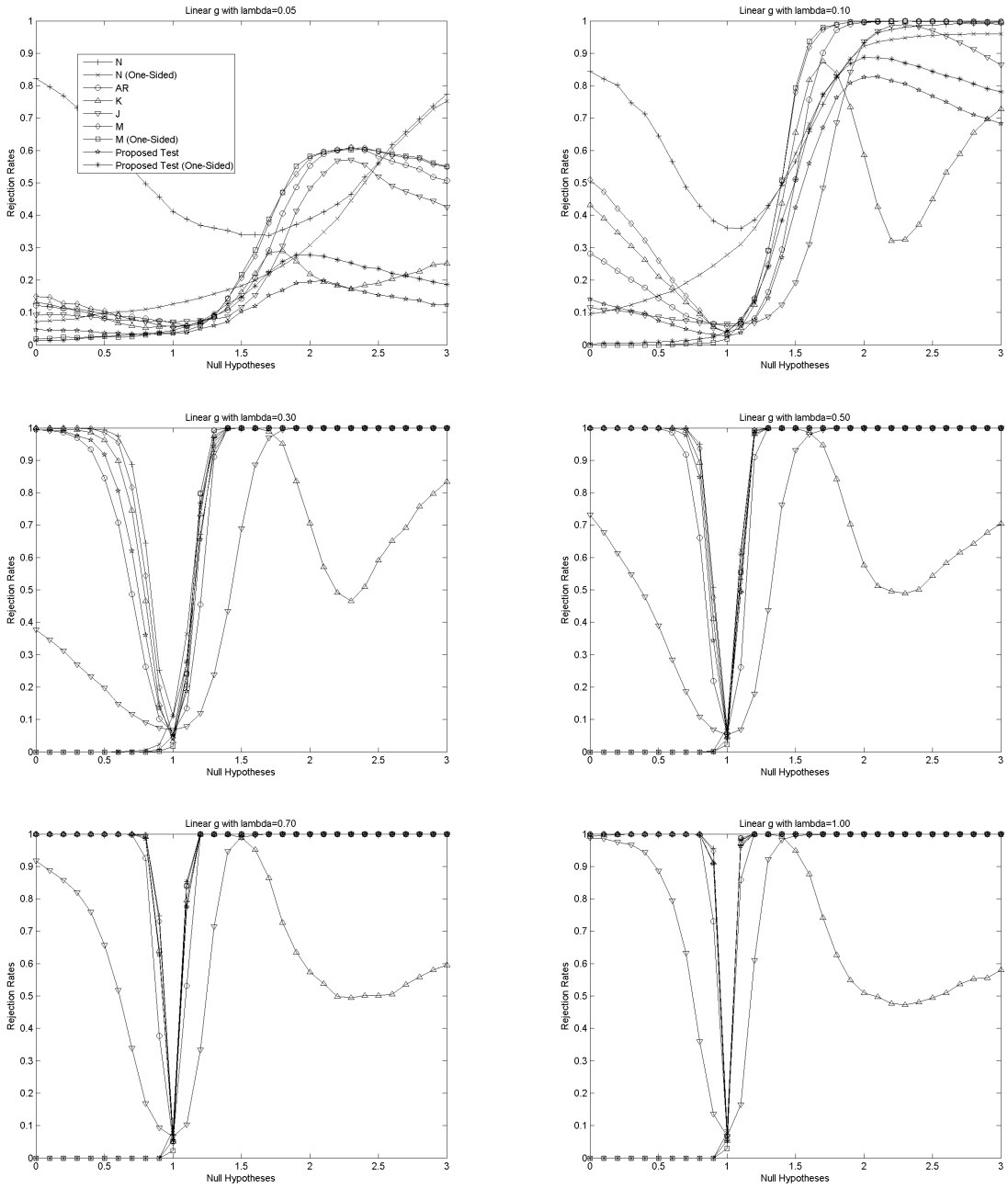

FIGURE 4. Monte Carlo rejection rates for linear g (DGP2)

The size of the M–test is moderately distorted, which may be attributable to the use of the heteroskedasticity–robust version of the test and would gradually disappear with an increase of the sample size. We feel encouraged that our test (which is also robust to heteroskedasticity) appears to have better size properties in these experiments. As we increase λ, the gain from

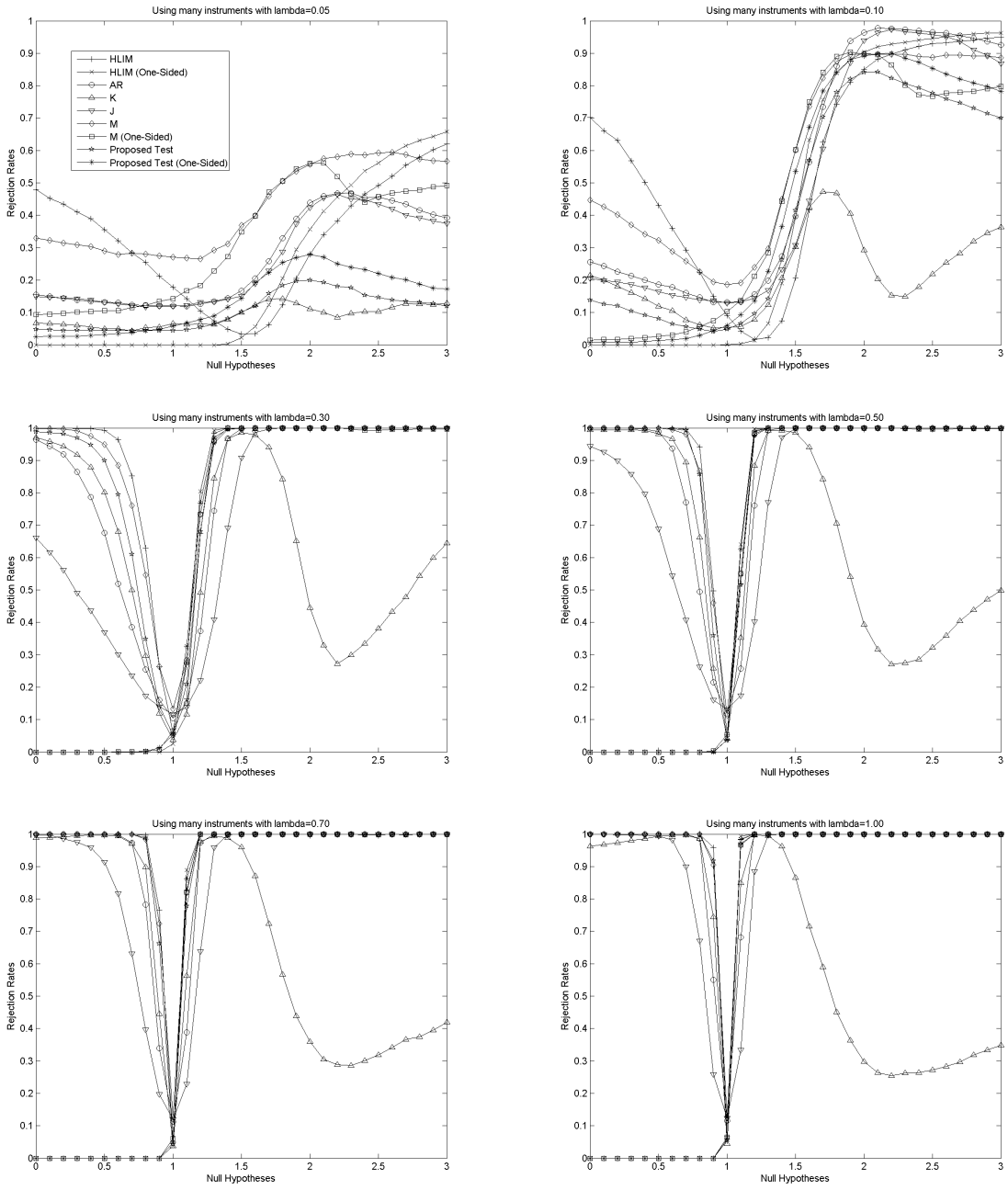

FIGURE 5. Monte Carlo rejection rates of the statistics using many instruments for linear g (DGP2)

using nonparametrically estimated instruments becomes apparent under DGP1 and our test is comparable to the N–test in this case.

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

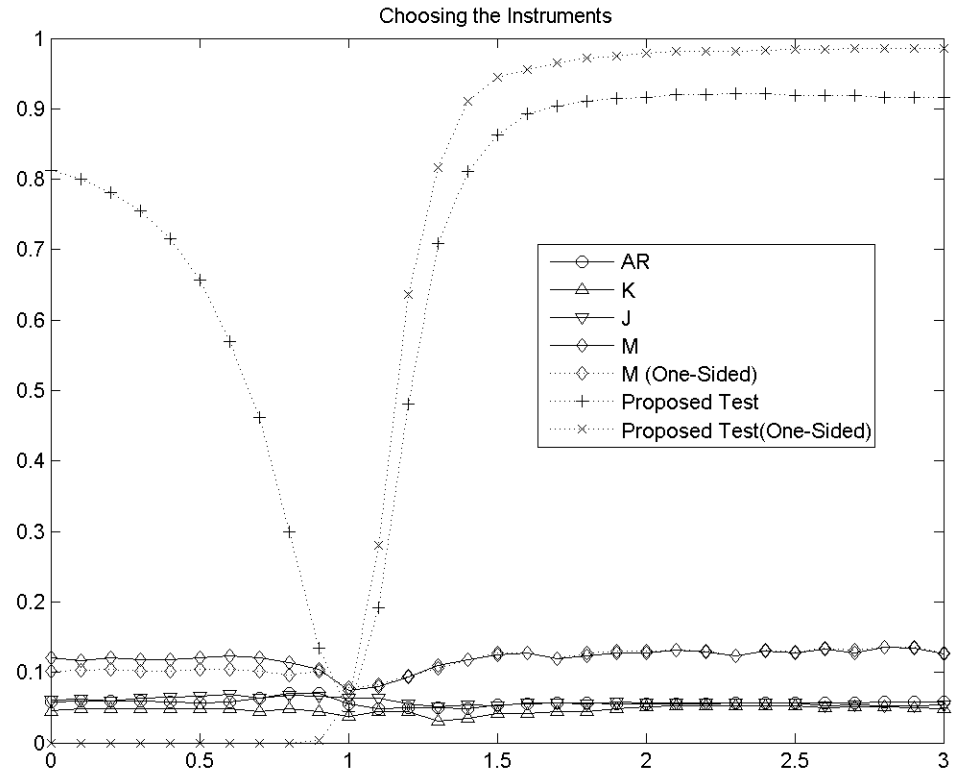

FIGURE 6. Monte Carlo rejection rates for quadratic g (DGP3)

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

In figure 5 we explore the effect of using more instruments, namely z_i, z_i^2 and z_i^3 on the rejection rates of the AR–, the K–, the M–tests, and of the t–test based on HLIM estimators.. The test based on HLIM is incorrectly centered when λ is small, which is due to the fact that the HLIM estimator is then inconsistent. Note that the AR– and the M–tests get size-distorted as more instruments are used and identification becomes weaker. Although these distortions disappear asymptotically,[23] they are a concern in moderate–size samples and do not occur with our statistic.

There is any number of nonlinear specifications we could have chosen for DGP1 and depending on our choice the relative performance of the tests considered will vary. We feel that DGP1 is

[23]Simulations (not graphed) suggest that distortions of the AR– and the M–tests indeed become substantially smaller as the sample size gets bigger, even when $\lambda = 0$.

favoring the parametric tests somewhat since g is a function of a linear index. To emphasize this point we now consider an extreme example to illustrate the fact that using arbitrary unconditional moment conditions generated from unconditional moment conditions can lead to poor inference.

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

where u_i, v_i are drawn from the same joint normal distribution as in DGP2. Note that here $\mathbb{E}\big[z_i(y_i - Y_i\theta)\big] = -\mathbb{E}\big[z_i(||z_i||^2 - 8)\big](\theta - \theta_0) = 0$ for any $\theta \in \mathbb{R}$ because the distribution of the z_i's is even. Therefore, the unconditional moment condition using linear instruments, $\mathbb{E}[z_i(y_i - Y_i\theta)]$, does not identify the parameter of interest, while using $g_i = ||z_i||^2 - 8$ does identify θ_0. Figure 6 shows that the power curves of the parametric tests are indeed almost flat, while the proposed test performs well. It is of course true that a parametric test using a polynomial expansion would have worked well in this particular case. But the whole point of using a nonparametric procedure is that one does not know the correct form of g, such that problems like the one illustrated here arise for any parametric choice for g.

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

5. Conclusions

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

References Cited

Anderson, T. W., and H. Rubin (1949): "Estimation of the parameters of a single equation in a complete system of stochastic equations," *Annals of Mathematical Statistics*, 20, 46–63.

Andrews, D. W., M. J. Moreira, and J. H. Stock (2004): "Optimal invariant similar tests for instrumental variables regression," Discussion paper, Cowles Foundation.

——— (2006): "Optimal two–sided invariant similar tests for instrumental variables regression," *Econometrica*, 74, 715–752.

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

Chao, J. C., and N. R. Swanson (2005): "Consistent estimation with weak instruments," *Econometrica*, 73, 1673–1692.

Domínguez, M., and I. Lobato (2004): "Consistent estimation of models defined by conditional moment restrictions," *Econometrica*, 72(5), 1601–1615.

Dufour, J.-M. (1997): "Some impossibility theorems in econometrics with applications to structural and dynamic models," *Econometrica*, 65, 1365–1387.

Han, C., and P. C. Phillips (2006): "GMM with many moment conditions," *Econometrica*, 74, 147–192.

Hausman, J. A., W. K. Newey, T. Woutersen, and J. C. Chao (2007): "Instrumental variable estimation with heterskedasticity and many instruments," Discussion paper, MIT.

Jun, S. J., and J. Pinkse (2009): "Semiparametric tests of conditional moment restrictions under weak or partial identification," *Journal of Econometrics*, 152, 3–18.

Kleibergen, F. (2002): "Pivotal statistics for testing structural parameters in instrumental variables regression," *Econometrica*, 70, 1781–1804.

——— (2005): "Testing parameters in GMM without assuming that they are identified," *Econometrica*, 73, 1103–1123.

Moreira, M. J. (2003): "A conditional likelihood ratio test for structural models," *Econometrica*, 71, 1027–1048.

Newey, W. K. (1990): "Efficient instrumental variables estimation of nonlinear models," *Econometrica*, 58, 809–837.

——— (1993): *Efficient estimation of models with conditional moment restrictions*pp. 419–454. North Holland.

——— (2004): "Many weak moment asymptotics for the continuously updated GMM estimator," Discussion paper, MIT.

Newey, W. K., and F. Windmeijer (2009): "GMM with many weak moment conditions," *Econometrica*, 77, 687–719.

Phillips, P. (1989): "Partially identified econometric models," *Econometric Theory*, 5(02), 181–240.

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

Staiger, D., and J. H. Stock (1997): "Instrumental variables regression with weak instruments," *Econometrica*, 65, 557–586.

Stock, J. H., and J. H. Wright (2000): "GMM with weak identification," *Econometrica*, 68, 1055–1096.

Stock, J. H., J. H. Wright, and M. Yogo (2002): "A survey of weak instruments and weak identification in generalized method of moments," *Journal of Business and Economic Statistics*, 20, 518–529.

Stock, J. H., and M. Yogo (2005): "Asymptotic distributions of instrumental variables statistics with many instruments," in *Identification and inference for econometric models: essays in honor of Thomas Rothenberg*, pp. 109–120. Cambridge University Press.

Stone, C. J. (1977): "Consistent nonparametric regression," *Annals of Statistics*, 5, 595–620.

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

Appendix A. Proofs of Theorems

Throughout the appendix we use the abbreviations RHS and LHS for *right hand side* and *left hand side*, respectively. Further, RHS1 means the first right hand side term, and similarly for RHS2, RHS3, etcetera. Further, we let $\mathscr{Z} = \{z_1, \ldots, z_n\}$.

A.1. **Proof of theorems 1 and 5.** We focus on $\boldsymbol{T}(\theta_0)$ defined in (17); the derivations for $\boldsymbol{t}(\theta_0)$ defined in (7) can be similarly shown by assuming that $\mathscr{V}_i(\theta_0)$ is known. Omitting the θ_0 argument, write $\boldsymbol{T} = \hat{N}^\intercal \hat{\tilde{D}}^{-2} \hat{N}$, where $\hat{N} = \sum_i \hat{G}_i^\intercal \hat{\mathscr{V}}_i^{-1} m_i$. Let $\hat{N}$ be defined as $\hat{N}$ but with $\mathscr{V}_i$ in lieu of $\hat{\mathscr{V}}_i$. Let $m_i^* = \mathscr{V}_i^{-1} m_i$ and $m_{\theta i}^* = \mathscr{V}_i^{-1} m_{\theta i}$. Let $\rho = \sqrt{n}(\lambda + 1/\sqrt{k})$. Define

$$\tilde{D}^2 = n \operatorname{Var}[G_1^\intercal m_1^*] + \sum_{ij} w_{ij}^2 \mathbb{E}[V_j^\intercal m_i^* m_i^{*\intercal} V_j | z_i, z_j] + \sum_{ij} w_{ij} w_{ji} \mathbb{E}[V_j^\intercal m_i^* m_j^{*\intercal} V_i | z_i, z_j].$$

Let further $B_i = B_i(\theta_0) = \zeta_i + \sum_{j<i} \xi_{ij}$ with $\zeta_i = G_i^\intercal m_i^*$ and $\xi_{ij} = \ell_{ij} + \ell_{ji}$, where $\ell_{ij} = w_{ij} V_j^\intercal m_i^*$. Let $\tilde{N} = \sum_i B_i$, where $\{\tilde{D}^{-1} B_i\}$ forms a martingale difference array due to the fact that $\mathbb{E}[B_i | B_{i-1}, \mathscr{Z}] = 0$ a.s..

(1) We first show that $\tilde{D}^{-1} \hat{N} \xrightarrow{d} N\left(0, I_{d_\theta}\right)$.
  (a) We show that $\|\hat{N} - \tilde{N}\| = o_p(\rho)$, which is established in lemma C8.
  (b) We show that $\|\hat{N} - \tilde{N}\| = o_p(\rho)$
    (i) Since $M_i(\theta_0) = 0$ a.s., lemma C15 establishes that $\hat{N} = \sum_i \zeta_i + \sum_{ij} \ell_{ij} + o_p(\rho)$.
    (ii) The result then follows from the fact that $\sum_i \sum_{j \neq i} \ell_{ij} = \sum_i \sum_{j<i} (\ell_{ij} + \ell_{ji})$.
  (c) We show that $\tilde{D}^{-1} \tilde{N} \xrightarrow{d} N\left(0, I_{d_\theta}\right)$.
    (i) Lemmas D1 to D4 imply that $\forall \|c\| = 1 : c^\intercal \tilde{D}^{-1} \tilde{N} \xrightarrow{d} N(0,1)$.
    (ii) The result then follows from the Cramér–Wold device.
  (d) Finally, note that $\|\tilde{D}^{-1}(\hat{N} - \tilde{N})\| \leq \|\tilde{D}^{-1}\| \|\hat{N} - \tilde{N}\| = O_p(\rho^{-1}) o_p(\rho) = o_p(1)$.
(2) We show that $\tilde{D}^{-1} \hat{\tilde{D}} \xrightarrow{p} I_{d_\theta}$.
  (a) $\hat{\tilde{D}}^2 - \tilde{D}^2 = o_p(\rho^2)$ is established in lemmas C14 and C22.
  (b) Since $\tilde{D}^{-1} = O_p(\rho^{-1})$ by lemma C25, we have $\|\tilde{D}^{-1} \hat{\tilde{D}} - I_{d_\theta}\| = \|\tilde{D}^{-1}(\hat{\tilde{D}} - \tilde{D})\| \leq \|\tilde{D}^{-1}\| \sqrt{\|\hat{\tilde{D}}^2 - \tilde{D}^2\|} = o_p(1)$.[24]
(3) So, the result follows from combining the results established in items (1) and (2).

A.2. **Proof of theorems 2 and 6.** Again, we focus on $\boldsymbol{T}(\theta)$ because $\boldsymbol{t}(\theta)$ can be similarly dealt with by assuming $\mathscr{V}_i(\theta)$ is known. We start with part (i).

(1) We show that $\hat{\tilde{N}}$ diverges at a rate of $n\lambda^2$.
  (a) We show that $\hat{\tilde{N}} = n\lambda^2 \mathbb{E}\left[\tilde{G}_1^\intercal \mathscr{V}_1^{-1} \tilde{M}_1\right] + o_p(\rho) + o_p(n\lambda^2)$, which follows from lemmas C8 and C17, noting that $n\lambda^2 \succ \rho$ by assumption.
  (b) $\mathbb{E}\left[\tilde{G}_1^\intercal \mathscr{V}_1^{-1} \tilde{M}_1\right] \neq 0$ by assumption CG.
(2) We show that $\hat{\tilde{D}}^2 = o_p(n^2\lambda^4)$, which implies that $\hat{\tilde{D}} = o_p(n\lambda^2)$.

[24] $\|\hat{\tilde{D}} - \tilde{D}\|^2 \leq \|\hat{\tilde{D}}^2 - \tilde{D}^2\|$ because if c is an eigenvector of $\hat{\tilde{D}} - \tilde{D}$ with eigenvalue ν then if $\nu > 0$, $c^\intercal(\hat{\tilde{D}} - \tilde{D})^2 c = c^\intercal(\hat{\tilde{D}}^2 - \tilde{D}^2)c - 2\nu c^\intercal \tilde{D} c \leq c^\intercal(\hat{\tilde{D}}^2 - \tilde{D}^2)c$ and if $\nu < 0$ then $c^\intercal(\hat{\tilde{D}} - \tilde{D})^2 c = c^\intercal(\tilde{D}^2 - \hat{\tilde{D}}^2)c + 2\nu c^\intercal \hat{\tilde{D}} c \leq c^\intercal(\tilde{D}^2 - \hat{\tilde{D}}^2)c$.

(a) Recall that $\|\hat{\tilde{D}}^2 - \tilde{D}^2\| = o_p(\rho^2)$ by lemmas C14 and C22.
(b) Since $\|\tilde{D}^2\| = O_p(\rho^2)$ by lemma C24, we have $\|\hat{\tilde{D}}^2\|/n^2\lambda^4 \leq o_p(\rho^2/n^2\lambda^4) + O_p(\rho^2/n^2\lambda^4) = o_p(1)$.

(3) Since $\boldsymbol{T}^{-1} = \left(\hat{\tilde{N}}^\intercal \hat{\tilde{D}}^{-2} \hat{\tilde{N}}\right)^{-1} \leq \left\{\|\hat{\tilde{N}}\|^2 \underline{\mathrm{ev}}(\hat{\tilde{D}}^{-2})\right\}^{-1} = \|\hat{\tilde{D}}^2\|/\|\hat{\tilde{N}}\|^2 = o_p(1)$, (i) hence holds.

(4) Now part (ii). By lemma C23, $\tilde{D}^2 = n \operatorname{Var}[G_1^\intercal \mathscr{V}_1^{-1} m_1] + o_p(n\lambda^2)$, and hence

$$\begin{aligned}\hat{\tilde{D}}^{-1}\hat{\tilde{N}} &= \left(n\lambda^2 \operatorname{Var}[\tilde{G}_1^\intercal \mathscr{V}_1^{-1} m_1] + o_p(n\lambda^2)\right)^{-1/2} \left(n\lambda^2 \mathbb{E}\left[\tilde{G}_1^\intercal \mathscr{V}_1^{-1} \tilde{M}_1\right] + o_p(n\lambda^2)\right) \\ &= \sqrt{n}\lambda \left(\operatorname{Var}[\tilde{G}_1^\intercal \mathscr{V}_1^{-1} m_1]\right)^{-1/2} \mathbb{E}\left[\tilde{G}_1^\intercal \mathscr{V}_1^{-1} \tilde{M}_1\right] + o_p(\sqrt{n}\lambda).\end{aligned}$$

A.3. **Proof of theorem 3.**

(1) We first have by lemmas E5 and E6,

$$\sum_i \hat{g}_i(\theta_n) m_i(\theta_n) = \sum_i \hat{g}_i m_i + \sqrt{n}\lambda \mathbb{E}\tilde{g}_1 \tilde{g}_1^\intercal \Delta + o_p(\rho). \tag{18}$$

(2) Also,

$$\hat{\tilde{D}}_s^2(\theta_n) = \left\{\sum_i \hat{g}_i \hat{g}_i^\intercal m_i^2 - n^{-1}\left(\sum_i \hat{g}_i m_i\right)\left(\sum_i \hat{g}_i^\intercal m_i\right) + \sum_{ij} w_{ij} w_{ji} m_j m_{\theta j}^\intercal m_i m_{\theta i}\right\} + o_p(\rho^2), \tag{19}$$

by lemmas E5 to E8.

(3) Combining the results of (18) and (19) with those of theorem 1, we obtain the stated result.

A.4. **Proof of theorem 4.** Express the test statistic $\hat{\boldsymbol{t}}$ as $\hat{\tilde{\mathscr{D}}}_N^{-1} \hat{\mathscr{N}}_N$ and let $\boldsymbol{t} = \hat{\tilde{\mathscr{D}}}_s^{-1} \hat{\mathscr{N}}_s$, where $\hat{\tilde{\mathscr{D}}}_s, \hat{\mathscr{N}}_s$ are defined as $\hat{\tilde{N}}_s$ and $\hat{\tilde{D}}_s$ with $\hat{g}_i$ replaced with $\hat{q}_i$; then, the results for $\hat{\tilde{N}}_s$ and $\hat{\tilde{D}}_s$ also hold for $\hat{\mathscr{N}}_s$ and $\hat{\tilde{\mathscr{D}}}_s$. We proceed as follows. The θ_0–argument is dropped in the list below.

(1) First asymptotic validity.
(a) Lemma F4 shows that $\hat{\mathscr{N}}_N - \hat{\mathscr{N}}_s = o_p(\rho)$.
(b) Lemma F8 establishes that $\hat{\tilde{\mathscr{D}}}_N^2 - \hat{\tilde{\mathscr{D}}}_s^2 = o_p(\rho^2)$.
(c) Consequently,

$$\left\{(\hat{\tilde{\mathscr{D}}}_N - \hat{\tilde{\mathscr{D}}}_s)/\rho + \hat{\tilde{\mathscr{D}}}_s/\rho\right\}^{-1}\left\{(\hat{\mathscr{N}}_N - \hat{\mathscr{N}}_s)/\rho + \hat{\mathscr{N}}_s/\rho\right\} = \boldsymbol{t} + o_p(1).$$

(d) Apply theorem 1.

(2) For consistency:
(a) Lemma F4 shows that $\hat{\mathscr{N}}_N - \hat{\mathscr{N}}_s = O_p(\rho) + o_p(n\lambda^2)$. Since $n\lambda^2 \succ \rho$ by assumption, $\hat{\mathscr{N}}_N = n\lambda^2 \mathbb{E}[\tilde{q}_i \tilde{M}_1] + o_p(n\lambda^2)$, which follows from applying lemma C17 to $\hat{\mathscr{N}}_s$, replacing g_i with q_i.
(b) Lemma F8 establishes that $\hat{\tilde{\mathscr{D}}}_N^2 - \hat{\tilde{\mathscr{D}}}_s^2 = o_p(\rho^2)$. Further, from the proof of theorem 2 with $\hat{\tilde{\mathscr{D}}}_s$ in lieu of $\hat{\tilde{D}}_s$, it follows that $\hat{\tilde{\mathscr{D}}}_s^2 = o_p(n^2\lambda^4)$. Hence $\hat{\tilde{\mathscr{D}}}_s = o_p(n\lambda^2)$.

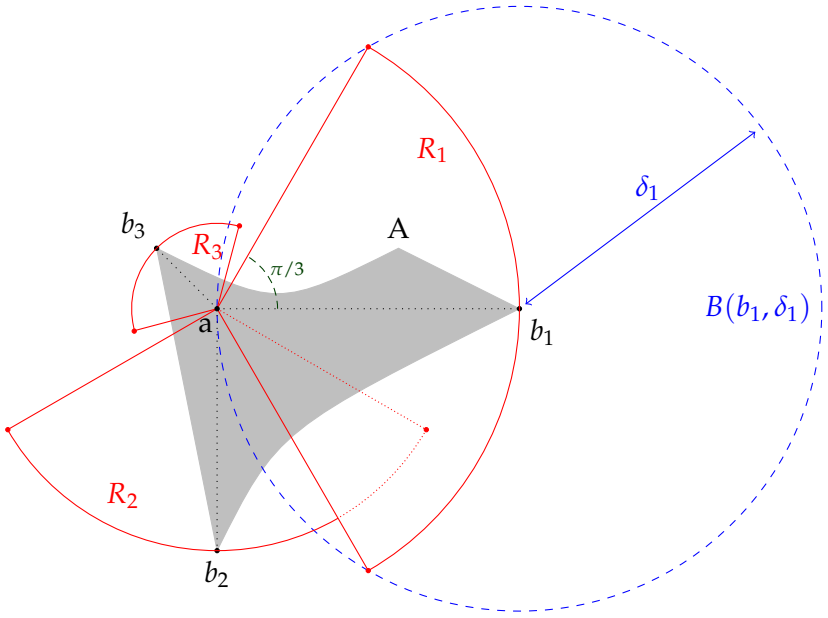

FIGURE 7. Proof of lemma B2

.

APPENDIX B. TECHNICAL LEMMAS

Lemma B1. *For $p_1 > 0$ let $f_i = f(z_i)$ be such that $\mathbb{E}|f_i|^{p_1 p_2} < \infty$ and $a_i > 0$ be such that $Ea_i^{p_2/(p_2-1)} < \infty$ for some $p_2 \geq 1$.*[25] *Then*

$$\mathbb{E}\left[w_{12}a_1|f_1 - f_2|^{p_1}\right] = o(n^{-1}). \tag{20}$$

Proof. Note first that by the Jensen inequality,

$$\left(\sum_j w_{ij}|f_i - f_j|^{p_1}\right)^{p_2} \leq \sum_j w_{ij}|f_i - f_j|^{p_1 p_2}. \tag{21}$$

The expectation of the RHS in (21) is $o(1)$ by lemma 1 of Robinson (1987). Hence the LHS in (20) is by the Hölder inequality bounded by

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

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

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

Proof. We establish the result for $z_i \in \mathbb{R}^2$; the case in which z_i is scalar–valued is simple and the case in which z_i has dimension greater than two is only more complicated in terms of exposition.

Let $A = A(a)$ and $b(t,\omega) = (a_1 + t\cos\omega, a_2 + t\sin\omega)$ and further $\delta_1 = \sup_{\omega,t:b(t,\omega)\in A} t$, ω_1 a corresponding ω and $b_1 = b(\delta_1, \omega_1)$; see figure 7. If $\delta_1 = 0$ there is nothing to prove so suppose that $\delta_1 > 0$. Then let $R_1 = \left\{b = b(t,\omega) : |\omega - \omega_1| < \pi/3, 0 \leq t \leq \delta_1\right\}$. So R_1 is a slice of a pie

[25] $p_2 = 1$ means that a_i is bounded a.s..

with radius δ_1 extending $\pi/3$ in both directions from ω; see again figure 7. By construction for all $|\omega - \omega_1| < \pi/3$ and $t \geq 0$: $b(t,\omega) \in A \Rightarrow b(t,\omega) \in R_1$.

Now let $\delta_2 = \sup_{\omega,t:b(t,\omega)\in A\setminus R_1} t$ and define ω_2, b_2, R_2 accordingly. Repeat this procedure until A is covered by at most 10 R_j–sets. By symmetry it suffices to show that $\mathbb{P}(z_2 \in R_1) \leq 2k/n$. Now let $A_1 = \{b(t,\omega_1) : 0 \leq t < \delta_1\}$. Then if $\mathcal{B}$ denotes a closed ball and $B(b_1,\delta_1) = \bigcup_{b\in A_1} \mathcal{B}(b, ||b-a||)$, then because $A_1 \subset A$,

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

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

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

$$\mathbb{E}[w_{21}w_{31}|z_1 = a] \leq \frac{C_w^2}{k^2}\mathbb{E}\big[I(1 \in \mathcal{N}_2)I(1 \in \mathcal{N}_3)|z_1 = a\big],$$

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

$$\begin{aligned}
&\mathbb{E}\big[I(1 \in \mathcal{N}_2)I(1 \in \mathcal{N}_3)|z_1 = a\big] \\
&\quad = \mathbb{E}\big[I(1 \in \mathcal{N}_2)I(1 \in \mathcal{N}_3)\big\{I(S_2^* \leq c(n-2)/2) + I(S_2^* > c(n-2)/2)\big\}|z_1 = a\big] \\
&\quad \leq \mathbb{P}\big[S_2^* \leq c(n-2)/2\big] + \frac{2k}{c(n-2)}\mathbb{E}\big[I(1 \in \mathcal{N}_3)|z_1 = a\big] \\
&\quad \leq \mathbb{P}\big[S_2^* \leq c(n-2)/2\big] + \frac{2k}{c(n-2)}\mathbb{P}\big[S_3^* \leq c(n-2)/2\big] + \frac{4k^2}{c^2(n-2)^2}. \qquad (22)
\end{aligned}$$

Now by the Hoeffding inequality,

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

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

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

we have

$$\begin{aligned}
\mathbb{E}\big[I(1 \in \mathcal{N}_2)I(1 \in \mathcal{N}_3)|z_1 = a\big] &\leq \mathbb{E}\big[I\big(\tau(a,z_2) \leq 2k/n\big)I\big(\tau(a,z_3) \leq 2k/n\big)|z_1 = a\big] \\
&\quad + \mathbb{P}\big[|S_2(a,z_2) - (n-2)\tau(a,z_2)| > k\big] + \mathbb{P}\big[|S_3(a,z_3) - (n-2)\tau(a,z_3)| > k\big]. \qquad (23)
\end{aligned}$$

RHS1 in (23) is $\sup_a \big[\mathbb{P}\big(z_1 \in A(a)\big)\big]^2 = O(k^2/n^2)$ by lemma B2. RHS2 and RHS3 in (23) are the same and are by the Hoeffding inequality bounded by $\exp\big(-2k^2/(n-2)\big)$, which goes to zero at an exponential rate. □

Lemma B4. $\max_j \sum_i w_{ij} = O_p(1)$.

Proof. Since the proof is similar to that of lemma B3, we only establish the result for the case in which the probability of ties is zero. Let $S_{ij}(a,b) = \sum_{t\neq i,j} I(\|z_t - b\| \leq \|a - b\|)$. Now, for any $0 < C^* < \infty$ and $C = C/3C_w$,

$$\begin{aligned}
&\mathbb{P}\Big[\max_j \sum_i w_{ij} > C^*\Big] \leq \mathbb{P}\Big[\max_j \sum_{i\neq j} I(j \in \mathcal{N}_i) > 3Ck\Big] \\
&\overset{(22)}{\leq} \mathbb{P}\Big[\max_j \sum_{i\neq j} I\big(\tau(z_j, z_i) \leq 2k/n\big) > 2Ck\Big] + \mathbb{P}\Big[\max_j \sum_{i\neq j} I(|S_{ij}(z_j, z_i) - (n-2)\tau(z_j, z_i)| > k) > Ck\Big] \\
&\leq n \sup_a \mathbb{P}\Big[\Big|\sum_i \big\{I\big(\tau(a, z_i) \leq 2k/n\big) - \mathbb{P}[\tau(a, z_i) \leq 2k/n]\big\}\Big| > Ck\Big] + \sup_a I\big(\mathbb{P}[\tau(a, z_1) \leq 2k/n] > Ck/n\big) \\
&\qquad + n^2 \sup_{a,b} \mathbb{P}\Big[\Big|\sum_t \big(I(\|z_t - b\| \leq \|a - b\|) - \tau(a, b)\big)\Big| > k\Big]. \qquad (24)
\end{aligned}$$

RHS1 and RHS3 in (24) are by theorem 1.1 of de la Peña (1999) respectively bounded by $2n\exp\big(-C^2k/(4+2C)\big) = o(1)$ and $2n^2\exp\big(-k^2/4n\big) = o(1)$. For RHS2, let $n \to \infty$ followed by $C \to \infty$. □

Lemma B5. *Suppose that for some $p \geq 1$, $\mathbb{E}|\mathbb{E}[b_1|z_1]|^p < \infty$ and $\mathbb{E}|\mathbb{E}[a_1|z_1]|^{p/(p-1)} < \infty$. Then*
(i) $\mathbb{E}[w_{12}a_1b_2] = n^{-1}\mathbb{E}\big[a_1\mathbb{E}[b_1|z_1]\big] + o(n^{-1}) = O(n^{-1})$,
(ii) for any $1 < p^* < \infty$: $\mathbb{E}[w_{12}^{p^*}a_1b_2) = O(n^{-1}k^{1-p^*}]$.
If $\mathbb{E}\big|\mathbb{E}[a_1|z_1]\big|^{2p} < \infty$, $\mathbb{E}\big|\mathbb{E}[b_1|z_1]|^{2p} < \infty$ and $\mathbb{E}\big|\mathbb{E}[c_1|z_1]\big|^{p/(p-1)} < \infty$ for some $p \geq 1$ then
(iii) $\mathbb{E}[w_{12}w_{13}c_1a_2b_3] = O(n^{-2})$,
(iv) $\sum_t \mathbb{E}[w_{12}w_{1t}c_1a_2b_t] = O(n^{-1})$.
Further, if for some $p_1 > 4$ and $p_2 \geq 1$, $\mathbb{E}\big|\mathbb{E}[a_1|z_1]\big|^{p_1p_2} + \mathbb{E}\big|\mathbb{E}[b_1|z_1]\big|^{p_1p_2/(p_2-1)} + \mathbb{E}|c_1| < \infty$, then
(v) $\mathbb{E}[w_{21}w_{31}c_1a_2b_3] = o(n^{-3/2}k^{-1/2})$, and
(vi) $\sum_t \mathbb{E}[w_{21}w_{t1}c_1a_2b_t] = o(1/\sqrt{nk})$.

Proof. First (i). Note that

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

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

$$\big|\mathbb{E}[w_{12}w_{13}c_1a_2b_3]\big| \leq \mathbb{E}\big[w_{12}w_{13}|c_1|(a_2^2 + b_3^2)\big] = n^{-1}\mathbb{E}\big[w_{12}|c_1|a_2^2\big] + n^{-1}\mathbb{E}\big[w_{12}|c_1|b_2^2\big] = O(n^{-2}),$$

by (i). For (iv) note that

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

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

$$\begin{aligned}\left|\mathbb{E}[w_{21}w_{31}c_1a_2b_3]\right| &= \left|\mathbb{E}\left[w_{21}w_{31}c_1\tilde{a}_2\tilde{b}_3\right]\right| = \left|\mathbb{E}\left\{\mathbb{E}\left[w_{21}w_{31}\tilde{a}_2\tilde{b}_3|z_1\right]c_1\right\}\right| \\ &\overset{\text{Hölder}}{\leq} \mathbb{E}\left\{\left(\mathbb{E}\left[(w_{21}w_{31})^{p_1/(p_1-1)}|z_1\right]\right)^{(p_1-1)/p_1}c_1\right\}\left(\mathbb{E}|\tilde{a}_2\tilde{b}_3|_1^p\right)^{1/p_1} \\ &\overset{\text{B3,Hölder}}{\leq} C_w^2n^{-2}(n/k)^{2/p_1}\mathbb{E}|c_1|\left(\mathbb{E}|\tilde{a}_2|^{p_1p_2}\right)^{1/(p_1p_2)}\left(\mathbb{E}|\tilde{b}_3|^{p_1p_2/(p_2-1)}\right)^{(p_2-1)/(p_1p_2)} \prec n^{-3/2}k^{-1/2}.\end{aligned}$$

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

Lemma B6. *Let a_i, b_i be such that for some $p \geq 1$, $\mathbb{E}|\mathbb{E}[b_1^2|z_1]|^p < \infty$ and $\mathbb{E}|\mathbb{E}[a_1^2|z_1]|^{p/(p-1)} < \infty$ and let p_1, p_2 be such that $\min(p_1, p_2) \geq 0$ and $\max(p_1, p_2) \geq 1$. Let furthermore $|c_{ij}| \leq 1$ be constants. Then*
(i) If $\mathbb{E}[a_i|z_i] = \mathbb{E}[b_i|z_i] = 0$ a.s. then $\sum_{ij} w_{ij}^{p_1}w_{ji}^{p_2}c_{ij}a_ib_j = O_p\left(n^{1/2}k^{1/2-p_1-p_2}\right)$,
(ii) If $\mathbb{E}[a_i|z_i] = 0$ a.s. and $p_1 \geq 1$ then $\sum_{ij} w_{ij}^{p_1}w_{ji}^{p_2}c_{ij}a_ib_j = O_p\left(n^{1/2}k^{1-p_1-p_2}\right)$,
(iii) if $\max(p_1, p_2) \geq 1$ then $\sum_{ij} w_{ij}^{p_1}w_{ji}^{p_2}c_{ij}\left[a_ib_j - \mathbb{E}[a_i|z_i]\mathbb{E}[b_j|z_j]\right] = O_p\left(n^{1/2}k^{1-p_1-p_2}\right)$,
If furthermore, for some $p_3 > 8$, $\mathbb{E}|\mathbb{E}[b_1|z_1]|^{p_3} < \infty$ then
(iv) if $\mathbb{E}[a_i|z_i] = 0$ a.s. then $\sum_{ij} w_{ji}^{p_2}c_{ij}a_ib_j = o_p\left(n^{3/4}k^{3/4-p_2}\right)$,
(v) $\sum_{ij} w_{ji}^{p_2}c_{ij}\left[a_ib_j - \mathbb{E}[a_i|z_i]\mathbb{E}[b_j|z_j]\right] = o_p\left(n^{3/4}k^{3/4-p_2}\right)$,
(vi) $\sum_{ij} w_{ij}^{p_1}w_{ji}^{p_2}a_ib_j = \sum_{ij} w_{ij}^{p_1}w_{ji}^{p_2}\mathbb{E}[a_i|z_i]\mathbb{E}[b_i|z_i] + o_p(nk^{1-p_1-p_2})$.

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

$$\begin{aligned}&\sum_{ij} c_{ij}\mathbb{E}\left[w_{ij}^{2p_1}w_{ji}^{2p_2}a_i^2b_j^2 + (w_{ij}w_{ji})^{p_1+p_2}a_ib_ia_jb_j\right] \\ &\qquad\leq (C_w/k)^{2p_1+2p_2-1}\sum_{ij} w_{ij}\left(a_i^2b_j^2 + |a_ib_ia_jb_j|\right) \overset{\text{B5}}{=} O\left(nk^{1-2p_1-2p_2}\right).\end{aligned}$$

For (ii) likewise square and take expectations which yields

$$\sum_{ijt} c_{ij}c_{it}\mathbb{E}\left[(w_{ij}w_{it})^{p_1}(w_{ji}w_{ti})^{p_2}a_i^2b_jb_t\right] \leq (C_w/k)^{2p_1+2p_2-2}\sum_{ijt}\mathbb{E}[w_{ij}w_{it}a_i^2b_jb_t] \overset{\text{B5}}{=} O\left(nk^{2(1-p_1-p_2)}\right).$$

For (iii) write $a_ib_j - \mathbb{E}[a_i|z_i]\mathbb{E}[b_j|z_j] = \left(a_i - \mathbb{E}[a_i|z_i]\right)\left(b_i - \mathbb{E}[b_i|z_i]\right) + \mathbb{E}[a_i|z_i]\left(b_i - \mathbb{E}[b_i|z_i]\right) + \left(a_i - \mathbb{E}[a_i|z_i]\right)\mathbb{E}[b_i|z_i]$ and apply parts (i) and (ii). For (iv) we get

$$\sum_{ijt} c_{ij}c_{it}\mathbb{E}\left[(w_{ji}w_{ti})^{p_2}a_i^2b_jb_t\right] \leq (C_w/k)^{2p_2-2}\sum_{ijt}\mathbb{E}[w_{ji}w_{ti}a_i^2b_jb_t] \overset{\text{B5}}{=} o\left(n^{3/2}k^{3/2-2p_2}\right).$$

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

$$\Big|\sum_{ij} w_{ij}^{p_1} w_{ji}^{p_2} \mathbb{E}[a_i|b_i]\big[\mathbb{E}[b_j|z_j] - \mathbb{E}[b_i|z_i]\big]\Big|$$
$$\leq n^2 (C_w/k)^{p_1+p_2-1} \mathbb{E}\big|w_{12}\mathbb{E}[a_1|z_1]\big(\mathbb{E}[b_2|z_2] - \mathbb{E}[b_1|z_1]\big)\big| \overset{\text{B1}}{=} o\big(nk^{1-p_1-p_2}\big). \qquad \square$$

Lemma B7. *Let $\{b_j\}$ be i.i.d. conditional on $\mathscr{Z}$ with $\mathbb{E}[b_j|\mathscr{Z}] = 0$ a.s., and let $c_{ij} = c(z_i, z_j, \mathscr{Z})$. Let further $\mathscr{S}_{ni} = \sum_j c_{ij} b_j$ and $\mathscr{Q}_{ni} = \sum_j c_{ij}^2 \mathbb{E}[b_j^2|\mathscr{Z}]$. Then if $\mathbb{E}|b_j|^p < \infty$ for some $2 \leq p < \infty$ and $\max_i \mathscr{Q}_{ni} = O_p(C_n^*)$, then for $C_n = (\sqrt{C_n^*} + n^{1/p}/k)(\log n)^2$, $\max_i |\mathscr{S}_{ni}| = o_p(C_n)$.*

Proof. First, for some $0 < C, C^* < \infty$,

$$\mathbb{P}\big[\max_i |\mathscr{S}_{ni}| > C_n\big] \leq \sum_i \mathbb{P}\big[|\mathscr{S}_{ni}| > C_n,\ \max_i \mathscr{Q}_{ni} \leq C^* C_n^*,\ \max_j |b_j| \leq C n^{1/p}\big]$$
$$+ \mathbb{P}\big[\max_i \mathscr{Q}_{ni} > C^* C_n^*\big] + \sum_i \mathbb{P}\big[|b_i| > C n^{1/p}\big]. \quad (25)$$

RHS1 in (25) is by theorem 1.1 of de la Peña (1999) bounded by $2n \exp\big(-C_n^2/2(C^* C_n^* + C n^{1/p} C_n/k)\big) = o(1)$ by the definition of C_n. RHS2 is also $o(1)$, which follows by taking the limit for $n \to \infty$ followed by $C^* \to \infty$. Finally, RHS3 is by the Markov inequality bounded by $\mathbb{E}|b_i|^p/C^p$. Take $C \to \infty$. $\square$

Lemma B8. *If $\{b_j\}$ is i.i.d. and $\mathbb{E}|b_j|^p < \infty$ for some $1 \leq p < \infty$ then $\max_i \big|\sum_j w_{ij} b_j\big| = O_p\big((n/k)^{1/p}\big)$.*

Proof. By the Hölder inequality for any i and some $C < \infty$ independent of i,
$\big|\sum_j w_{ij} b_j\big| \leq \big(\sum_j w_{ij}^{p/(p-1)}\big)^{1-1/p} \big(\sum_j |b_j|^p\big)^{1/p} \leq C\big(\sum_j |b_j|^p\big)^{1/p}/k^{1/p}$. $\square$

Lemma B9. *Let $\{b_j\}$ be such that $\{(b_j, z_j)\}$ is i.i.d. with $\mathbb{E}[b_j|z_j] = 0$ a.s. and $\mathbb{E}|b_j|^p < \infty$ for some $2 \leq p < \infty$ and $\mathbb{E}\big[\mathbb{E}[b_j^2|z_j]\big]^{p^*} < \infty$ for some $1 \leq p^* < \infty$. Then $\max_i \big|\sum_j w_{ij} b_j\big| = o_p\big((n^{1/p}/k + n^{1/2p^*}/k^{(p^*+1)/2p^*})(\log n)^2\big)$.*

Proof. Take $c_{ij} = w_{ij}$ in B7 and note that $\max_i \mathscr{Q}_{ni} = O_p(n^{1/p^*}/k^{1+1/p^*})$ by B8. $\square$

Lemma B10. *For an arbitrary random sequence $\{b_n\}$, if $\mathbb{E}[|b_n|\big|\mathscr{Z}] = O_p(1)$, then $b_n = O_p(1)$.*

Proof. Choose an arbitrary $\epsilon > 0$. We show that for some $C_1 > 0$, $\sup_n \mathbb{P}[|b_n| > C_1] < \epsilon$. Since $a_n = \mathbb{E}[|b_n|\big|\mathscr{Z}] = O_p(1)$, there exists a $C_2 > 0$ such that $\sup_n \mathbb{P}[a_n > C_2] < \epsilon/2$. For this value of C_2, choose $C_1 > 0$ such that $C_2/C_1 \leq \epsilon/2$. We then have

$$\sup_n \mathbb{P}[|b_n| > C_1] = \sup_n \int \mathbb{P}[|b_n| > C_1\big|\mathscr{Z}]\big(I(a_n > C_2) + I(a_n \leq C_2)\big) dP_{\mathscr{Z}}$$
$$\leq \sup_n \mathbb{P}[a_n > C_2] + \frac{1}{C_1} \sup_n \int a_n I(a_n \leq C_2) dP_{\mathscr{Z}} \leq \frac{\epsilon}{2} + \frac{C_2}{C_1} \leq \epsilon. \quad \square$$

Appendix C. Approximations

Suppressing θ, let $u_i = m_i - M_i$, $u_i^* = \mathscr{V}_i^{-1} u_i$, $G_i^* = \mathscr{V}_i^{-1} m_{\theta i}$, and $V_i^* = m_{\theta i}^* - G_i^* = \mathscr{V}_i^{-1} V_i$.

C.1. $\hat{\mathscr{V}}_i$ **and** $\mathscr{V}_i$**.** Let $\hat{u}_i(\theta) = \sum_j w_{ij} u_j(\theta)$, $\hat{M}_i(\theta) = \sum_j w_{ij} M_j(\theta)$, $\check{\mathscr{V}}_i(\theta) = \sum_j w_{ij}\big(M_j(\theta) - \hat{M}_j(\theta)\big)\big(M_j(\theta) - \hat{M}_j(\theta)\big)^\intercal + \sum_j w_{ij}\mathscr{V}_j(\theta)$. In the lemmas below we will frequently drop the argument θ when this can be done without losing clarity.

Lemma C1. $\exists \epsilon > 0 : \mathbb{P}\big[\max_i \|\check{\mathscr{V}}_i^{-1}\| > 1/\epsilon\big] = 0$.

Proof. For sufficiently small $\epsilon > 0$, $\|\check{\mathscr{V}}_i^{-1}\| \leq \big\|\big(\sum_j w_{ij}\mathscr{V}_j\big)^{-1}\big\| \leq 1/\underline{\mathrm{ev}}(\mathscr{V}_i) < 1/\epsilon$ a.s. by assumption A. □

Lemma C2. $\max_i \|\hat{\mathscr{V}}_i - \check{\mathscr{V}}_i\| = o_p\big((\log n)^2 n^{1/2}/k\big)$.

Proof. Note that

$$\hat{\mathscr{V}}_i - \check{\mathscr{V}}_i = \sum_j w_{ij}(u_j u_j^\intercal - \mathscr{V}_j) - \sum_j w_{ij}(u_j \hat{u}_j^\intercal + \hat{u}_j u_j^\intercal) + \sum_j w_{ij}\hat{u}_j\hat{u}_j^\intercal + \sum_j w_{ij}\big((u_j - \hat{u}_j)(M_j - \hat{M}_j)^\intercal + (M_j - \hat{M}_j)(u_j - \hat{u}_j)^\intercal\big). \quad (26)$$

We show that the desired uniform rate holds for RHS1 and RHS2 in (26), where similar arguments can be used for RHS3. For RHS1, take $p = 2, p^* = 4$ in B9. For RHS2, note that

$$\max_i \Big\|\sum_j w_{ij}(u_j\hat{u}_j^\intercal + \hat{u}_j u_j^\intercal)\Big\| \leq 2\max_i |\hat{u}_i| \max_i \sum_j w_{ij}|u_j| = o_p\big((n^{1/4}/k)(n/k)^{1/4}(\log n)^2\big) = o_p\big(n^{1/2}k^{-5/4}(\log n)^2\big),$$

where the penultimate equality follows by taking $p = 4, p^* = 8$ in B8 and B9. □

Lemma C3. $\max_i \|\hat{\mathscr{V}}_i^{-1}\| = O_p(1)$.

Proof. Follows from lemmas C1 and C2 since $\underline{\mathrm{ev}}(\hat{\mathscr{V}}_i) \geq \underline{\mathrm{ev}}(\check{\mathscr{V}}_i) - \|\hat{\mathscr{V}}_i - \check{\mathscr{V}}_i\|$. □

Lemma C4. *(i) For* $1 \leq p \leq 2$, $\sum_i \|\hat{\mathscr{V}}_i - \check{\mathscr{V}}_i\|^p = O_p(n/k^{p/2})$ *and (ii) for* $p > 2$, $\sum_i \|\hat{\mathscr{V}}_i - \check{\mathscr{V}}_i\|^p = o_p\big((\log n)^{2p-4} n^{p/2}/k^{p-1}\big)$.

Proof. For (i) each RHS term in the expansion (26) can be dealt with separately. We only work on the dominant term, i.e. RHS1, where the other terms follow similarly. We have for some $C < \infty$,

$$\sum_i \mathbb{E}\Big\|\sum_j w_{ij}(u_j u_j^\intercal - \mathscr{V}_j)\Big\|^p \overset{\text{Burkholder}}{\leq}{}^{26} C\sum_i \Big(\sum_j \mathbb{E}[w_{ij}^2\|u_j\|^4]\Big)^{p/2} \leq \frac{C^2}{k^{p/2}}\sum_i \Big(\sum_j \mathbb{E}[w_{ij}\|u_j\|^4]\Big)^{p/2} = O\Big(\frac{n}{k^{p/2}}\Big).$$

(ii) then follows from (i) and C3. □

Lemma C5. $\sum_i \hat{G}_i^\intercal(\hat{\mathscr{V}}_i^{-1} - \check{\mathscr{V}}_i^{-1})m_i = \begin{cases} o_p(\rho), & \theta = \theta_0, \\ o_p(\rho + n\lambda^2), & \theta \neq \theta_0. \end{cases}$.

[26] See theorem 15.18 of Davidson (1994) and comments following it.

Proof. It suffices to show that

$$\sum_{ij} w_{ij} V_j^\intercal (\hat{\mathscr{V}}_i^{-1} - \check{\mathscr{V}}_i^{-1}) u_i = o_p(\rho), \tag{27}$$

$$\sum_{ij} w_{ij} G_j^\intercal (\hat{\mathscr{V}}_i^{-1} - \check{\mathscr{V}}_i^{-1}) u_i = o_p(\rho), \tag{28}$$

$$\sum_{ij} w_{ij} V_j^\intercal (\hat{\mathscr{V}}_i^{-1} - \check{\mathscr{V}}_i^{-1}) M_i = o_p(\rho), \tag{29}$$

$$\sum_{ij} w_{ij} G_j^\intercal (\hat{\mathscr{V}}_i^{-1} - \check{\mathscr{V}}_i^{-1}) M_i = o_p(n\lambda^2). \tag{30}$$

The LHS in (30) is bounded by $\lambda^2 \max_i \|\hat{\mathscr{V}}_i^{-1} - \check{\mathscr{V}}_i^{-1}\| \sum_{ij} w_{ij} \|\tilde{G}_j\| \|\tilde{M}_i\| = o_p(n\lambda^2)$ by lemmas C1, C2, and C3. The LHS of (29) is by the Schwarz inequality plus lemmas C1, C3 bounded by

$$O_p(\lambda) \sqrt{\sum_i \Big\| \sum_j w_{ij} V_j \Big\|^2 \|\tilde{M}_i\|^2 \sum_i \|\hat{\mathscr{V}}_i - \check{\mathscr{V}}_i\|^2} = O_p\big(\lambda \sqrt{n/k} \sqrt{n/k}\big) = O_p(\rho \sqrt{n}/k) = o_p(\rho),$$

where the first equality follows from lemma C4 and from taking expectations of the first factor under the square root. Letting $\hat{\tilde{G}}_i = \sum_j w_{ij} \tilde{G}_j$, the LHS in (28) equals

$$-\lambda \sum_i \hat{\tilde{G}}_i^\intercal \check{\mathscr{V}}_i^{-1} (\hat{\mathscr{V}}_i - \check{\mathscr{V}}_i) \check{\mathscr{V}}_i^{-1} u_i + \lambda \sum_i \hat{\tilde{G}}_i^\intercal \hat{\mathscr{V}}_i^{-1} (\hat{\mathscr{V}}_i - \check{\mathscr{V}}_i) \check{\mathscr{V}}_i^{-1} (\hat{\mathscr{V}}_i - \check{\mathscr{V}}_i) \check{\mathscr{V}}_i^{-1} u_i, \tag{31}$$

The second term in (31) is by C1, C3 bounded in norm by a constant times

$$\lambda \sum_i \|\hat{\tilde{G}}_i\| \|u_i\| \|\hat{\mathscr{V}}_i - \check{\mathscr{V}}_i\|^2 \overset{\text{Hölder}}{\leq} \lambda \Big(\sum_i (\|\hat{\tilde{G}}_i\| \|u_i\|)^{p/(p-1)} \Big)^{1-1/p} \Big(\sum_i \|\hat{\mathscr{V}}_i - \check{\mathscr{V}}_i\|^{2p} \Big)^{1/p}$$
$$\overset{\text{C4}}{=} o_p\big(\lambda n^{1-1/p} (\log n)^{4-4/p} n / k^{2-1/p}\big) = o_p\big(\rho n^{3/2-1/p} (\log n)^{4-4/p} / k^{2-1/p}\big).$$

Take $p = 4/3$. A bound for the norm of the first term in (31) can be found by expanding $\hat{\mathscr{V}}_i - \check{\mathscr{V}}_i$ using (26). For the first term in the expansion (26), we get

$$\lambda^2 \mathbb{E} \Big\| \sum_{ij} w_{ij} \hat{\tilde{G}}_i^\intercal \check{\mathscr{V}}_i^{-1} (u_j u_j^\intercal - \mathscr{V}_j) \check{\mathscr{V}}_i^{-1} u_i \Big\|^2 \overset{\text{C1}}{\leq} C\lambda^2 \sum_{ij} \mathbb{E}\Big[w_{ij}^2 \|\hat{\tilde{G}}_i\|^2 \|u_j\|^4 \|u_i\|^2 \Big] = O(n\lambda^2/k) = o(\rho^2).$$

The other expansion terms can be dealt with similarly. Finally, the LHS in (27) equals

$$-\sum_{ij} w_{ij} V_j^\intercal \check{\mathscr{V}}_i^{-1} (\hat{\mathscr{V}}_i - \check{\mathscr{V}}_i) \check{\mathscr{V}}_i^{-1} u_i + \sum_{ij} w_{ij} V_j^\intercal \hat{\mathscr{V}}_i^{-1} (\hat{\mathscr{V}}_i - \check{\mathscr{V}}_i) \check{\mathscr{V}}_i^{-1} (\hat{\mathscr{V}}_i - \check{\mathscr{V}}_i) \check{\mathscr{V}}_i^{-1} u_i. \tag{32}$$

The second term in (32) is by the Hölder inequality bounded in norm by a constant times

$$\Big\{ \sum_i \Big(\Big\| \sum_j w_{ij} V_j \Big\| \|u_i\| \Big)^{p/(p-1)} \Big\}^{1-1/p} \Big\{ \sum_i \|\hat{\mathscr{V}}_i - \check{\mathscr{V}}_i\|^{2p} \Big\}^{1/p}$$
$$= o_p\big((n/k)^{1-1/p} (n/k^{2-1/p}) (\log n)^{4-4/p}\big) = o_p\big(n^{2-1/p} (\log n)^{4-4/p} / k^{3-2/p}\big) = o_p(\rho),$$

for sufficiently large p, where the first equality follows with a Burkholder type argument like the one used earlier and lemma C4. For the first term in (32) we can once again use the expansion (26)

and deal with each RHS term in the expansion separately. For RHS1 in (26) we get

$$\mathbb{E}\Big\|\sum_{ijt} w_{ij}w_{it}V_j^\intercal \check{\mathscr{V}}_i^{-1}(u_tu_t^\intercal - \mathscr{V}_t)\check{\mathscr{V}}_i^{-1}u_i\Big\|$$
$$\leq \mathbb{E}\Big\|\sum_{ij} w_{ij}^2 V_j^\intercal \check{\mathscr{V}}_i^{-1}(u_ju_j^\intercal - \mathscr{V}_j)\check{\mathscr{V}}_i^{-1}u_i\Big\| + \mathbb{E}\Big\|\sum_{ij;t\neq j} w_{ij}w_{it}V_j^\intercal \check{\mathscr{V}}_i^{-1}(u_tu_t^\intercal - \mathscr{V}_t)\check{\mathscr{V}}_i^{-1}u_i\Big\|. \quad (33)$$

By the Schwarz inequality, the square of RHS2 in (33) is bounded by a constant times $\sum_{ij;t\neq j}\mathbb{E}\left[w_{ij}^2w_{it}^2\|V_j\|^2\|u_t\|^4\|u_i\|^2\right] = O(n/k^2) = o(\rho^2)$. For RHS1 in (33) note that for $p = 4/3$, some $C < \infty$, and $\mathscr{R}_{ni} = \sum_{j<i} w_{ij}^2 V_j^\intercal \check{\mathscr{V}}_i^{-1}(u_ju_j^\intercal - \mathscr{V}_j)\check{\mathscr{V}}_i^{-1}$ (the $\sum_{j>i}$ case is analogous), by the Burkholder and Loève inequalities,

$$\mathbb{E}\Big\|\sum_i \mathscr{R}_{ni}u_i\Big\|^p \leq C\mathbb{E}\Big[\sum_i\|\mathscr{R}_{ni}u_i\|^2\Big]^{p/2} \leq C^2\sum_i\mathbb{E}\|\mathscr{R}_{ni}u_i\|^p \leq \frac{C^3}{k^p}\sum_i\sum_{j<i}\mathbb{E}\left[w_{ij}\|V_j\|^p\|u_j\|^{2p}\|u_i\|^p\right]$$
$$\overset{\text{Hölder}}{\leq} \frac{C^3}{k^p}\sum_i\sum_{j<i}\mathbb{E}\Big[w_{ij}\big(\mathbb{E}[\|V_j\|^4|z_j]\big)^{1/3}\big(\mathbb{E}[\|u_j\|^4|z_j]\big)^{2/3}\mathbb{E}[\|u_i\|^{4/3}|z_i]\Big] = O(n/k^p) = o(\rho^p). \quad \square$$

Lemma C6. *For any $p \geq 1$, $\sum_i\|\check{\mathscr{V}}_i^{-1} - \mathscr{V}_i^{-1}\|^p = o_p(n)$.*

Proof. Since $\|\check{\mathscr{V}}_i^{-1} - \mathscr{V}_i^{-1}\|^p = \|\check{\mathscr{V}}_i^{-1} - \mathscr{V}_i^{-1}\|^{p-1}\|\check{\mathscr{V}}_i^{-1}(\check{\mathscr{V}}_i - \mathscr{V}_i)\mathscr{V}_i^{-1}\|$ it suffices by lemma C1 to establish the result for $p = 1$. For some $C < \infty$,

$$\frac{1}{C}\sum_i\|\check{\mathscr{V}}_i^{-1} - \mathscr{V}_i^{-1}\| \leq \sum_i\|\check{\mathscr{V}}_i - \mathscr{V}_i\| \leq \sum_{ij}w_{ij}\|\mathscr{V}_j - \mathscr{V}_i\| + \sum_{ij}w_{ij}\|M_j - \hat{M}_j\|^2$$
$$\leq o_p(n) + \sum_{ijt}w_{ij}w_{jt}\|M_j - M_t\|^2 = o_p(n),$$

by lemmas B1 and B4. $\square$

Lemma C7. $\sum_i\hat{G}_i^\intercal(\check{\mathscr{V}}_i^{-1} - \mathscr{V}_i^{-1})m_i = \begin{cases} o_p(\rho), & \theta = \theta_0, \\ o_p(\rho + n\lambda^2), & \theta \neq \theta_0. \end{cases}$

Proof. It suffices to show that

$$\sum_{ij}w_{ij}V_j^\intercal(\check{\mathscr{V}}_i^{-1} - \mathscr{V}_i^{-1})u_i = o_p(\rho), \quad (34)$$
$$\sum_{ij}w_{ij}G_j^\intercal(\check{\mathscr{V}}_i^{-1} - \mathscr{V}_i^{-1})u_i = o_p(\rho), \quad (35)$$
$$\sum_{ij}w_{ij}V_j^\intercal(\check{\mathscr{V}}_i^{-1} - \mathscr{V}_i^{-1})M_i = o_p(\rho), \quad (36)$$
$$\sum_{ij}w_{ij}G_j^\intercal(\check{\mathscr{V}}_i^{-1} - \mathscr{V}_i^{-1})M_i = o_p(n\lambda^2). \quad (37)$$

The LHS in (37) is by the Schwarz inequality and lemma C1 bounded in norm by $\lambda^2\sqrt{\sum_i\|\hat{\tilde{G}}_i\|^2\|\tilde{M}_i\|^2\sum_i\|\check{\mathscr{V}}_i^{-1} - \mathscr{V}_i^{-1}\|} = o_p(n\lambda^2)$ by lemma C6. The expectation of the norm of the

LHS in (36) is by repeated application of the Schwarz inequality bounded by

$$\lambda\Big\{\sum_i\Big(\sum_{jt} w_{ij}w_{tj}\|V_j\|^2\|\tilde{M}_i\|^2\Big)^2\sum_i\|\check{\mathscr{V}}_i^{-1}-\mathscr{V}_i^{-1}\|^4\Big\}^{1/4}=o_p(\sqrt{n}\lambda)=o_p(\rho),$$

by lemmas B4 and C6. The derivation for (34) and (35) is similar to that for (36). □

Lemma C8. $\sum_i \hat{G}_i^\intercal(\hat{\mathscr{V}}_i^{-1}-\mathscr{V}_i^{-1})m_i=\begin{cases} o_p(\rho), & \theta=\theta_0,\\ o_p(\rho+n\lambda^2), & \theta\neq\theta_0.\end{cases}$.

Proof. Combine lemmas C5 and C7. □

Let $\hat{\hat{\alpha}}_i=\hat{G}_i^\intercal\hat{\mathscr{V}}_i^{-1}m_i$, $\hat{\alpha}_i=\hat{G}_i^\intercal\mathscr{V}_i^{-1}m_i$.

Lemma C9. $n^{-1}\big\|\big(\sum_i\hat{\hat{\alpha}}_i\big)\big(\sum_i\hat{\hat{\alpha}}_i^\intercal\big)-\big(\sum_i\hat{\alpha}_i\big)\big(\sum_i\hat{\alpha}_i^\intercal\big)\big\|=o_p(\rho^2)$.

Proof. The LHS is bounded by $n^{-1}\big\|\sum_i(\hat{\hat{\alpha}}_i-\hat{\alpha}_i)\big\|^2+2n^{-1}\big\|\sum_i(\hat{\hat{\alpha}}_i-\hat{\alpha}_i)\big\|\big\|\sum_i\hat{\alpha}_i\big\|=o_p\big(n^{-1}\rho^2+n\lambda^4\big)+o_p(\rho+n\lambda^4)=o_p(\rho^2)$ by lemma C8. □

Lemma C10. $\sum_i\|\hat{\alpha}_i\|^2=O_p(\rho^2)$.

Proof. The LHS is bounded by a constant times $\lambda^2\sum_i\big\|\hat{\hat{G}}_i^\intercal\mathscr{V}_i^{-1}m_i\big\|^2+\sum_i\big\|\sum_j w_{ij}V_j^\intercal\mathscr{V}_i^{-1}m_i\big\|^2=O_p(n\lambda^2+n/k)=O_p(\rho^2)$. □

Lemma C11. $\sum_i\|\hat{\hat{\alpha}}_i-\hat{\alpha}_i\|^2=o_p(\rho^2)$.

Proof. The LHS is bounded by a constant times

$$\lambda^2\sum_i\big\|\hat{\hat{G}}_i^\intercal(\hat{\mathscr{V}}_i^{-1}-\mathscr{V}_i^{-1})m_i\big\|^2+\sum_i\Big\|\sum_j w_{ij}V_j^\intercal(\hat{\mathscr{V}}_i^{-1}-\check{\mathscr{V}}_i^{-1})m_i\Big\|^2+\sum_i\Big\|\sum_j w_{ij}V_j^\intercal(\check{\mathscr{V}}_i^{-1}-\mathscr{V}_i^{-1})m_i\Big\|^2. \tag{38}$$

The first term in (38) is bounded by $\lambda^2\sqrt{\sum_i\|\hat{\hat{G}}_i\|^2\|m_i\|^2\sum_i\|\hat{\mathscr{V}}_i^{-1}-\mathscr{V}_i^{-1}\|^2}=o_p(n\lambda^2)$ by lemmas C1, C4, C6. The expectation of RHS3 conditional on $\mathscr{Z}$ is bounded by

$$\begin{aligned}&\sum_{ij}w_{ij}^2\mathbb{E}[\|V_j\|^2|z_j]\mathbb{E}[\|m_i\|^2|z_i]\|\check{\mathscr{V}}_i^{-1}-\mathscr{V}_i^{-1}\|^2]\\ &\overset{\text{Hölder,Jensen}}{\leq}\frac{C_w}{k}\Big(\sum_{ij}w_{ij}\big(\mathbb{E}[\|V_j\|^2|z_j]\mathbb{E}[\|m_i\|^2|z_i]\big)^p\Big)^{1/p}\Big(\sum_i\|\check{\mathscr{V}}_i^{-1}-\mathscr{V}_i^{-1}\|^{2p/(p-1)}\Big)^{1-1/p}.\end{aligned}$$

Pick p no greater than two to obtain a rate of $o_p(n/k)$ by lemmas B1, C6. RHS2 in (38) is by the Schwarz inequality bounded by $\sqrt{\sum_i\big\|\sum_j w_{ij}V_j\big\|^4\|m_i\|^4\sum_i\|\hat{\mathscr{V}}_i^{-1}-\check{\mathscr{V}}_i^{-1}\|^4}$. By the Jensen inequality, $\sum_i\mathbb{E}\big\|\sum_j w_{ij}V_j\big\|^4\leq(C_w/k)^2\sum_{ij}\mathbb{E}\big[w_{ij}\|V_j\|^4\big]=O(n/k^2)$ by lemma B1. Finally, using $\hat{\mathscr{V}}_i^{-1}-\check{\mathscr{V}}_i^{-1}=\hat{\mathscr{V}}_i^{-1}(\check{\mathscr{V}}_i-\hat{\mathscr{V}}_i)\check{\mathscr{V}}_i^{-1}$, we have that $\sum_i\|\hat{\mathscr{V}}_i^{-1}-\check{\mathscr{V}}_i^{-1}\|^4\leq\max_i\|\hat{\mathscr{V}}_i^{-1}\|^4\max_i\|\check{\mathscr{V}}_i^{-1}\|^4\sum_i\|\hat{\mathscr{V}}_i-\check{\mathscr{V}}_i\|^4=o_p\big(n^2(\log n)^4/k^3\big)$, by lemmas C1, C3, and C4. □

Lemma C12. $\sum_i(\hat{\hat{\alpha}}_i\hat{\hat{\alpha}}_i^\intercal-\hat{\alpha}_i\hat{\alpha}_i^\intercal)=o_p(\rho^2)$.

Proof. The LHS is bounded in norm by a constant times $\sum_i\|\hat{\hat{\alpha}}_i-\hat{\alpha}_i\|^2+\sqrt{\sum_i\|\hat{\alpha}_i\|^2\sum_i\|\hat{\hat{\alpha}}_i-\hat{\alpha}_i\|^2}=o_p(\rho^2)$ by lemmas C10, C11. □

Let $\hat{\alpha}^*_{ij} = w_{ij} m_{\theta j}^\intercal \hat{\mathscr{V}}_i^{-1} m_i$, $\alpha^*_{ij} = w_{ij} m_{\theta j}^\intercal \mathscr{V}_i^{-1} m_i$.

Lemma C13. $\sum_{ij}(\hat{\alpha}^*_{ij}\hat{\alpha}^{*\intercal}_{ji} - \alpha^*_{ij}\alpha^{*\intercal}_{ji}) = o_p(\rho^2)$.

Proof. It suffices to establish that (i) $\sum_{ij}\|\alpha^*_{ij}\|^2 = O_p(\rho^2)$ and that (ii) $\sum_{ij}\|\hat{\alpha}^*_{ij} - \alpha^*_{ij}\|^2 = o_p(\rho^2)$. The LHS in (i) is bounded by $(C_w/k)\sum_{ij} w_{ij}\|m_{\theta j}\|^2\|\mathscr{V}_i^{-1}\|^2\|m_i\|^2 = O_p(n/k)$. The LHS in (ii) is bounded by

$$\frac{2C_w}{k}\sqrt{\sum_i\Big(\sum_j w_{ij}\|m_{\theta j}\|^2\Big)^2\|m_i\|^4\Big(\sum_i\|\hat{\mathscr{V}}_i^{-1} - \check{\mathscr{V}}_i^{-1}\|^2 + \sum_i\|\check{\mathscr{V}}_i^{-1} - \mathscr{V}_i^{-1}\|^2\Big)} \overset{\text{Schwarz,C6}}{\leq}$$
$$\frac{2C_w}{k}\sqrt{\sum_{ij} w_{ij}\|m_{\theta j}\|^4\|m_i\|^4\Big(\max_i\|\hat{\mathscr{V}}_i^{-1}\|^2\max_i\|\check{\mathscr{V}}_i^{-1}\|^2\sum_i\|\hat{\mathscr{V}}_i - \check{\mathscr{V}}_i\|^2 + o_p(n)\Big)} = o_p\Big(\frac{n}{k}\Big),$$

by lemmas B1,C1,C3,C4. □

Lemma C14. $\hat{\hat{D}}^2 - \hat{D}^2 = o_p(\rho^2)$.

Proof. Follows from lemmas C9, C12, C13. □

C.2. $\mathscr{V}_i$. Let $M^*_i(\theta) = \mathscr{V}_i^{-1}(\theta)M_i(\theta)$, $\tilde{M}^*_i(\theta) = \mathscr{V}_i^{-1}(\theta)\tilde{M}_i(\theta)$, and $u^*_i(\theta) = m^*_i(\theta) - M^*_i(\theta)$.

Lemma C15. $\sum_i \hat{G}_i^\intercal(\theta)u^*_i(\theta) = \sum_i G_i^\intercal(\theta)u^*_i(\theta) + \sum_{ij} w_{ij}V_j^\intercal(\theta)u^*_i(\theta) + o_p(\sqrt{n\lambda})$.

Proof. Since $\hat{G}_i = \sum_j w_{ij}m_{\theta j} = G_i + \sum_j w_{ij}(G_j - G_i) + \sum_j w_{ij}V_j$, we consider (omitting θ)

$$\mathbb{E}\Big\|\sum_{ij} w_{ij}(G_j - G_i)^\intercal u^*_i\Big\|^2 \leq \sum_{ijs}\mathbb{E}\big[w_{ij}w_{is}\|G_j - G_i\|\|G_s - G_i\|\|u^*_i\|^2\big]$$
$$\leq \sum_{ijs}\mathbb{E}\left[w_{ij}w_{is}\big(\|G_j - G_i\|^2 + \|G_s - G_i\|^2\big)\|u^*_i\|^2\right] = 2\sum_{ij}\mathbb{E}\big[w_{ij}\|G_j - G_i\|^2\|u^*_i\|^2\big] \overset{\text{B1}}{=} o(n\lambda^2). \qquad \square$$

Lemma C16. $\sum_i \hat{G}_i^\intercal(\theta)M^*_i(\theta) = n\mathbb{E}\big[G_1^\intercal(\theta)M^*_1(\theta)\big] + o_p(n\lambda^2 + \rho)$.

Proof. The LHS equals (omitting θ)

$$\lambda^2\sum_i(\tilde{G}_i^\intercal\tilde{M}^*_i - \mathbb{E}[\tilde{G}_1^\intercal\tilde{M}^*_1]) + \lambda\sum_{ij} w_{ij}V_j^\intercal\tilde{M}^*_i + \lambda^2\sum_{ij} w_{ij}(\tilde{G}_j - \tilde{G}_i)^\intercal\tilde{M}^*_i. \tag{39}$$

RHS1 in (39) is $O_p(\sqrt{n}\lambda^2)$ and RHS3 is $o_p(n\lambda^2)$ by lemma B1. Finally, take the norm of RHS2, square and take its expectation conditional on $\mathscr{Z}$ (using lemma B10) to obtain a bound of

$$\lambda^2\sum_{ijt} w_{ij}w_{tj}\mathbb{E}\big[\|V_j\|^2|z_j\big]\|\tilde{M}^*_i\|\|\tilde{M}^*_t\| \leq 2\lambda^2\max_j\Big(\sum_t w_{tj}\Big)\sum_{ij} w_{ij}\mathbb{E}\big[\|V_j\|^2|z_j\big]\|\tilde{M}^*_i\|^2 \overset{\text{B4}}{=} O_p(n\lambda^2).$$

□

Recall the definition of $\hat{\tilde{\alpha}}_i, \hat{\alpha}_i$, defined prior to lemma C9 and let $\alpha_i(\theta) = G_i^\intercal(\theta)\mathscr{V}_i^{-1}(\theta)m_i(\theta)$.

Lemma C17. $\sum_i \hat{\alpha}_i(\theta) - n\mathbb{E}\alpha_1(\theta) = o_p(n\lambda^2) + O_p(\rho)$.

Proof. In view of lemmas C15, C16, it suffices to show that (omitting θ) (i) $\sum_i G_i^\intercal u_i^* = O_p(\rho)$ and that (ii) $\sum_{ij} w_{ij} V_j^\intercal u_i^* = O_p(\rho)$. The expectation of the squared norm of the LHS in (i) is $\lambda^2 \sum_i \mathbb{E}\|\tilde{G}_i^\intercal u_i^*\|^2 = O(n\lambda^2) = O(\rho^2)$ and the expectation of the squared norm of (ii) is bounded by twice $\sum_{ij} \mathbb{E}\left[w_{ij}^2 \|V_j\|^2 \|u_i^*\|^2\right] = O(n/k) = O(\rho^2)$ by lemma B1. □

Lemma C18. $\sum_i \hat{\alpha}_i(\theta) \sum_i \hat{\alpha}_i^\intercal(\theta) - n^2 \mathbb{E}\alpha_1(\theta) \mathbb{E}\alpha_1^\intercal(\theta) = o_p(\rho^4)$.

Proof. Omitting θ, the LHS is bounded in norm by $\sum_i \|\hat{\alpha}_i - \mathbb{E}\alpha_1\|^2 + 2\|\mathbb{E}\alpha_1\| \sum_i \|\hat{\alpha}_i - \mathbb{E}\alpha_1\| = o_p(\rho^4)$ by lemma C17 and the fact that $\|\mathbb{E}\alpha_1\| = O(\lambda^2)$. □

Let $\hat{\alpha}_{Vi}(\theta) = \sum_j w_{ij} V_j^\intercal(\theta) m_i^*(\theta)$ and $\hat{\alpha}_i^*(\theta) = \left(G_i(\theta) + \sum_j w_{ij} V_j(\theta)\right)^\intercal m_i^*(\theta)$.

Lemma C19. $\sum_i \left(\hat{\alpha}_i(\theta)\hat{\alpha}_i^\intercal(\theta) - \alpha_i(\theta)\alpha_i^\intercal(\theta) - \hat{\alpha}_{Vi}(\theta)\hat{\alpha}_{Vi}^\intercal(\theta)\right) = o_p(\rho^2)$.

Proof. We show separately that (i) $\sum_i (\hat{\alpha}_i \hat{\alpha}_i^\intercal - \hat{\alpha}_i^* \hat{\alpha}_i^{*\intercal}) = o_p(\rho^2)$ and that (ii) $\sum_i (\hat{\alpha}_i^* \hat{\alpha}_i^{*\intercal} - \alpha_i \alpha_i^\intercal - \hat{\alpha}_{Vi} \hat{\alpha}_{Vi}^\intercal) = o_p(\rho^2)$. The LHS in (i) is bounded by twice $\sum_i \|\hat{\alpha}_i - \hat{\alpha}_i^*\|^2 + \sqrt{\sum_i \|\hat{\alpha}_i - \hat{\alpha}_i^*\|^2 \sum_i \|\hat{\alpha}_i^*\|^2}$. Now, $\sum_i \|\hat{\alpha}_i^*\|^2/2 \le \lambda^2 \sum_i \|\tilde{G}_i\|^2 \|m_i^*\|^2 + \sum_i \left\|\sum_j w_{ij} V_j\right\|^2 \|m_i^*\|^2 = O_p(\rho^2)$. Further, by the Jensen inequality $\sum_i \|\hat{\alpha}_i - \hat{\alpha}_i^*\|^2 \le \lambda^2 \sum_{ij} w_{ij} \|\tilde{G}_j - \tilde{G}_i\|^2 \|m_i^*\|^2 = o_p(\rho^2)$ by lemma B1. The norm of the LHS in (ii) is twice $\left\|\sum_i \alpha_i \hat{\alpha}_{Vi}^\intercal\right\|$. Thus, using lemma B10,

$$\begin{aligned}
\mathbb{E}\left[\left\|\sum_i \alpha_i \hat{\alpha}_{Vi}^\intercal\right\|^2 | \mathscr{Z}\right] &= \lambda^2 \mathbb{E}\left[\left\|\sum_{ij} w_{ij} \tilde{G}_i^\intercal m_i^* m_i^{*\intercal} V_j\right\|^2 | \mathscr{Z}\right] \\
&\le \lambda^2 \sum_{ijt} w_{ij} w_{tj} \|\tilde{G}_i\| \|\tilde{G}_t\| \mathbb{E}\left[\|\tilde{m}_i^*\|^2 \|\tilde{m}_t^*\|^2 | z_i, z_t\right] \mathbb{E}\left[\|V_j\|^2 | z_j\right] \\
&\le 2\lambda^2 \sum_{ijt} w_{ij} w_{tj} \|\tilde{G}_i\|^2 \mathbb{E}\left[\|\tilde{m}_i^*\|^4 | z_i\right] \mathbb{E}\left[\|V_j\|^2 | z_j\right] \overset{B4}{=} O_p(\lambda^2 n) = o_p(\rho^4).
\end{aligned}$$ □

Lemma C20. $\sum_i \hat{\alpha}_{Vi}(\theta) \hat{\alpha}_{Vi}^\intercal(\theta) - \sum_{ij} w_{ij}^2 \mathbb{E}\left[V_j^\intercal(\theta) m_i^*(\theta) m_i^{*\intercal}(\theta) V_j(\theta) | z_i, z_j\right] = o_p(\rho^2)$.

Proof. Note first that the LHS is

$$\left(\sum_{ij} w_{ij}^2 \left(V_j^\intercal m_i^* m_i^{*\intercal} V_j - \mathbb{E}\left[V_j^\intercal m_i^* m_i^{*\intercal} V_j | z_i, z_j\right]\right)\right) + \sum_{ij} \sum_{t \ne j} w_{ij} w_{it} V_j^\intercal m_i^* m_i^{*\intercal} V_t. \tag{40}$$

Using lemma B10, we have for the second term in (40) that for some $C < \infty$,

$$\begin{aligned}
\mathbb{E}\left[\left\|\sum_{ij} \sum_{t \ne j} w_{ij} w_{it} V_j^\intercal m_i^* m_i^{*\intercal} V_t\right\|^2 | \mathscr{Z}\right] &\le C \sum_{ijts} w_{ij} w_{it} w_{sj} w_{st} \mathbb{E}\left[\|V_j\|^2 | z_j\right] \mathbb{E}\left[\|V_t\|^2 | z_t\right] \mathbb{E}\left[\|m_i^*\|^2 \|m_s^*\|^2 | z_i, z_s\right] \\
&\overset{B4}{\le} O_p(1/k) \sum_{ij} w_{ij} \mathbb{E}\left[\|V_j\|^2 | z_j\right]^2 \mathbb{E}\left[\|m_i^*\|^4 | z_i\right] = O_p(n/k) = o_p(\rho^4).
\end{aligned}$$

Again by lemma B10, we have that the expectation of the squared norm of the first term in (40) given $\mathscr{Z}$ is for some $C < \infty$ bounded by

$$C \sum_{ijt} w_{ij}^2 w_{it}^2 \mathbb{E}\left[\|V_j\|^2 | z_j\right] \mathbb{E}\left[\|V_t\|^2 | z_t\right] \mathbb{E}\left[\|m_i^*\|^4 | z_i\right] \le \frac{C^2}{k^2} \sum_{ij} w_{ij} \left(\mathbb{E}\left[\|V_j\|^2 | z_j\right]\right)^2 \mathbb{E}\left[\|m_i^*\|^4 | z_i\right] = o_p(\rho^4).$$

□

Lemma C21. $\sum_{ij} w_{ij}w_{ji}m_{\theta j}^{\intercal}(\theta)m_i^*(\theta)m_j^{*\intercal}(\theta)m_{\theta i}(\theta) - \sum_{ij} w_{ij}w_{ji}\mathbb{E}\big[V_j^{\intercal}(\theta)m_i^*(\theta)m_j^{*\intercal}(\theta)V_i(\theta)|z_i,z_j\big] = o_p(\rho^2)$.

Proof. Write the LHS as

$$\sum_{ij} w_{ij}w_{ji}\big(m_{\theta j}^{\intercal}m_i^*m_j^{*\intercal}m_{\theta i} - \mathbb{E}\big[m_{\theta j}^{\intercal}m_i^*m_j^{*\intercal}m_{\theta i}|z_i,z_j\big]\big) + \lambda\sum_{ij} w_{ij}w_{ji}\mathbb{E}\big[V_j^{\intercal}m_i^*m_j^{*\intercal}\tilde{G}_i|z_i,z_j\big] + \lambda\sum_{ij} w_{ij}w_{ji}\mathbb{E}\big[\tilde{G}_j^{\intercal}m_i^*m_j^{*\intercal}V_i|z_i,z_j\big] + \lambda^4\sum_{ij} w_{ij}w_{ji}\tilde{G}_j^{\intercal}\tilde{M}_i^*\tilde{M}_j^{*\intercal}\tilde{G}_i. \quad (41)$$

By lemma B6(iii) the first term in (41) is $O_p(n^{1/2}k^{-1}) = o_p(\rho^2)$. The expectation of the fourth term in (41) is bounded by $(n^4\lambda^4C_w/k)\mathbb{E}\big[w_{12}\|\tilde{M}_1^*\|\|\tilde{M}_2^*\|\|\tilde{G}_1\|\|\tilde{G}_2\|\big] \overset{B5}{=} O(n\lambda^4/k) = o(\rho^2)$. Finally, the second and third terms in (41) are $O_p(n\lambda^2/k) = o_p(\rho^2)$, which can be similarly established. □

Lemma C22. $\hat{D}^2(\theta) - \tilde{D}^2(\theta) = o_p(\rho^2)$.

Proof. Omitting the θ argument, the LHS is

$$\underbrace{\sum_i(\hat{\alpha}_i\hat{\alpha}_i^{\intercal} - \alpha_i\alpha_i^{\intercal} - \hat{\alpha}_{Vi}\hat{\alpha}_{Vi}^{\intercal})}_{\overset{C19}{=}o_p(\rho^2)} - \underbrace{n^{-1}\Big(\sum_i\hat{\alpha}_i\sum_i\hat{\alpha}_i^{\intercal} - n^2\mathbb{E}\alpha_1\mathbb{E}\alpha_1^{\intercal}\Big)}_{\overset{C18}{=}o_p(\rho^4/n)=o_p(\rho^2)} + \underbrace{\sum_i(\alpha_i\alpha_i^{\intercal} - \mathbb{E}[\alpha_1\alpha_1^{\intercal}])}_{=O_p(\sqrt{n}\lambda^2)}$$
$$+ \underbrace{\Big(\sum_i\hat{\alpha}_{Vi}\hat{\alpha}_{Vi}^{\intercal} - \sum_{ij}w_{ij}^2\mathbb{E}\big[V_jm_i^*m_i^{*\intercal}V_j^{\intercal}|z_i,z_j\big]\Big)}_{\overset{C20}{=}o_p(\rho^2)} + \underbrace{\sum_{ij}w_{ij}w_{ji}\big(m_{\theta j}^{\intercal}m_i^*m_j^{*\intercal}m_{\theta i} - \mathbb{E}\big[V_j^{\intercal}m_i^*m_j^{*\intercal}V_i|z_i,z_j\big]\big)}_{\overset{C21}{=}o_p(\rho^2)}$$
$$= o_p(\rho^2). \quad \square \quad (42)$$

Lemma C23. $\tilde{D}^2(\theta) - n\operatorname{Var}[\alpha_1(\theta)] = O_p(n/k)$.

Proof. The expectation of the norm of the LHS is bounded by a constant times $\sum_{ij}w_{ij}^2\mathbb{E}\big[\|V_j\|^2|z_j\big]\mathbb{E}\big[\|m_i^*\|^2|z_i\big] = O_p(n/k)$ by lemma B5. □

Lemma C24. $\tilde{D}(\theta) = O_p(\rho)$.

Proof. Apply lemma C23 and note that $n\operatorname{Var}[\alpha_1] = O(n\lambda^2)$, such that $\tilde{D}^2 = O_p(\rho^2)$. □

Lemma C25. $\tilde{D}^{-1} = O_p(\rho^{-1})$.

Proof. Recall that $\ell_{ij} = w_{ij}V_j^{\intercal}m_i^*$, such that $\tilde{D}^2 = n\operatorname{Var}[\alpha_1] + \sum_{ij}\mathbb{E}[\ell_{ij}\ell_{ij}^{\intercal}|\mathscr{Z}] + \sum_{ij}\mathbb{E}[\ell_{ij}\ell_{ji}^{\intercal}|\mathscr{Z}]$. Now,

$$\frac{1}{\|\tilde{D}^{-2}\|} \geq \max\Big\{\underline{\mathrm{ev}}\big(n\operatorname{Var}[\alpha_1]\big), \underline{\mathrm{ev}}\Big(\sum_{ij}\mathbb{E}[\ell_{ij}\ell_{ij}^{\intercal}|\mathscr{Z}] + \sum_{ij}\mathbb{E}[\ell_{ij}\ell_{ji}^{\intercal}|\mathscr{Z}]\Big)\Big\}$$

Now, $\underline{\mathrm{ev}}\big(n\operatorname{Var}[\alpha_1]\big) \geq n\lambda^2\underline{\mathrm{ev}}\big(\mathbb{E}\big[\tilde{G}_1^{\intercal}\mathscr{V}_1^{-1}\tilde{G}_1\big]\big) \geq Cn\lambda^2$ for some $C > 0$ by assumption B. We show that for any fixed $\|c\| = 1$ and some $C > 0$,

$$\sum_{ij}c^{\intercal}\mathbb{E}[\ell_{ij}\ell_{ij}^{\intercal}|\mathscr{Z}]c + \sum_{ij}c^{\intercal}\mathbb{E}[\ell_{ij}\ell_{ji}^{\intercal}|\mathscr{Z}]c \geq Cn/k - o_p(n/k). \quad (43)$$

First, letting $A_{ij} = w_{ij}\mathbb{E}[m_i^* c^\intercal V_i^\intercal | z_i]$, note that the second term in (43) is equal to

$$\begin{aligned}\sum_{ij} w_{ij}w_{ji}\mathbb{E}\left[c^\intercal V_j^\intercal m_i^* m_j^{*\intercal} V_i c | z_i, z_j\right] &= \sum_{ij}\operatorname{tr}(A_{ij}A_{ji}) = \frac{1}{2}\sum_{ij}\operatorname{tr}(A_{ij}A_{ji} + A_{ji}^\intercal A_{ij}^\intercal) \\ &= \frac{1}{2}\sum_{ij}\operatorname{tr}\left\{(A_{ij}+A_{ji}^\intercal)(A_{ij}+A_{ji}^\intercal)^\intercal - (A_{ij}A_{ij}^\intercal + A_{ji}A_{ji}^\intercal)\right\} \geq -\sum_{ij}\operatorname{tr}(A_{ij}A_{ij}^\intercal) \\ &= -\sum_{ij} w_{ij}^2 \operatorname{tr}\big(\mathbb{E}[m_i^* c^\intercal V_i^\intercal | z_i]\mathbb{E}[V_i c m_i^{*\intercal} | z_i]\big). \quad (44)\end{aligned}$$

The first term in (43) equals

$$\begin{aligned}\sum_{ij} w_{ij}^2 \mathbb{E}\left[(c^\intercal V_j^\intercal m_i^*)^2 | z_i, z_j\right] &= \operatorname{tr}\Big(\sum_{ij} w_{ij}^2 \mathbb{E}\left[m_i^* m_i^{*\intercal} V_j c c^\intercal V_j^\intercal | z_i, z_j\right]\Big) \\ &\overset{\text{B1,B6(iii)}}{=} \sum_{ij} w_{ij}^2 \operatorname{tr}\big(\mathbb{E}[m_i^* m_i^{*\intercal}|z_i]\mathbb{E}[V_i c c^\intercal V_i^\intercal | z_i]\big) + o_p(n/k). \quad (45)\end{aligned}$$

Noting that $m_i^* = u_i^* + \lambda \tilde{M}_i^*$, it follows that the sum of the RHS in (44) and that in (45) is bounded below by

$$\begin{aligned}&\sum_{ij} w_{ij}^2 \operatorname{tr}\big(\operatorname{Var}[u_i^*|z_i]\operatorname{Var}[V_i c|z_i] - \operatorname{Cov}[u_i^*, V_i c|z_i]\operatorname{Cov}[V_i c, u_i^*|z_i]\big) + o_p(n/k) \\ &\quad = \sum_{ij} w_{ij}^2 \operatorname{tr}\left\{\mathscr{V}_i^{-2}\big(\operatorname{Var}[u_i|z_i]\operatorname{Var}[V_i c|z_i] - \operatorname{Cov}[u_i, V_i c|z_i]\operatorname{Cov}[V_i c, u_i|z_i]\big)\right\} + o_p(n/k) \\ &\geq \frac{C_w^- n}{k}\mathbb{E}\operatorname{tr}\left\{\mathscr{V}_1^{-2}\big(\operatorname{Var}[u_1|z_1]\operatorname{Var}[V_1 c|z_1] - \operatorname{Cov}[u_1, V_1 c|z_1]\operatorname{Cov}[V_1 c, u_1|z_1]\big)\right\} - o_p(n/k) \\ &\geq Cn/k - o_p(n/k),\end{aligned}$$

for some $C > 0$ by assumption A. □

Appendix D. Martingale Differences

Lemma D1. *Let $\{X_{ni}, \mathcal{F}_{ni}\}$ be a martingale difference array for which (i) $\sup_n \mathbb{E}\left[\max_{i\leq n} X_{ni}^2\right] < \infty$, (ii) $\sum_{i=1}^n X_{ni}^2 \xrightarrow{p} 1$ and (iii) $\max_{i\leq n}|X_{ni}| \xrightarrow{p} 0$. Then $\sum_{i=1}^n X_{ni} \xrightarrow{d} N(0,1)$.*

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

Lemma D2. $\max_i \|\tilde{D}^{-1}B_i\| = o_p(1)$.

Proof. Since $\|\tilde{D}^{-1}\| = O_p(\rho^{-1})$ by lemma C25, if suffices to show that $\max_i \|B_i\| = o_p(\rho)$. Pick any $\epsilon > 0$. Then for any $2 < p \leq 4$ (suppressing θ_0),

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

So by the Minkowski inequality it suffices to show that

$$n\mathbb{E}\|\zeta_1\|^p = o(\rho^p), \quad \sum_i \mathbb{E}\Big\|\sum_{j<i}\ell_{ij}\Big\|^p = o(\rho^p), \quad \sum_i \mathbb{E}\Big\|\sum_{j<i}\ell_{ji}\Big\|^p = o(\rho^p). \tag{46}$$

The LHS in the first condition in (46) is $O(n\lambda^p) = o(\rho^p)$. The second and third conditions in (46) are similar; we only show the second. We have for some fixed $C_p < \infty$,

$$\sum_i \mathbb{E}\Big\|\sum_{j<i}\ell_{ij}\Big\|^p = \sum_i \mathbb{E}\Big\|\sum_{j<i} w_{ij}V_j^\intercal m_i^*\Big\|^p \overset{\text{Burkholder}}{\leq} C_p \sum_i \mathbb{E}\Big[\sum_{j<i} w_{ij}^2\|m_i^*\|^2\|V_j\|^2\Big]^{p/2}$$
$$\overset{\text{Jensen}}{\leq} C_p \sum_i \mathbb{E}\Big[i^{p/2-1}\sum_{j<i} w_{ij}^p\|m_i^*\|^p\|V_j\|^p\Big] \leq C_p n^{p/2+1}\mathbb{E}\big[w_{12}^p\|m_1^*\|^p\|V_2\|^p\big] \overset{\text{B5}}{=} O(n^{p/2}k^{1-p}) = o(\rho^p).$$

□

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

Proof. Noting that $\tilde{D}^2(\theta_0) = n\mathbb{E}[\zeta_1\zeta_1^\intercal] + \sum_i\sum_{j<i}\mathbb{E}[\xi_{ij}\xi_{ij}^\intercal|\mathscr{Z}]$, we have (omitting the θ_0–argument)

$$\sum_i B_iB_i^\intercal = \tilde{D}^2 + \sum_i(\zeta_i\zeta_i^\intercal - \mathbb{E}[\zeta_i\zeta_i^\intercal]) + \sum_i\sum_{j<i}(\zeta_i\xi_{ij}^\intercal + \xi_{ij}\zeta_i^\intercal)$$
$$+ \sum_i\sum_{j<i}\{\xi_{ij}\xi_{ij}^\intercal - \mathbb{E}[\xi_{ij}\xi_{ij}^\intercal|\mathscr{Z}]\} + 2\sum_i\sum_{j<i}\sum_{t<j}\xi_{ij}\xi_{it}^\intercal. \tag{47}$$

For RHS2 in (47), note that $\sum_i \mathrm{Var}\big[\mathrm{vec}(\zeta_i\zeta_i^\intercal)\big] = O(n\lambda^4) = o(\rho^4)$. For RHS3, note that (the case $\sum_{j>i}$ is symmetric)

$$\sum_i\sum_{j<i} w_{ij}\zeta_i m_i^{*\intercal}V_j + \sum_i\sum_{j<i} w_{ji}\zeta_i m_j^{*\intercal}V_i \overset{\text{B6(ii),(iv)}}{=} o_p(n^{3/4}k^{-1/4}\lambda) + O_p(n^{1/2}\lambda) = o_p(\rho^2),$$

where lemma B6 is applied element–wise with the c_{ij}–weights taken as $c_{ij} = I(j<i)$. RHS4 in (47) is $o_p(\rho^2)$ because using lemma B6(iii),

$$\sum_i\sum_{j<i}(\ell_{ij}\ell_{ij}^\intercal - \mathbb{E}[\ell_{ij}\ell_{ij}^\intercal|\mathscr{Z}]) = \sum_i\sum_{j<i} w_{ij}^2(V_j^\intercal m_i^*m_i^{*\intercal}V_j - \mathbb{E}[V_j^\intercal m_i^*m_i^{*\intercal}V_j|z_i,z_j]) = O_p(n^{1/2}/k) = o_p(\rho^2),$$
$$\sum_i\sum_{j<i}(\ell_{ij}\ell_{ji}^\intercal - \mathbb{E}[\ell_{ij}\ell_{ji}^\intercal|\mathscr{Z}]) = \sum_i\sum_{j<i} w_{ij}w_{ji}(V_j^\intercal m_i^*m_j^{*\intercal}V_i - \mathbb{E}[V_j^\intercal m_i^*m_j^{*\intercal}V_i|z_i,z_j]) = O_p(n^{1/2}/k) = o_p(\rho^2).$$

Finally, RHS5 in (47) is $o_p(\rho^2)$ because for arbitrary elements $\xi_{ij}^*, \xi_{it}^{**}$ of ξ_{ij}, ξ_{it}, $\mathbb{E}\big[\sum_i\sum_{j<i}\sum_{t<j}\xi_{ij}^*\xi_{it}^{**}\big]^2 = \sum_i\sum_{j<i}\sum_{t<j}\mathbb{E}[\xi_{ij}^{*2}\xi_{it}^{**2}] = o(\rho^4)$, by lemma B5. □

Lemma D4. $\sup_n \mathbb{E}\big[\max_{i\leq n}\|B_i^\intercal(\theta_0)\tilde{D}^{-1}(\theta_0)\|^2\big] < \infty$.

Proof. Omit θ_0 throughout. Note that

$$\tilde{D}^{-1}\sum_i\mathbb{E}[B_iB_i^\intercal|\mathscr{Z}]\tilde{D}^{-1} = \tilde{D}^{-1}\Big(\tilde{D}^2 + \sum_i(G_i^\intercal\mathscr{V}_i^{-1}G_i - \mathbb{E}[G_1^\intercal\mathscr{V}_1^{-1}G_1])\Big)\tilde{D}^{-1}$$
$$= I + \sum_i\tilde{D}^{-1}(G_i^\intercal\mathscr{V}_i^{-1}G_i - \mathbb{E}[G_1^\intercal\mathscr{V}_1^{-1}G_1])\tilde{D}^{-1}.$$

For $\lambda = 0$ the result is immediate. For $\lambda > 0$, note that for some $C < \infty$, $\|\tilde{D}^{-2}\| \leq \left\|\left(nV[G_1 m_1^*]\right)^{-1}\right\| \leq C/n\lambda^2$. Thus,

$$\max_{i\leq n} \|B_i^\intercal \tilde{D}^{-1}\|^2 \leq \sum_i \|B_i^\intercal \tilde{D}^{-1}\|^2 \leq 1 + \frac{C}{n}\left\|\sum_i \left(\tilde{G}_i^\intercal \mathscr{V}_i^{-1}\tilde{G}_i - \mathbb{E}[\tilde{G}_1^\intercal \mathscr{V}_1^{-1}\tilde{G}_1]\right)\right\| \leq 1 + C\mathbb{E}\|\mathscr{V}_1^{-1}\tilde{G}_1\|^4 < \infty. \quad \square$$

Appendix E. Local Alternatives

E.1. **Uniform Lemmas.** Let $\mathscr{N}$ be a convex compact set including θ_0 and $|\mathscr{N}| = \sup_{\theta\in\mathscr{N}} \|\theta - \theta_0\|$.

Lemma E1. *For $p_1 > 0$ let $f_i = f_i(\theta) = f(z_i, \theta)$ be such that $\mathbb{E}\left|\sup_{\mathscr{N}} f_i\right|^{p_1 p_2} < \infty$ and $a_i = a_i(\theta) > 0$ be such that $\mathbb{E}\sup_{\mathscr{N}} a_i^{p_2/(p_2-1)} < \infty$ for some $p_2 \geq 1$. If $\lim_{|\mathscr{N}|\to 0} \mathbb{E}\left[\sup_{\mathscr{N}} f_i - \inf_{\mathscr{N}} f_i\right]^{p_2} = 0$ then*

$$\lim_{|\mathscr{N}|\to 0} n\mathbb{E}\Big[w_{12} \sup_{\theta\in\mathscr{N}} \left(a_1|f_1 - f_2|\right)\Big] = 0.$$

Proof. Let sup and inf be taken over $\theta \in \mathscr{N}$ for the remainder of this lemma. Now,

$$n\mathbb{E}\left[w_{12}\sup\left(a_1|f_1 - f_2|\right)\right] \leq n\mathbb{E}\left[w_{12}\sup a_1|\sup f_1 - \inf f_2|\right] + n\mathbb{E}\left[w_{12}\sup a_1|\inf f_1 - \sup f_2|\right]. \tag{48}$$

Both RHS terms in (48) can be dealt with using the same argument, so we concentrate on RHS1 which is by the triangle inequality bounded by

$$\mathbb{E}\left[\sup a_1|\sup f_1 - \inf f_1|\right] + n\mathbb{E}\left[w_{12}\sup a_1|\inf f_1 - \inf f_2|\right] \overset{\text{B1}}{=} \mathbb{E}\left[\sup a_1|\sup f_1 - \inf f_1|\right] + o(1)$$
$$\overset{\text{Hölder}}{\leq} \left(\mathbb{E}\left[\sup a_1^{p_2/(p_2-1)}\right]\right)^{1-1/p_2}\left(\mathbb{E}\left|\sup f_1 - \inf f_1\right|^{p_2}\right)^{1/p_2} + o(1) \to 0 \text{ as } |\mathscr{N}| \to 0. \quad \square$$

Lemma E2. *Consider a process $\mathscr{T}_n$ which for any polynomial $\mathscr{P}_n$ and any $1 = O(\epsilon_n)$, $\sup_\theta \mathbb{P}[|\mathscr{T}_n(\theta)| > \epsilon_n] = o(\mathscr{P}_n^{-1})$ and for some polynomial $\mathscr{P}_n^*$ satisfies $\sup_\theta \|\mathscr{T}_{n\theta}(\theta)\| = o_p(\mathscr{P}_n^*)$ then $\sup_\theta |\mathscr{T}_n(\theta)| = o_p(\epsilon_n)$.*

Proof. Divide Θ into convex sets $\Theta_1, \ldots, \Theta_{\mathscr{P}_n}$ such that for all $t = 1, \ldots, \mathscr{P}_n$ and all $\theta, \tilde{\theta} \in \Theta_t$: $\|\theta - \tilde{\theta}\| \leq \epsilon_n/\mathscr{P}_n^*$. Then if $\theta_t \in \Theta_t$,

$$\sup_\theta |\mathscr{T}_n(\theta)| \leq \max_{t=1,\ldots,\mathscr{P}_n} \sup_{\theta\in\Theta_t} |\mathscr{T}_n(\theta) - \mathscr{T}_n(\theta_t)| + \max_{t=1,\ldots,\mathscr{P}_n} |\mathscr{T}_n(\theta_t)|. \tag{49}$$

RHS1 in (49) is by the mean value theorem bounded by $\sup_\theta \|\mathscr{T}_{n\theta}(\theta)\|\epsilon_n/\mathscr{P}_n^* = o_p(\epsilon_n)$. For RHS2 note that

$$\mathbb{P}\Big[\max_{t=1,\ldots,\mathscr{P}_n} |\mathscr{T}_n(\theta_t)| > \epsilon_n\Big] \leq \mathscr{P}_n \sup_\theta \mathbb{P}\left[|\mathscr{T}_n(\theta)| > \epsilon_n\right] = o(1). \quad \square$$

Lemma E3. *Let $\{a_i, b_i\}$ be an i.i.d. sequence such that for some $p_a, p_b > 2$ with $1/p_a + 1/p_b < 1/2$, $\mathbb{E}\sup_\theta |a_i(\theta)|^{p_a} < \infty$, $\mathbb{E}\sup_\theta |b_i(\theta)|^{p_b} < \infty$, $\mathbb{E}\sup_\theta \|a_{\theta i}(\theta)\| < \infty$, $\mathbb{E}\sup_\theta \|b_{\theta i}(\theta)\| < \infty$, and further that $\mathbb{E}[a_i(\theta)|\mathscr{Z}] = 0$ a.s., $b_i(\theta) = b(z_i, \theta)$. Then $\sup_\theta\left|\sum_{ij} w_{ij} b_j(\theta) a_i(\theta)\right| = o_p(\sqrt{n}\log n)$.*

Proof. Let $\tilde{a}_i = a_i I(|a_i| \leq Cn^{1/p_a}) - \mathbb{E}[a_i I(|a_i| \leq Cn^{1/p_a})|z_i]$, $\sigma_{ai}^2 = \mathrm{Var}[\tilde{a}_i|z_i]$, and $\epsilon_n = \sqrt{n}\log n$. We first show that $\sup_\theta |\tilde{\mathscr{S}}_n - \mathscr{S}_n| = o_p(\epsilon_n)$ for $\tilde{\mathscr{S}}_n = \sum_{ij} w_{ij} b_j \tilde{a}_i$ and $\mathscr{S}_n = \sum_{ij} w_{ij} b_j a_i$. Now, since

for any two random variables $\mathbb{E}_1, \mathbb{E}_2$, $\mathbb{P}[|\mathbb{E}_1| + |\mathbb{E}_2| > \epsilon_n] \leq \mathbb{P}[\mathbb{E}_1 \neq 0] + \mathbb{P}[|\mathbb{E}_2| > \epsilon_n]$ and noting that $\mathbb{E}\left[a_i I(|a_i| > Cn^{1/p_a})|z_i\right] = -\mathbb{E}\left[a_i I(|a_i| \leq Cn^{1/p_a})|z_i\right]$ a.s.,

$$\begin{aligned}\mathbb{P}\Big[\sup_\theta |\mathscr{S}_n - \tilde{\mathscr{S}}_n| > \epsilon_n\Big] \\ \leq \mathbb{P}\Big[\max_i \sup_\theta |a_i| > Cn^{1/p_a}\Big] + \mathbb{P}\Big[\sum_{ij} w_{ij} \sup_\theta |b_j| \sup_\theta \mathbb{E}[|a_i| I(|a_i| > Cn^{1/p_a})|z_i] > \epsilon_n\Big].\end{aligned} \tag{50}$$

RHS1 in (50) is by the Markov inequality bounded by $\mathbb{E}\sup|a_i|^{p_a}/C^{p_a} \to 0$ as $C \to \infty$. For RHS2 in (50), note first that $\max_j \sup_\theta |b_j| = O_p(n^{1/p_b})$, such that

$$\sum_{ij} w_{ij} \sup_\theta |b_j| \sup_\theta \mathbb{E}[|a_i| I(|a_i| > Cn^{1/p_a})|z_i] = O_p(n^{1/p_b}) \sum_i \sup_\theta \mathbb{E}[|a_i| I(|a_i| > Cn^{1/p_a})|z_i]. \tag{51}$$

Since for $a_i^* = \sup_\theta |a_i|$ by the Hölder and Markov inequalities,

$$\mathbb{E}[a_i^* I(a_i^* > Cn^{1/p_a})] \leq \left(\mathbb{E}|a_i^*|^{p_a}\right)^{1/p_a} \left(\mathbb{P}[a_i^* > Cn^{1/p_a}]\right)^{1-1/p_a} \leq C^{1-p_a} n^{1/p_a - 1} \mathbb{E}|a_i^*|^{p_a},$$

it follows that the RHS in (51) is $O_p(n^{1/p_b + 1/p_a}) = o_p(\sqrt{n})$ such that RHS2 in (50) is $o(1)$.

It remains to be shown that $\sup_\theta |\tilde{\mathscr{S}}_n| = o_p(\epsilon_n)$. Let $\mathscr{Q}_n = \sum_{ij} w_{ij} b_j^2 \sigma_{ai}^2$. Then

$$\begin{aligned}\mathbb{P}\Big[\sup_\theta |\tilde{\mathscr{S}}_n| > \epsilon_n\Big] \leq \mathbb{P}\Big[\sup_\theta |\tilde{\mathscr{S}}_n| > \epsilon_n,\ \sup_\theta \mathscr{Q}_n \leq Cn,\ \max_i \sup_\theta |b_i| \leq Cn^{1/p_b}\Big] \\ + \mathbb{P}\Big[\sup_\theta \mathscr{Q}_n > Cn\Big] + \mathbb{P}\Big[\max_i \sup_\theta |b_i| > Cn^{1/p_b}\Big].\end{aligned} \tag{52}$$

Note that by taking C to ∞ after taking the supremum over n, RHS3 in (52) tends to zero because $\max_i \sup_\theta |b_i| = O_p(n^{1/p_b})$. The same is true for RHS2 in (52) because $\sup_\theta \mathscr{Q}_n \leq \sum_{ij} w_{ij} \sup_\theta b_j^2 \sup_\theta \sigma_{ai}^2 = O_p(n)$ by lemma B1 and the moment conditions.

Finally RHS1 in (52). Note first that by theorem 1.1 of de la Peña (1999),

$$\begin{aligned}\sup_\theta \mathbb{P}\Big[|\tilde{\mathscr{S}}_n| > \epsilon_n \mid \mathscr{Z},\ \sup_\theta \mathscr{Q}_n \leq Cn,\ \max_i \sup_\theta |b_i| \leq Cn^{1/p_b}\Big] \\ \leq 2\exp\left(-\frac{\epsilon_n^2/2}{Cn + C^2 n^{1/p_a + 1/p_b}\epsilon_n}\right) \leq 2\exp\left(-\frac{(\log n)^2}{4C}\right),\end{aligned}$$

which converges faster than any polynomial of n. Finally, since by the conditions of this lemma, $\sum_{ij} w_{ij}\left(\sup_\theta \|b_{\theta j}\| \sup_\theta |a_i| + \sup_\theta |b_j| \sup_\theta \|a_{\theta i}\|\right)$ increases at a polynomial rate, both conditions of lemma E2 are satisfied for $\mathscr{T}_n(\theta) = \tilde{\mathscr{S}}_n(\theta) I\left(\max_i \sup_{\tilde{\theta}} |a_i(\tilde{\theta})| \leq Cn^{1/p_a}\right)$.[27] Since by the Markov inequality,

$$\mathbb{P}\Big[\sup_\theta I\left(\mathscr{T}_n \neq \tilde{\mathscr{S}}_n\right) = 1\Big] = \mathbb{P}\Big[\max_i \sup_\theta |a_i| > Cn^{1/p_a}\Big] \leq C^{-p_a} \mathbb{E} \sup_\theta |a_i|^{p_a} \to 0,$$

as $C \to \infty$, the lemma statement holds. □

[27] When $\max_i \sup_\theta |a_i(\theta)| \leq Cn^{1/p_a}$, $\tilde{\mathscr{S}}_n = \mathscr{S}_n$, and hence $\mathscr{T}_n$ is differentiable.

Lemma E4. *Let $\{a_i, b_i\}$ be an i.i.d. sequence such that for some $p_a, p_b > 2$ with $1/p_a + 1/p_b < 1/2$, $n^{2/p_b}/k = o(1)$, $\mathbb{E}\sup_\theta |a_i(\theta)|^{p_a} < \infty$, $\mathbb{E}\sup_\theta |b_i(\theta)|^{p_b} < \infty$, $\mathbb{E}\sup_\theta \|a_{\theta i}(\theta)\| < \infty$, $\mathbb{E}\sup_\theta \|b_{\theta i}(\theta)\| < \infty$, and further that $\mathbb{E}[a_i(\theta)|\mathscr{Z}] = \mathbb{E}[b_i(\theta)|\mathscr{Z}] = 0$ a.s.. Then $\sup_\theta \left|\sum_{ij} w_{ij} b_j(\theta) a_i(\theta)\right| = o_p\big(\sqrt{n/k}(\log n)^2\big)$.*

Proof. The arguments surrounding truncation of b_j, a_i are the same as those for a_i in lemma E3, so they are omitted here. We moreover take $\sigma_{ai}^2 = \mathrm{Var}[\tilde{a}_i|z_i]$ and $\sigma_{bi}^2 = \mathrm{Var}[\tilde{b}_i|z_i]$ to have an a.s. upper bound of one, uniformly in θ, n, because the same truncation argument can be used otherwise. The discussion of the application of lemma E2 is similar to that in lemma E3 and is omitted here, also. We hence limit ourselves to showing that for any polynomial $\mathscr{P}_n$,

$$\mathscr{P}_n \sup_\theta \mathbb{P}\big[|\sum_{ij} w_{ij}\tilde{a}_i\tilde{b}_j| > 2\sqrt{n/k}(\log n)^2\big] = o(1), \tag{53}$$

where $\tilde{a}_i = \bar{a}_i - \mathbb{E}[\bar{a}_i|\mathscr{Z}]$ with $\bar{a}_i = a_i I(|a_i| \le Cn^{1/p_a})$ and $\tilde{b}_i$ is similarly defined. Instead of (53) we in fact show that

$$\mathscr{P}_n \sup_\theta \mathbb{P}\Big[\Big|\sum_i \sum_{j<i} w_{ij}\tilde{b}_j\tilde{a}_i\Big| > \sqrt{n/k}(\log n)^2\Big] = o(1), \tag{54}$$

where the derivation for $\sum_{j>i}$ obtains by reversing the sum. Let $\mathscr{A}_i = \sum_{j<i} w_{ij}\tilde{b}_j\tilde{a}_i$, such that $\{\mathscr{A}_i\}$ is a martingale sequence. Then for $\epsilon_{An} = n^{1/p_a+1/p_b}(\log n)/\sqrt{k}$ by theorem 1.1 of de la Peña (1999),

$$\begin{aligned}\sup_\theta \mathbb{P}\Big[\max_i |\mathscr{A}_i| > \epsilon_{An}\Big] &\le \sum_{i=1}^n \sup_\theta \mathbb{E}\Big[\mathbb{P}\Big[\Big|\sum_j w_{ij}\tilde{b}_j\tilde{a}_i\Big| > \epsilon_{An}|\mathscr{Z}, \tilde{a}_i\Big]\Big] \\ &\le 2n\exp\Big(-\frac{\epsilon_{An}^2/2}{C_w C^4 n^{2/p_a+2/p_b} + \epsilon_{An} C_w C^2 n^{1/p_a+1/p_b}}\Big) = o\big(\mathscr{P}_n^{-1}\big). \end{aligned} \tag{55}$$

Further, for $\mathscr{Q}_n = \sum_i \mathrm{Var}[\mathscr{A}_i|\mathscr{Z}, \tilde{b}_1, \cdots, \tilde{b}_{i-1}] = \sum_i \big(\sum_{j<i} w_{ij}\tilde{b}_j\big)^2 \sigma_{ai}^2$ and $\epsilon_{Qn} = n(\log n)^2/k$, we have

$$\begin{aligned}\sup_\theta \mathbb{P}\big[\mathscr{Q}_n > \epsilon_{Qn}\big] &\le \sum_{i=1}^n \sup_\theta \mathbb{P}\Big[\Big|\sum_{j<i} w_{ij}\tilde{b}_j\Big| > \sqrt{\frac{\epsilon_{Qn}}{n}}\Big] \\ &\le 2n\mathbb{E}\exp\Big(-\frac{(\log n)^2/(2k)}{\sum_{j<i} w_{ij}^2 + C_w C n^{1/p_b}(\log n)/k^{3/2}}\Big) = o\big(\mathscr{P}_n^{-1}\big). \end{aligned} \tag{56}$$

Finally, letting $\epsilon_n = \sqrt{n/k}(\log n)^2$,

$$\begin{aligned}&\sup_\theta \mathbb{P}\Big[\Big|\sum_i \mathscr{A}_i\Big| > \epsilon_n\Big] \le \\ &\sup_\theta \mathbb{P}\Big[\Big|\sum_i \mathscr{A}_i\Big| > \epsilon_n,\ \mathscr{Q}_n \le \epsilon_{Qn}, \max_i |\mathscr{A}_i| \le \epsilon_{An}\Big] + \sup_\theta \mathbb{P}\big[\mathscr{Q}_n > \epsilon_{Qn}\big] + \sup_\theta \mathbb{P}\Big[\max_i |\mathscr{A}_i| > \epsilon_{An}\Big] \\ &\le 2\exp\Big(-\frac{\epsilon_n^2/2}{\epsilon_{Qn} + \epsilon_{An}\epsilon_n}\Big) + o\big(\mathscr{P}_n^{-1}\big) + o\big(\mathscr{P}_n^{-1}\big) = o\big(\mathscr{P}_n^{-1}\big), \end{aligned} \tag{57}$$

where the last inequality in (57) follows from theorem 1.2A of de la Peña (1999) plus (55), (56), and the equality by the definition of ϵ_n. So (54) holds. □

E.2. **Specifics.** Let below $\mathscr{N} = \{\theta : \|\theta - \theta_0\| \leq \|\theta_n - \theta_0\|\}$.

Lemma E5. $\sum_{ij} w_{ij} g_j(\theta_n) m_i(\theta_n) = \sum_{ij} w_{ij} g_j m_i + \sqrt{n}\lambda \mathbb{E}\tilde{g}_1 \tilde{g}_1^\intercal \Delta + o_p(\rho)$.

Proof. Using the mean value theorem,

$$\begin{aligned}
&\Big\|\sum_{ij} w_{ij}\big(g_j(\theta_n)m_i(\theta_n) - g_j m_i\big) - \sqrt{n}\lambda \mathbb{E}\tilde{g}_1\tilde{g}_1^\intercal\Delta\Big\| \\
&\qquad \leq \frac{\|\Delta\|}{\sqrt{n}} \sup_{\mathscr{N}} \Big\|\sum_{ij} w_{ij}\tilde{M}_{\theta\theta j}(\theta)u_i(\theta)\Big\| + \frac{\|\Delta\|}{\sqrt{n}} \sup_{\mathscr{N}} \Big\|\sum_{ij} w_{ij}\tilde{g}_j(\theta)v_i^\intercal(\theta)\Big\| \\
&\qquad + \frac{\|\Delta\|}{\sqrt{n}} \sup_{\mathscr{N}} \Big\|\sum_{ij} w_{ij}\tilde{g}_j(\theta)M_i(\theta)\Big\| + \frac{\|\Delta\|\lambda}{\sqrt{n}} \sup_{\mathscr{N}} \Big\|\sum_{ij} w_{ij}\tilde{g}_j(\theta)\tilde{g}_i^\intercal(\theta) - n\mathbb{E}\tilde{g}_1\tilde{g}_1^\intercal\Big\|. \qquad (58)
\end{aligned}$$

RHS1 and RHS2 in (58) are $o_p(\log n) = o_p(\rho)$ by lemma E3. By the mean value theorem RHS3 is bounded by $\|\Delta\|^2 n^{-1} \sup_{\theta,\theta^* \in \mathscr{N}} \big\|\sum_{ij} w_{ij}\tilde{g}_j(\theta)\tilde{g}_i^\intercal(\theta^*)\big\| = O_p(1) = o_p(\rho)$. Finally RHS4 is by lemma E1 equal to $o_p(\sqrt{n}\lambda) = o_p(\rho)$. □

Lemma E6. $\sum_{ij} w_{ij} v_j(\theta_n) m_i(\theta_n) - \sum_{ij} w_{ij} v_j m_i = o_p(\rho)$.

Proof. The norm of the LHS is by the mean value theorem bounded by

$$\begin{aligned}
&\frac{\|\Delta\|}{\sqrt{n}\lambda} \sup_{\mathscr{N}} \Big\|\sum_{ij} w_{ij}\big(m_{\theta\theta j}(\theta) - M_{\theta\theta j}(\theta)\big)u_i(\theta)\Big\| + \frac{\|\Delta\|}{\sqrt{n}\lambda} \sup_{\mathscr{N}} \Big\|\sum_{ij} w_{ij} v_j(\theta) v_i^\intercal(\theta)\Big\| \\
&\qquad + \frac{\|\Delta\|}{\sqrt{n}} \sup_{\mathscr{N}} \Big\|\sum_{ij} w_{ij}\big(m_{\theta\theta j}(\theta) - M_{\theta\theta j}(\theta)\big)\tilde{M}_i(\theta)\Big\| + \frac{\|\Delta\|}{\sqrt{n}} \sup_{\mathscr{N}} \Big\|\sum_{ij} w_{ij} v_j(\theta)\tilde{g}_i^\intercal(\theta)\Big\|. \qquad (59)
\end{aligned}$$

The first two terms in (59) are $o_p\big((\log n)^2/(\sqrt{k}\lambda)\big) = o_p\big((\log n)^2\big) = o_p(\rho)$ by lemma E4. The last two terms in (59) are $o_p(\log n) = o_p(\rho)$ by lemma E3, making minor adjustments for the fact that a, b are interchanged. □

Lemma E7. $\sum_i \hat{g}_i(\theta_n)\hat{g}_i^\intercal(\theta_n) m_i^2(\theta_n) = \sum_i \hat{g}_i \hat{g}_i^\intercal m_i^2 + o_p(\rho)$.

Proof. By the mean value theorem,

$$\begin{aligned}
\Big\|\sum_i \hat{g}_i(\theta_n)\hat{g}_i^\intercal(\theta_n) m_i^2(\theta_n) - \sum_i \hat{g}_i\hat{g}_i^\intercal m_i^2\Big\| &\leq \frac{\|\Delta\|}{\sqrt{n}\lambda} \sup_{\mathscr{N}} \sum_i \|\hat{g}_{\theta i}(\theta)\| \|\hat{g}_i(\theta)\| m_i^2(\theta) \\
&+ \frac{\|\Delta\|}{\sqrt{n}\lambda} \sup_{\mathscr{N}} \sum_i \|\hat{g}_i(\theta)\| \|\hat{g}_{\theta i}(\theta)\| m_i^2(\theta) + \frac{2\|\Delta\|}{\sqrt{n}\lambda} \sup_{\mathscr{N}} \Big\|\sum_i \hat{g}_i(\theta)\hat{g}_i(\theta) m_{\theta i}(\theta) m_i(\theta)\Big\|. \qquad (60)
\end{aligned}$$

RHS1, RHS2, and RHS3 in (60) are similar, and we only obtain a result for RHS3, which is bounded by

$$\frac{2\|\Delta\|\lambda}{\sqrt{n}} \sum_i \Big\|\sum_j w_{ij}\tilde{g}_j(\theta)\Big\|^2 \|m_{\theta i}(\theta) m_i(\theta)\| + \frac{2\|\Delta\|}{\sqrt{n}\lambda} \sum_i \Big\|\sum_j w_{ij} v_j(\theta)\Big\|^2 \|m_{\theta i}(\theta) m_i(\theta)\|. \qquad (61)$$

The first term in (61) is $O_p(\sqrt{n}\lambda) = o_p(\rho^2)$. For the second term, note that by arguments similar to (but simpler than) those made in lemma E3, $\sup_{\mathscr{N}} \max_i \big\|\sum_j w_{ij} v_j\big\| = o_p\big((\log n)/\sqrt{k}\big)$. Hence the second term in (61) is $o_p\big((\log n)^2\sqrt{n}/(k\lambda)\big) = o_p(\rho^2)$. □

Lemma E8. $\sum_{ij} w_{ij}w_{ji}m_j(\theta_n)m_{\theta j}^\intercal(\theta_n)m_i(\theta_n)m_{\theta i}(\theta_n) - \sum_{ij} w_{ij}w_{ji}m_j m_{\theta j}^\intercal m_i m_{\theta i} = o_p(\rho^2)$.

Proof. By the mean value theorem,

$$\begin{aligned}
&\Big\|\sum_{ij} w_{ij}w_{ji}m_j(\theta_n)m_{\theta j}^\intercal(\theta_n)m_i(\theta_n)m_{\theta i}(\theta_n) - \sum_{ij} w_{ij}w_{ji}m_j m_{\theta j}^\intercal m_i m_{\theta i}\Big\| \\
&\quad\le \frac{\|\Delta\|}{\sqrt{n}\lambda}\sup_{\mathcal{N}}\Big\{\sum_{ij} w_{ij}w_{ji}\|m_{\theta j}(\theta)\|^2\|m_i(\theta)m_{\theta i}(\theta)\| + \sum_{ij} w_{ij}w_{ji}\|m_j(\theta)m_{\theta\theta}(\theta)\|\|m_i(\theta)m_{\theta i}(\theta)\| \\
&\qquad + \sum_{ij} w_{ij}w_{ji}\|m_j(\theta)m_{\theta j}(\theta)\|\|m_{\theta i}(\theta)\|^2 + \sum_{ij} w_{ij}w_{ji}\|m_j(\theta)m_{\theta j}(\theta)\|\|m_i(\theta)m_{\theta\theta i}(\theta)\|\Big\} \\
&\qquad = O_p\Big(\frac{\sqrt{n}}{k\lambda}\Big) = o_p(\rho) = o_p(\rho^2). \qquad \square
\end{aligned}$$

Appendix F. Nuisance Parameters

Lemma F1. *If f, f^* are such that for a neighborhood $\aleph$ of $\tilde\beta$, $\mathbb{E}\big(\mathbb{E}\big[\sup_{\beta\in\aleph}\|f_1(\theta,\beta)\|\,|z_1\big]\big)^2 < \infty$ and $\mathbb{E}\big(\mathbb{E}\big[\sup_{\beta\in\aleph}\|f_1^*(\theta,\beta)\|\,|z_1\big]\big)^2 < \infty$, then for any β^* converging to $\tilde\beta$, $\sum_{ij} w_{ij}f_i^*(\theta,\beta^*)f_j(\theta,\beta^*) = O_p(n)$.*

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

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

by lemma B5. $\square$

Lemma F2. $\hat\beta(\theta) \xrightarrow{p} \beta(\theta)$.

Proof. Since $\mathbb{E}[h_1 m_1]$ is continuous and yields a unique solution $\beta(\theta)$ by assumption D, we are left to show uniform convergence (in β at θ) of $n^{-1}\sum_i \hat h_i m_i$ to $\mathbb{E}[h_1 m_1]$. We first show pointwise convergence and then stochastic equicontinuity. By lemma B6(vi) at fixed θ, β,

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

Stochastic equicontinuity then follows from the fact that

$$\begin{aligned}
&\sup_{\tilde\beta}\sup_{\beta:||\beta-\tilde\beta||\le\delta}\Big\|n^{-1}\sum_i[\hat h_i(\theta,\tilde\beta)m_i(\theta,\tilde\beta) - \hat h_i(\theta,\beta)m_i(\theta,\beta)]\Big\| \\
&\qquad \overset{\text{M.V.T.}}{\le} \Big(n^{-1}\sum_{ij} w_{ij}\sup_\beta||m_{\beta\beta j}(\theta,\beta)||\sup_\beta|m_i(\theta,\beta)| \\
&\qquad + n^{-1}\sum_{ij} w_{ij}\sup_\beta||m_{\beta j}(\theta,\beta)||\sup_\beta|m_{\beta i}(\theta,\beta)|\Big) \times \sup_{\tilde\beta}\sup_{\beta:||\beta-\tilde\beta||\le\delta}||\tilde\beta-\beta|| \overset{\text{B5,E}}{=} O_p(\delta).
\end{aligned}$$

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

Lemma F3. $\hat{\beta}(\theta) - \beta(\theta) = \begin{cases} O_p(n^{-1/2}), & \theta = \theta_0, \\ O_p(n^{-1/2}) + o_p(\lambda), & \theta \neq \theta_0. \end{cases}$.

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

$$\left(\hat{\beta}(\theta) - \beta(\theta)\right)^\intercal \sum_i \Big(\hat{h}_{\beta\beta it}(\theta, \beta^*) m_i(\theta, \beta^*) + 2\hat{h}_{\beta it}(\theta, \beta^*) m_{\beta i}^\intercal(\theta, \beta^*) + \hat{h}_{it}(\theta, \beta^*) m_{\beta\beta i}(\theta, \beta^*) \Big) \left(\hat{\beta}(\theta) - \beta(\theta)\right)/2,$$

where $\hat{h}_{it}$ is the t–th element of $\hat{h}_i$ and $\hat{h}_{\beta it}, \hat{h}_{\beta\beta it}$ are its first and second partial derivatives with respect to β. Then by lemmas F1 and F2, $S_n(\theta, \beta^*) = O_p(n||\hat{\beta}(\theta) - \beta(\theta)||^2)$ for any β^* between $\hat{\beta}(\theta)$ and $\beta(\theta)$. Hence by the mean value theorem (omitting θ from hereon),

$$0 = \sum_i \hat{\tilde{h}}_i \hat{m}_i = \sum_i \hat{h}_i m_i + \sum_i m_i \hat{h}_{\beta i}^\intercal (\hat{\beta} - \beta) + \sum_i \hat{h}_i m_{\beta i}^\intercal (\hat{\beta} - \beta) + O_p(n||\hat{\beta} - \beta||^2). \tag{62}$$

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

$$\sum_i \hat{h}_i m_i = \underbrace{\sum_i \hat{h}_i (m_i - M_i)}_{\overset{\text{B6(ii)}}{=} O_p(\sqrt{n})} + \underbrace{\lambda \sum_i (\hat{h}_i - h_i)\tilde{M}_i}_{\overset{\text{B6(vi)}}{=} o_p(n\lambda)} + \underbrace{\lambda \sum_i h_i \tilde{M}_i}_{= O_p(\sqrt{n}\lambda)} = \begin{cases} O_p(\sqrt{n}), & \theta = \theta_0, \\ O_p(\sqrt{n}) + o_p(n\lambda), & \theta \neq \theta_0. \end{cases} \tag{63}$$

Further,

$$\sum_i \hat{h}_{\beta i} m_i = \underbrace{\sum_i \hat{h}_{\beta i}(m_i - M_i)}_{\overset{\text{B6(ii)}}{=} O_p(\sqrt{n})} + \underbrace{\lambda \sum_i (\hat{h}_{\beta i} - h_{\beta i})\tilde{M}_i}_{\overset{\text{B6(vi)}}{=} o_p(n\lambda)} + \underbrace{\lambda \sum_i (h_{\beta i}\tilde{M}_i - \mathbb{E}[h_{\beta 1}\tilde{M}_1])}_{= O_p(\sqrt{n}\lambda)} + n\mathbb{E}[h_{\beta 1} M_1]$$

$$= nE[m_{\beta\beta 1} M_1] + \begin{cases} O_p(\sqrt{n}), & \theta = \theta_0, \\ O_p(\sqrt{n}) + o_p(n\lambda), & \theta \neq \theta_0. \end{cases} \tag{64}$$

And

$$\sum_i \hat{h}_i m_{\beta i}^\intercal \overset{\text{B6(vi)}}{=} \sum_i h_i h_i^\intercal + o_p(n) = n\mathbb{E}[h_1 h_1^\intercal] + o_p(n). \tag{65}$$

Plugging (63)–(65) into (62) we get

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

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

Lemma F4. $\sum_i \hat{\tilde{q}}_i(\theta)\hat{m}_i(\theta) = \sum_i \hat{q}_i(\theta) m_i(\theta) + \begin{cases} o_p(\rho), & \theta = \theta_0, \\ O_p(\rho) + o_p(n\lambda^2), & \theta \neq \theta_0. \end{cases}$

Proof. Since we can consider each element of $\hat{\tilde{q}}_i$ separately, we may assume that $\hat{\tilde{q}}_i$ is a scalar without loss of generality. Note then that for any β^* between $\hat{\beta}(\theta)$ and $\beta(\theta)$,

$$\Big|\Big|(\hat{\beta}(\theta)-\beta(\theta))^{\intercal}\sum_i\Big[\hat{g}_{\beta\beta i}(\theta,\beta^*)m_i(\theta,\beta^*)+2\hat{g}_{\beta i}(\theta,\beta^*)m_{\beta i}^{\intercal}(\theta,\beta^*) + \hat{g}_i(\theta,\beta^*)m_{\beta\beta i}(\theta,\beta^*)\Big](\hat{\beta}(\theta)-\beta(\theta))\Big|\Big| \overset{\text{F1}}{=} O_p(n||\hat{\beta}(\theta)-\beta(\theta)||^2).$$

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

$$\begin{aligned}\sum_i\hat{q}_i\hat{m}_i=\sum_i\hat{g}_i\hat{m}_i&=\sum_i\hat{g}_i\big[m_i+m_{\beta i}^{\intercal}(\hat{\beta}-\beta)\big]+\sum_i m_i\hat{g}_{\beta i}^{\intercal}(\hat{\beta}-\beta)+O_p(n||\hat{\beta}-\beta||^2)\\&=\sum_i\hat{q}_i m_i+\sum_i\hat{q}_i m_{\beta i}^{\intercal}(\hat{\beta}-\beta)+\kappa^{\intercal}\sum_i\hat{h}_i\big[m_i+m_{\beta i}^{\intercal}(\hat{\beta}-\beta)\big]\\&\quad+\sum_i m_i\hat{g}_{\beta i}^{\intercal}(\hat{\beta}-\beta)+\begin{cases}O_p(1), & \theta=\theta_0,\\ O_p(1)+o_p(n\lambda^2), & \theta\neq\theta_0.\end{cases}\end{aligned}\tag{66}$$

We deal with each of the RHS2–RHS4 terms separately. For RHS3 in (66) note that by (62),

$$\begin{aligned}\Big|\Big|\sum_i\hat{h}_i(m_i+m_{\beta i}^{\intercal}(\hat{\beta}-\beta))\Big|\Big|&\leq\Big|\Big|\sum_i m_i\hat{h}_{\beta i}\Big|\Big|\,||\hat{\beta}-\beta||+O_p(n||\hat{\beta}-\beta||^2)\\&\overset{\text{F3, (64)}}{=}\begin{cases}O_p(n^{1/2})O_p(n^{-1/2})+O_p(1), & \theta=\theta_0,\\ \{O_p(n^{1/2})+o_p(n\lambda)\}\{O_p(n^{-1/2})+o_p(\lambda)\}+O_p(1)+o_p(n\lambda^2), & \theta\neq\theta_0.\end{cases},\end{aligned}$$

which is $\begin{cases}O_p(1), & \theta=\theta_0,\\ o_p(\rho+n\lambda^2), & \theta\neq\theta_0.\end{cases}$. For RHS4 in (66) we have

$$\sum_i\hat{g}_{\beta i}m_i=\underbrace{\sum_i\hat{g}_{\beta i}(m_i-M_i)}_{\overset{\text{B6(ii)}}{=}O_p(\sqrt{n})}+\underbrace{\lambda\sum_i(\hat{g}_{\beta i}-g_{\beta i})\tilde{M}_i}_{\overset{\text{B6(vi)}}{=}o_p(n\lambda)}+\underbrace{\lambda\sum_i g_{\beta i}\tilde{M}_i}_{=O_p(n\lambda)}=\begin{cases}O_p(\sqrt{n}), & \theta=\theta_0,\\ O_p(\sqrt{n}+n\lambda), & \theta\neq\theta_0.\end{cases}\tag{67}$$

Now combine (67) with lemma F3 to obtain rates for RHS4 in (66) of $O_p(1)$ if $\theta=\theta_0$ and $O_p(\rho)+o_p(n\lambda^2)$ if $\theta\neq\theta_0$. Finally RHS2 in (66). We have

$$\begin{aligned}\sum_i\hat{q}_i m_{\beta i}&\overset{\text{B6(ii)}}{=}\sum_i\hat{q}_i h_i+O_p(\sqrt{n})\overset{\mathbb{E}[q_1h_1]=0}{=}\sum_i(\hat{q}_i-q_i)h_i+O_p(\sqrt{n})\\&=\underbrace{\lambda\sum_{ij}w_{ij}(\tilde{q}_j-\tilde{q}_i)h_i}_{\overset{\text{B6(vi)}}{=}o_p(n\lambda)}+\underbrace{\sum_{ij}w_{ij}(m_{\theta j}^{\intercal}-\kappa^{\intercal}m_{\beta j}-q_j)h_i}_{\overset{\text{B6(iv)}}{=}o_p(n^{3/4}k^{-1/4})}+O_p(\sqrt{n})\\&=O_p(\sqrt{n})+o_p(\sqrt{n\rho})+o_p(n\lambda),\end{aligned}$$

such that RHS2 in (66) is $\begin{cases}o_p(\rho), & \theta=\theta_0,\\ o_p(\rho+n\lambda^2), & \theta\neq\theta_0.\end{cases}$ □

Lemma F5. $\hat{\kappa}(\theta)-\kappa(\theta)=o_p(\rho/\sqrt{n})$.

Proof. Note that (omitting θ)

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

Since we can consider each column of $\hat{\hat{\kappa}} - \kappa$ separately, we may assume that $\hat{\hat{g}}_i$ (and hence q_i) is a scalar without loss of generality. By Slutsky it then suffices to show that (i) $n^{-1}\sum_i \hat{\hat{h}}_i \hat{\hat{h}}_i^\intercal - \mathbb{E}[h_1 h_1^\intercal] = o_p(1)$ and (ii) $n^{-1}\sum_i \hat{\hat{h}}_i(\hat{\hat{g}}_i - \hat{\hat{h}}_i^\intercal \kappa) = o_p(\rho/\sqrt{n})$. Establishing (i) is similar to but simpler than showing (ii) so we only show (ii) here. We show that

$$\sum_i [\hat{\hat{h}}_i(\hat{\hat{g}}_i - \hat{\hat{h}}_i^\intercal \kappa) - \hat{h}_i \hat{q}_i)] = o_p(\sqrt{n}\rho), \tag{68}$$

$$\sum_i (\hat{h}_i \hat{q}_i - h_i q_i) = o_p(\sqrt{n}\rho), \tag{69}$$

$$\sum_i h_i q_i = o_p(\sqrt{n}\rho). \tag{70}$$

Since $\mathbb{E}[h_1 q_1] = 0$ by construction, the LHS in (70) is $O_p(\sqrt{n}\lambda)$. For (69) note that

$$\sum_i (\hat{h}_i \hat{q}_i - h_i q_i) = \sum_i (\hat{h}_i - h_i)(\hat{q}_i - q_i) + \sum_i (\hat{h}_i - h_i) q_i + \sum_i h_i(\hat{q}_i - q_i). \tag{71}$$

First RHS3 in (71). Let $\epsilon_i = m_{\theta i}^\intercal - \kappa^\intercal m_{\beta i} - q_i$. Then

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

But

$$\mathbb{E}\left|\left|\sum_{ij} w_{ij} h_i \epsilon_j\right|\right|^2 \leq \sum_{ij} \mathbb{E}[w_{ij}^2 ||h_i||^2 ||\epsilon_j||^2] + \sum_{ij}\sum_{t\neq i} \mathbb{E}[w_{ij} w_{tj} ||h_i||\,||h_t|| \,||\epsilon_j||^2] \overset{B5}{=} o(n^{3/2}k^{-1/2}) = o(n\rho).$$

The derivation for RHS2 in (71) is similar to that for RHS3 and is omitted. By the Schwarz inequality, sufficient conditions for RHS1 in (71) to be $o_p(\sqrt{n}\rho)$ are that (a) $\sum_i ||\hat{h}_i - h_i||^2 = o_p(n)$ and (b) $\sum_i ||\hat{q}_i - q_i||^2 = O_p(\rho^2)$. First (b). Since $\hat{q}_i - q_i = \sum_j w_{ij}(q_j - q_i + \epsilon_j)$, we have

$$\begin{aligned} 2^{-1}\sum_i \mathbb{E}||\hat{q}_i - q_i||^2 &\leq \sum_i \mathbb{E}\left\|\sum_j w_{ij}(q_j - q_i)\right\|^2 + \sum_i \mathbb{E}\left\|\sum_j w_{ij}\epsilon_j\right\|^2 \\ &\overset{\text{Schwarz}}{\leq} n^2\lambda^2 \mathbb{E}\left[w_{12}||\tilde{q}_2 - \tilde{q}_1||^2\right] + n^2 \mathbb{E}[w_{12}^2 ||\epsilon_2||^2] \overset{B1,B5}{=} o(n\lambda^2) + O(n/k) = O(\rho^2). \end{aligned}$$

Requirement (a) follows with the same derivation resulting in a rate of $o(n) + O(n/k) = o(n)$. Finally apply the mean value theorem to the LHS in (68) and obtain for some β^* between $\hat{\beta}$ and β

$$\sum_i \left[\hat{h}_{\beta i}(\hat{g}_i - \hat{h}_i^\intercal \kappa) + \hat{h}_i(\hat{g}_{\beta i} - \hat{h}_{\beta i}^\intercal \kappa)^\intercal\right](\hat{\beta} - \beta) \quad \text{all at } (\theta, \beta^*). \tag{72}$$

By lemma F3, $||\hat{\beta} - \beta|| = o_p(\rho/\sqrt{n})$. It is hence sufficient to show that the sum in (72) is $O_p(n)$, which follows from repeated application of lemma F1. □

Lemma F6. $\sum_i \hat{\hat{q}}_i(\theta)\hat{\hat{q}}_i^\intercal(\theta)\hat{\hat{m}}_i^2(\theta) = \sum_i \hat{q}_i(\theta)\hat{q}_i^\intercal(\theta) m_i^2(\theta) + o_p(\rho^2)$.

Proof. Since for generic a, b, $\|aa^\intercal - bb^\intercal\| \leq \|a-b\|^2 + 2\|b\|\|a-b\|$, and $\sum_i \|\hat{q}_i m_i\|^2 = O_p(\rho^2)$ by lemmas C22 and C24 (substituting $\hat{q}$ for $\hat{g}$), we need to show that $\sum_i \|\hat{\hat{q}}_i \hat{m}_i - \hat{q}_i m_i\|^2 = o_p(\rho^2)$. Since we can consider each element of $\hat{\hat{q}}_i \hat{m}_i - \hat{q}_i m_i$ separately, we may assume that $\hat{\hat{q}}_i$ and $\hat{q}_i$ are scalars without loss of generality, and we only need to show $\sum_i (\hat{\hat{q}}_i \hat{m}_i - \hat{q}_i m_i)^2 = o_p(\rho)$. Now,

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

By lemma F5 it hence suffices to show that

$$\sum_i (\hat{\hat{g}}_i \hat{m}_i - \hat{g}_i m_i)^2 = o_p(\rho^2), \quad \sum_i ||\hat{\hat{h}}_i \hat{m}_i - \hat{h}_i m_i||^2 = o_p(\rho^2), \quad \sum_i ||\hat{h}_i m_i||^2 = O_p(n). \tag{73}$$

The third condition in (73) follows from lemma B6. Now the first condition. We have for some β^* between $\hat{\beta}$ and β,

$$\begin{aligned}\sum_i (\hat{\hat{g}}_i \hat{m}_i - \hat{g}_i m_i)^2 &= \sum_i \Big(\sum_j w_{ij} (\hat{m}_{\theta j} \hat{m}_i - m_{\theta j} m_i) \Big)^2 \\ &\leq ||\hat{\beta} - \beta||^2 \sum_{ijt} w_{ij} w_{it} \big| \big| m_t(\theta, \beta^*) m_{\theta\beta j}(\theta, \beta^*) + m_{\theta j}(\theta, \beta^*) m_{\beta t}(\theta, \beta^*) \big| \big|. \end{aligned} \tag{74}$$

Now since for generic a_i, b_i

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

it follows from lemmas F1 and F3 that the RHS in (74) is $o_p(\rho^2/n) O_p(n) = o_p(\rho^2)$. So we have established the first condition in (73), where the derivation for the second condition follows similarly. □

Lemma F7. $\sum_{ij} w_{ij} w_{ji} \hat{m}_j(\theta) \hat{m}_{\theta j}^\intercal(\theta) \hat{m}_i(\theta) \hat{m}_{\theta i}(\theta) = \sum_{ij} w_{ij} w_{ji} m_j(\theta) m_{\theta j}^\intercal(\theta) m_i(\theta) m_{\theta i}(\theta) + o_p(\rho^2)$.

Proof. By the mean value theorem, we have

$$\begin{aligned}&\Big| \Big| \sum_{ij} w_{ij} w_{ji} (\hat{m}_j \hat{m}_{\theta j}^\intercal \hat{m}_i \hat{m}_{\theta i}^\intercal - m_j m_{\theta j}^\intercal m_i m_{\theta i}) \Big| \Big| \\ &\qquad \leq (C_w/k) ||\hat{\beta} - \beta|| \sum_{ij} w_{ij} \big| \big| m_{\beta j}(\theta, \beta^*) m_{\theta j}^\intercal(\theta, \beta^*) m_i(\theta, \beta^*) m_{\theta i}(\theta, \beta^*) + \ldots + \\ &\qquad\qquad m_j(\theta, \beta^*) m_{\theta j}^\intercal(\theta, \beta^*) m_i(\theta, \beta^*) m_{\theta\beta i}(\theta, \beta^*) \big| \big|. \end{aligned}$$

Apply the triangle inequality and lemmas F1 and F3 to obtain a rate of $o_p(\rho/\sqrt{n}) O_p(n/k) = o_p(\rho^2)$. □

Lemma F8. $\hat{\hat{\mathscr{D}}}_N^2(\theta) = \hat{\mathscr{D}}_N^2(\theta) + o_p(\rho^2)$.

Proof. By lemmas F4, F6 and F7 (omitting θ),

$$\hat{\hat{\mathscr{D}}}_N^2 - \hat{\mathscr{D}}_N^2 = \underbrace{\Big[\sum_i \hat{\hat{q}}_i \hat{\hat{q}}_i^\intercal \hat{m}_i^2 - \sum_i \hat{q}_i \hat{q}_i^\intercal m_i^2 \Big]}_{\overset{\text{F6}}{=} o_p(\rho^2)} - n^{-1} \Big[\Big(\sum_i \hat{\hat{q}}_i \hat{m}_i \Big) \Big(\sum_i \hat{\hat{q}}_i^\intercal \hat{m}_i \Big) - \Big(\sum_i \hat{q}_i m_i \Big) \Big(\sum_i \hat{q}_i^\intercal m_i \Big) \Big]$$

$$+\underbrace{\Big[\sum_{ij} w_{ij}w_{ji}\hat{m}_j\hat{m}_{\theta j}^{\intercal}\hat{m}_i\hat{m}_{\theta i} - \sum_{ij} w_{ij}w_{ji}m_j m_{\theta j}^{\intercal} m_i m_{\theta i}\Big]}_{\overset{\text{F7}}{=} o_p(\rho^2)} \tag{75}$$

Finally, the norm of RHS2 in (75) is bounded by

$$\underbrace{n^{-1}\Big\|\sum_i(\hat{\hat{q}}_i\hat{m}_i - \hat{q}_i m_i)\Big\|^2}_{\overset{\text{F4}}{=} O_p(\rho^2/n)+o_p(n\lambda^4)} + 2n^{-1}\underbrace{\Big\|\sum_i \hat{q}_i m_i\Big\|}_{=O_p(\rho)}\underbrace{\Big\|\sum_i(\hat{\hat{q}}_i\hat{m}_i - \hat{q}_i m_i)\Big\|}_{\overset{\text{F4}}{=} O_p(\rho)+o_p(n\lambda^2)} = o_p(\rho^2). \qquad \square$$

Appendix G. Justification of example I

We first justify that

$$\begin{bmatrix} n^{-1/2}\sum_i \tilde{g}_i u_i \\ n^{-1/2}\sum_i \tilde{g}_i v_i \\ (n/k)^{-1/2}\sum_{ij} w_{ij}v_j u_i \\ (n/k)^{-1/2}\sum_{ij} w_{ij}v_j v_i \end{bmatrix} \xrightarrow{d} \begin{bmatrix} \Psi_{\tilde{g}u} \\ \Psi_{\tilde{g}v} \\ \Psi_{vu} \\ \Psi_{vv} \end{bmatrix} \sim N(0, V_\Psi), \tag{76}$$

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

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

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

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

$$n^{-1/2}\lambda^{-1}\hat{N}_I = n^{-1/2}\sum_i \tilde{g}_i u_i, \tag{77}$$

$$\begin{aligned} n^{-1/2}\lambda^{-1}\hat{D}_I = n^{-1/2}\sum_i \tilde{g}_i Y_i &= \sqrt{n}\lambda n^{-1}\sum_i \tilde{g}_i^2 + n^{-1/2}\sum_i \tilde{g}_i v_i \\ &= (\sqrt{n}\lambda)\mathbb{E}\tilde{g}_i^2 + n^{-1/2}\sum_i \tilde{g}_i v_i + O_p(\lambda), \end{aligned} \tag{78}$$

where we used the fact that $n^{-1}\sum_i \tilde{g}_i^2 = \mathbb{E}[\tilde{g}_i^2] + O_p(n^{-1/2})$. The result stated in (3) then follows directly from combining (76)–(78) and applying the continuous mapping theorem. We now turn to $\hat{\theta} - \theta_0 = \hat{N}/\hat{D}$, for which

$$\hat{N} = \sum_i \hat{g}_i u_i = \sum_{ij} w_{ij}Y_j u_i = \lambda\sum_{ij} w_{ij}\tilde{g}_j u_i + \sum_{ij} w_{ij}v_j u_i = \lambda\sum_{ij} w_{ij}(\tilde{g}_j - \tilde{g}_i)u_i$$

$$+\lambda\sum_i \tilde{g}_i u_i + \sum_{ij} w_{ij} v_j u_i = \underbrace{\lambda\sum_i \tilde{g}_i u_i}_{O_p(\sqrt{n}\lambda)} + \underbrace{\sum_{ij} w_{ij} v_j u_i}_{O_p(\sqrt{n/k})} + o_p(\sqrt{n}\lambda). \quad (79)$$

Similarly,

$$\hat{D} = \sum_i \hat{g}_i Y_i = \lambda\sum_i \hat{g}_i\tilde{g}_i + \sum_i \hat{g}_i v_i = \underbrace{\lambda^2\sum_i \tilde{g}_i^2}_{O_p(n\lambda^2)} + \underbrace{\lambda^2\sum_{ij} w_{ij}(\tilde{g}_j - \tilde{g}_i)\tilde{g}_i}_{o_p(n\lambda^2)} + \underbrace{\lambda\sum_{ij} w_{ij}\tilde{g}_i v_j}_{o_p(n\lambda^2+\sqrt{n}\lambda+\sqrt{n/k})}$$
$$+ \underbrace{\lambda\sum_i \tilde{g}_i v_i}_{O_p(\sqrt{n}\lambda)} + \underbrace{\lambda\sum_{ij} w_{ij}(\tilde{g}_j - \tilde{g}_i)v_i}_{o_p(\sqrt{n}\lambda)} + \underbrace{\sum_{ij} w_{ij} v_i v_j}_{O_p(\sqrt{n/k})} \quad (80)$$

For RHS3 in (80), note that by lemma B5,

$$\mathbb{E}\Big(\sum_{ij} w_{ij}\tilde{g}_i v_j\Big)^2 = \sum_{ijt}\mathbb{E}[w_{ij}w_{tj}\tilde{g}_i\tilde{g}_t v_j^2] = o(n^{3/2}k^{-1/2}),$$

which implies

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

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

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

The dominant numerator terms converge at the same rate if $\lambda \sim 1/\sqrt{k}$ and the denominator terms when $\lambda \sim 1/\sqrt[4]{nk}$. Combining (79) and (80) with (76) and applying the continuous mapping theorem for each of the three cases in (4) gives the results stated therein. □