Available online at www.sciencedirect.com
SCIENCE DIRECT®

Journal of Econometrics 117 (2003) 55–93

JOURNAL OF Econometrics

www.elsevier.com/locate/econbase

Empirical likelihood estimation and consistent tests with conditional moment restrictions☆

Stephen G. Donald[a], Guido W. Imbens[b], Whitney K. Newey[c,*]

[a] Department of Economics, University of Texas, TX, USA
[b] Department of Economics, UCLA, USA
[c] Department of Economics, MIT, E52-262D, Cambridge, MA, USA

Accepted 6 February 2003

Abstract

This paper is about efficient estimation and consistent tests of conditional moment restrictions. We use unconditional moment restrictions based on splines or other approximating functions for this purpose. Empirical likelihood estimation is particularly appropriate for this setting, because of its relatively low bias with many moment conditions. We give conditions so that efficiency of estimators and consistency of tests is achieved as the number of restrictions grows with the sample size. We also give results for generalized empirical likelihood, generalized method of moments, and nonlinear instrumental variable estimators.

JEL classification: C13; C30

Keywords: Nonlinear instrumental variables; Generalized method of moments; Generalized empirical likelihood; Consistent specification testing

1. Introduction

Models with conditional moment restrictions are important in econometrics. These models arise in many econometric settings, including rational expectations, panel data, and instrumental variable settings. This paper is about efficient estimation of parameters of these models and consistent tests of their restrictions. We construct these estimators

☆ This research was partially completed while the third author was a fellow at the Center for Advanced Study in the Behavioral Sciences. The NSF provided partial financial support, as did the Alfred P. Sloan Foundation for the first author. T. Amemiya and two referees provided useful comments.
* Corresponding author. Tel.: 617-253-6420; fax: 617-253-1130.
E-mail address: wnewey@mit.edu (W.K. Newey).

0304-4076/03/$ - see front matter
doi:10.1016/S0304-4076(03)00118-0

and tests using general approximating functions, such as splines, to form unconditional moments that grow in number and variety with the sample size.

Empirical likelihood (EL) (Owen, 1988; Qin and Lawless, 1994; Imbens, 1997) is particularly interesting in our setting where there may be many moments used in estimation. As shown by Newey and Smith (2002), higher order asymptotic bias of these estimators is smaller than GMM with many moments, which should lead to better asymptotic approximations to finite sample distributions, and bias-corrected EL is higher-order efficient among bias-corrected estimators. Thus, in cases where bias is a concern, such as models with endogeneity and covariance models for panel data, EL has good theoretical properties. Here we give explicit limits on the growth rate for the number of moments for asymptotic efficiency of the estimator and for conditional moment tests to be asymptotically normal. In particular, for B-splines where the density of the conditioning variables is bounded and bounded away from zero we find that $K^3/n \to 0$ suffices to obtain asymptotic efficiency of EL and $K^4/n \to 0$ to obtain asymptotic normality of the EL overidentification test statistic, where K is the number of approximating functions used in estimation. We also give analogous results for generalizations of empirical likelihood that include the continuous updating estimator of Hansen et al. (1996) and the exponential tilting estimator of Imbens et al. (1998) and Kitamura and Stutzer (1997).

Other estimators may also be useful in imposing many moment conditions. The two-step generalized method of moments (GMM) (Hansen, 1982) estimator is computationally simpler than the EL estimator and is widely used. Also, the GMM overidentification test statistic provides a simple test of the conditional moment restrictions. We show that under slightly weaker conditions than for EL, $K^2/n \to 0$ suffices for efficiency of the estimator and $K^3/n \to 0$ for asymptotic normality of the overidentification test statistic, with regression splines. Nonlinear instrumental variables estimation (IV) (Amemiya, 1974, 1977) is also useful. It is efficient under the auxiliary assumption of homoskedasticity, and is known to have better small sample properties than GMM in some settings (e.g. see Arellano and Bond, 1991 for dynamic panel data). In the homoskedastic case this estimator requires the weakest regularity conditions for efficiency. The marginal distribution of the conditioning variables is not restricted; it can even be discrete. The estimator will be efficient if $K^2/n \to 0$.

The asymptotic theory for EL is new. These results provide theory for a conditional version of empirical likelihood that is based on many unconditional moment restrictions. This approach is complementary to that of LeBlanc and Crowley (1995) and Kitamura et al. (2001). Both approaches produce an efficient estimator and a consistent empirical likelihood test of conditional moment restrictions. Our approach is computationally simpler, producing a smaller dimensional likelihood, but does not lead to an estimator of the conditional distribution of the data.

The testing theory is also new, although it is related to previous work. Bierens (1982) and Newey (1985) suggested testing conditional moment restrictions by many unconditional ones and de Jong and Bierens (1994) and Hong and White (1995) developed asymptotic theory for such tests. We provide a result that is more general than existing ones in several respects. It applies to models where there are endogenous right-hand side variables. It also has weaker rate restrictions than some in the literature, e.g. only requiring that $K^3/n \to 0$ for splines rather than $K^5/n \to 0$ as in Hong and White (1995). Also, the conditional moment test statistics have an interesting form. In addition to including the GMM overidentification statistic, they include the empirical log-likelihood ratio, and others. Each of these will have the same asymptotic distribution under the null hypothesis, and be consistent against violations of the conditional moment restrictions, although they may have very different finite sample distributions.

The asymptotic theory for IV and GMM estimators is closely related to previous results. Newey (1990) and Newey and Chipty (2000) give asymptotic efficiency results for linearized versions of IV and GMM, respectively, under exactly the same rate conditions we have. Our contribution is to obtain these results for the fully iterated IV and GMM estimators. One nice feature of these results is that they only use the minimal identification condition from the conditional moment restrictions. Theoretically speaking, these results "close the loop" by showing that for IV and GMM we get a $\sqrt{n}$-consistent estimator (which previously had just been assumed for initial estimators) under minimal conditions. Also, the asymptotic results for EL are based in part on those for IV and GMM. These results have been included in the body of the paper, rather than the appendix, because we thought they might be of some independent interest.

There is other previous work on asymptotic efficiency of linearized GMM. Newey (1993) and Hahn (1997) put much stronger restrictions on the growth of K but weaken the restrictions on the distribution of the instruments. Koenker and Machado (1999) give general results for a linear model and give primitive conditions for Fourier series. We obtain efficiency and consistent asymptotic variance estimation under $K^2/n \to 0$ rather than $K^3/n \to 0$. The consistency result for IV is similar to that of Newey and Powell (1989). The GMM consistency result under growing numbers of moments is new.

Section 2 of the paper will set up the model we consider and briefly discuss the virtues of different types of approximating functions for use in forming the moment conditions. Also, regularity conditions for the approximating functions are discussed. Section 3 describes empirical likelihood estimation and inference. Section 4 briefly reviews IV and GMM. Section 5 gives consistency, asymptotic normality, and asymptotic efficiency results for all of the estimators. Section 6 gives limiting distribution results for tests of conditional moment restrictions. All the proofs are given in the appendix.

2. Moment restrictions

The environment we consider is one where there is a vector of conditional moment restrictions depending on unknown parameters. To describe this setting let z denote a single observation, β a $p \times 1$ vector of parameters, and $\rho(z,\beta)$ a $J \times 1$ vector of functions, that often can be thought of as residuals. We specify that there is a subvector x of z, acting as conditioning variables, and a value of the parameters β_0 that satisfy

$$E[\rho(z,\beta_0)|x] = 0, \tag{2.1}$$

where E[.] denotes expectation taken with respect to the distribution of z.

It is well known that a conditional moment restriction is equivalent to a countable number of unconditional moment restrictions, under certain circumstances, see especially Bierens (1982) and Chamberlain (1987). We briefly discuss the completeness conditions for this equivalence in our setting. For each positive integer K let $q^K(x)=(q_{1K}(x),\dots,q_{KK}(x))'$ be a $K\times 1$ vector of approximating functions. We impose the following fundamental condition on the sequence $q^K(x)$ and the distribution of x:

Assumption 1. For all K, $\mathrm{E}[q^K(x)'q^K(x)]$ is finite, and for any $a(x)$ with $\mathrm{E}[a(x)^2]<\infty$ there are $K\times 1$ vectors γ_K such that as $K\to\infty$,

$$\mathrm{E}[\{a(x)-q^K(x)'\gamma_K\}^2]\to 0. \tag{2.2}$$

Lemma 2.1. *Suppose that Assumption* 1 *is satisfied and* $\mathrm{E}[\rho(z,\beta_0)'\rho(z,\beta_0)]$ *is finite. If Eq.* (2.1) *is satisfied then* $\mathrm{E}[\rho(z,\beta_0)\otimes q^K(x)]=0$ *for all* K*. Furthermore, if Eq.* (2.1) *is not satisfied then* $\mathrm{E}[\rho(z,\beta_0)\otimes q^K(x)]\neq 0$ *for all* K *large enough.*

The consequence of this result is that the conditional moment restriction is equivalent to a sequence of unconditional moment restrictions. Consequently, an efficient estimator under the conditional moment restrictions can be constructed from the sequence of unconditional restrictions. By letting K grow with the sample size all of the information in the conditional moment restrictions is eventually accounted for. Also, a consistent test of the conditional moment restrictions can be constructed from a sequence of tests of unconditional ones. If the conditional moment restriction is not satisfied then neither are the unconditional ones for K large enough, so the test detects all violations of the conditional moment restrictions as K grows. This result is an extension of Chamberlain (1987) and de Jong and Bierens (1994) to the case where $q_{kK}(x)$ can depend on K.

The specific role of Assumption 1 in the efficiency of the estimator is to ensure that linear combinations of $q^K(x)$ can approximate certain functions of x. From Chamberlain (1987) we know that an estimator obtained as the solution to

$$\sum_{i=1}^{n} B(x_i)\rho(z_i,\beta)=0, \qquad B(x)=\mathrm{E}[\partial\rho(z,\beta_0)/\partial\beta|x]'\{\mathrm{E}[\rho(z,\beta_0)\rho(z,\beta_0)'|x]\}^{-1},$$

achieves the semiparametric efficiency bound. Newey (1993) showed that the asymptotic variance of the optimal GMM estimator based on the moment function $\rho(z,\beta)\otimes q^K(x)$ corresponds to a minimum mean-square error approximation of $B(x)\rho(z,\beta_0)$ by linear combinations of $\rho(z,\beta_0)\otimes q^K(x)$. Thus, if Assumption 1 is satisfied linear combinations of $q^K(x)$ can approximate each component of $B(x)$, so that as K grows the asymptotic variance of the GMM estimator approaches the semiparametric efficiency bound.

If the spanning condition of Assumption 1 is not satisfied, the estimators we consider will still be asymptotically normal, but they will generally not be asymptotically efficient. Instead their asymptotic variance will be the same as a GMM estimator where $B(x)\rho(z,\beta_0)$ is replaced by its best mean-square error approximation by linear combinations of $\rho(z,\beta_0)\otimes q^K(x)$. This situation could arise, for example, if only certain components (or linear combinations) of x were used in forming the instruments. To avoid further complications we will not explicitly allow for this possibility in the results, though it would be straightforward to do so. We would find asymptotic normality of this estimator under the same conditions given here and could characterize the efficiency of the estimator.

There are many possible choices of $q_{kK}(x)$, including splines, power series, and Fourier series. Although each will be allowed for in at least some of our conditions, in several ways splines are the most attractive of these. Unlike power and Fourier series, spline approximations are not severely affected by singularities (e.g. discontinuities) in the function being approximated. Like power series, splines have faster approximation rates for smoother functions (up to the order of the spline), i.e. for functions where more derivatives exist. In addition, in our results they allow for restrictions on the growth of K with the sample size that are as weak as Fourier series, without the periodicity of Fourier series. As discussed in Gallant and Souza (1991) and Hong and White (1995), avoiding periodicity by using additional approximating functions or transforming x to be strictly inside their domain leads to very strong restrictions on the rate of growth of K.

For splines $q_{kK}(x)$ will consist of functions such that linear combinations are piecewise polynomials with join points referred to as knots. To describe splines consider first the scalar x case. Let s be a positive scalar giving the order of the spline. The most common specification is $s=3$. Let $t_1,\ldots,t_{K-s-1}$ denote knots and let $\xi(x)=1(x>0)x$, where $1(A)$ denotes the indicator function for the event A. Then a vector of spline approximating functions is given by

$$q^K(x)=(1,x,\ldots,x^s,\xi(x-t_1)^s,\ldots,\xi(x-t_{K-s-1})^s)'. \tag{2.3}$$

Some of our regularity conditions will require that the knots t_j be placed in the support of x. In practice this is done by placing them within the range of the observed data. Although the theory does not allow explicitly for knots chosen in this data-based way, it can be shown that the results are unaffected by such a use of the data. The growth rate conditions will also require that the knots be evenly spaced, although other kinds of knot spacings would also give the same results, as long as the ratio of the smallest distance between knots to the largest distance is bounded away from zero. Further, possible multicollinearity can be mitigated by using nonsingular linear transformations of the spline functions known as B-splines, e.g. see Schumaker (1981).

When x is multivariate, spline approximating functions can be formed from products of univariate splines for individual components of x. These functions can either be taken from Eq. (2.3) or can be B-splines.

Except for IV, all the results will depend on a normalization for the second moment matrix of the approximating functions, as specified in the following condition. Let X denote the support of x_i.

Assumption 2. For each K there is a constant scalar $\zeta(K)$ and matrix B such that $\tilde{q}^K(x)=Bp^K(x)$ for all $x\in X$, $\sup_{x\in X}\|\tilde{q}^K(x)\|\leqslant\zeta(K)$, $\mathrm{E}[\tilde{q}^K(x)\tilde{q}^K(x)']$ has smallest eigenvalue bounded away from zero uniformly in x, and $\sqrt{K}\leqslant\zeta(K)$.

This assumption is a normalization like that adopted by Newey (1997). The bound $\zeta(K)$ is meant to be explicit, and plays a crucial role in the theory for EL, GMM, and the overidentification test statistics. For example, asymptotic efficiency of EL is obtained under the condition that $\zeta(K)^2K^2/n \to 0$. Thus, explicit formulae for $\zeta(K)$ will be required to obtain explicit limits on the growth rate of K that are sufficient for efficiency. The need for these conditions comes from the need for explicit rates of convergence in probability for estimates of the second moment matrix of $g(z, \beta_0)$ combined with a need to bound it away from singularity. Only for IV will Assumption 2 not be needed.

Explicit formula for $\zeta(K)$ are available in a number of different cases. For splines, when x is continuously distributed with rectangular support and density bounded away from zero on its support, $\tilde{q}^K(x)$ can be a vector of products of B-splines multiplied by $\sqrt{K}$. As shown by Stone (1985), Burman and Chen (1989), and Newey (1997) Assumption 2 is then satisfied for $\zeta(K) = C\sqrt{K}$ and a constant C. Under the same conditions on x Newey (1988a) and Andrews (1991) showed that for power series this condition is satisfied for $\tilde{q}^K(x)$ equal to products of polynomials that are orthonormal with respect the uniform distribution, with $\zeta(K) = CK$ for another constant C. Andrews (1991) also showed that $\zeta(K) = C\sqrt{K}$ for Fourier series.

It is possible to weaken the conditions on the distribution of x at the expense of much larger bounds on $\zeta(K)$. As shown in Newey (1988b), if the density of x is bounded away from zero over some open ball, and not necessarily the whole support, then for power series the smallest eigenvalue of $\mathrm{E}[q^K(x)q^K(x)']$ is bounded below by K^{-CK}/C for some constant C. Then $\tilde{q}^K(K) = K^{CK}q^K(x)$ will satisfy Assumption 2 with $\zeta(K) = K^{CK}$ for some constant C. Also, when Fourier series are combined with power series, or the Fourier series have restricted domain, the smallest eigenvalue of $\mathrm{E}[q^K(x)q^K(x)']$ is bounded below by C^K as shown by Gallant and Souza (1991), so that Assumption 2 will be satisfied with $\zeta(K) = C^K$. Implied restrictions on the rate of growth of K with n implied by each of these conditions are outlined in the conclusion.

For the density bounded away from zero, among the approximations we consider, the smallest bound $\zeta(K)$ is obtained for splines. Consequently, splines will require the weakest restrictions on the growth rate of the number of terms. Furthermore, splines are well known to have good approximation properties. These features of splines mean they have the nicest theoretical properties.

3. Empirical likelihood and generalizations

One approach to empirical likelihood with conditional moment restrictions is to use approximating functions to make the conditional moment restrictions approximately be satisfied in the sample. This approach is complementary to smoothing the empirical likelihood to obtain a conditional distribution estimator, as in Kitamura et al. (2001). Only one maximization is required for our approach rather than the n maximizations when smoothing the empirical likelihood. On the other hand, we only estimate the marginal distribution of a single observation z, rather than the conditional distribution given x, so we lose some of the richness of the smoothing approach. Our approach does suffice for the purposes of efficiently estimating β and testing the conditional moment restrictions consistently. Also, the conditional moment restrictions will be approximately satisfied in the sample, in a sense discussed below.

The basic idea is to use empirical likelihood as in Qin and Lawless (1994) and Imbens (1997) with unconditional moment restrictions of the form $\mathrm{E}[\rho(z,\beta_0)\otimes q^K(x)]=0$, where $q^K(x)$ is a vector of approximating functions as discussed in the last section. The EL estimator of β and of the distribution of a single observations z is obtained as the solution to

$$\max_{\pi_i>0,\beta\in B}\sum_{i=1}^{n}\ln\pi_i \text{ s.t. } \sum_{i=1}^{n}\pi_i\rho(z_i,\beta)\otimes q^K(x_i)=0, \qquad \sum_{i=1}^{n}\pi_i=1. \tag{3.1}$$

The distribution of z_i is estimated by $\Pr(z=z_i)=\hat{\pi}_i$. The positivity constraints will be satisfied whenever the unit simplex has a nonempty intersection with the null space of the matrix $[g_1(\beta),\ldots,g_n(\beta)]$, for $g_i(\beta)=\rho(z_i,\beta)\otimes q^K(x_i)$. The theory will guarantee that such a nonempty intersection exists with probability approaching one when the conditional moment restrictions are satisfied. In practice, positivity may be a problem when the moment restrictions are severely violated in the data.

For this estimator the conditional moment restrictions will be approximately satisfied, in the sense that functions of x_i are approximately orthogonal to the residuals in the sample. To describe this property, let $a(x)$ denote some function of x_i. Consider an approximation of $a(x)$ by a linear combination $q^K(x)'\gamma$ of the approximating functions. Then by the constraints in Eq. (3.1) and the Cauchy–Schwartz inequality, it follows that for $\hat{\rho}_i=\rho(z_i,\hat{\beta})$.

$$\begin{aligned}\left\|\sum_{i=1}^{n}\hat{\pi}_i a(x_i)\hat{\rho}_i\right\| &= \left\|\sum_{i=1}^{n}\hat{\pi}_i[a(x_i)-q^K(x_i)'\gamma]\hat{\rho}_i\right\| \\ &\leqslant \left\{\sum_{i=1}^{n}\hat{\pi}_i[a(x_i)-q^K(x_i)'\gamma]^2\right\}^{1/2}\left\{\sum_{i=1}^{n}\hat{\pi}_i\|\hat{\rho}_i\|^2\right\}^{1/2}.\end{aligned}$$

The approximation error $\sum_{i=1}^{n}\hat{\pi}_i[a(x_i)-q^K(x_i)'\gamma]^2$ will be small, and hence the sample expectation $\sum_{i=1}^{n}\hat{\pi}_i a(x_i)\hat{\rho}_i$ close to zero, by virtue of the approximation properties of $q^K(x)'\gamma$. The sample expectation $\sum_{i=1}^{n}\hat{\pi}_i a(x_i)\hat{\rho}_i$ will even be uniformly small over classes of functions that can be uniformly approximated by a linear combination of $q^K(x)$. For example, for univariate x and splines it is known that such a uniform approximation holds over the class of all functions with uniformly bounded derivatives. Thus, the conditional moment condition is approximately satisfied for empirical likelihood, in the sense that functions of x are approximately uncorrelated with the residual.

The maximization problem in Eq. (3.1) is high dimensional. This computational burden can be avoided by solving a corresponding dual saddle point problem given by

$$\hat{\beta}=\underset{\beta\in\mathscr{B}}{\operatorname{argmin}}\max_{\lambda\in\hat{\Lambda}(\beta)}\sum_{i=1}^{n}\ln[1-\lambda' g_i(\beta)], \qquad \hat{\pi}_i=1/\{n[1+\hat{\lambda}' g_i(\hat{\beta})]\},$$

where $\mathscr{B}$ is a set of possible values for β and $\hat{\Lambda}(\beta) = \{\lambda\colon\ \lambda' g_i(\beta) < 1,\ i = 1,\dots,n\}$. As discussed in Qin and Lawless (1994), the estimators obtained from this problem are the same as those obtained from the original one.

The constraint $\lambda \in \hat{\Lambda}_n(\beta)$ bounds the argument of the log function to be within its domain. This constraint has a linear inequality form and requires that λ be an element of the intersection of n open half spaces for each β. In practice, although it is not too difficult to impose this constraint, it is often possible to just ignore it. Ignoring it will only lead to problems when the moment conditions are far from being satisfied in the data. Indeed, in the theory we show that at $\hat{\lambda}$ it will not be binding with probability approaching one, when the conditional moment restrictions hold. By continuity, we would then expect that the constraint is also not binding for small variation around the optimum with small misspecification.

For inference purposes, it is useful to have a consistent estimator of the asymptotic variance of $\sqrt{n}(\hat{\beta} - \beta_0)$. We consider here the estimator of Qin and Lawless (1994). Let $\hat{g}_i = \rho(z_i, \hat{\beta}) \otimes q^K(x_i)$ and $\hat{g}_{\beta i} = \partial\rho(z_i, \hat{\beta})/\partial\beta \otimes q^K(x_i)$. An estimator of the asymptotic variance like that of Qin and Lawless (1994) is

$$\hat{V} = (\hat{G}'\hat{\Omega}\hat{G})^{-1}, \quad \hat{G} = \sum_{i=1}^{n} \hat{\pi}_i \hat{g}_{\beta i}, \quad \hat{\Omega} = \sum_{i=1}^{n} \hat{\pi}_i \hat{g}_i \hat{g}_i'.$$

Of course, one could also use sample averages where $\hat{\pi}_i$ is replaced by $1/n$. The estimator with $\hat{\pi}_i$ will have the theoretical property that it is an efficient semiparametric estimator of the asymptotic variance of$\sqrt{n}(\hat{\beta} - \beta_0)$ under the conditional moment restrictions, as discussed in Brown and Newey (1998).

Another interesting inference issue is testing of the conditional moment restrictions of Eq. (2.1). Indeed, in some applications such a test may be the primary goal, as motivated by a test of some econometric model. We can form such a test from the empirical log-likelihood ratio of Owen (1988). Let

$$\hat{T} = 2\left[n\ln(1/n) - \sum_{i=1}^{n} \ln\hat{\pi}_i \right] = 2\sum_{i=1}^{n} \ln(1 - \hat{\lambda}' g_i(\hat{\beta})).$$

This statistic is the difference of the log-likelihood for the empirical distribution, which places probability $1/n$ on each observation, and the restricted distribution with $\hat{\pi}_i$. For fixed K, the statistic $\hat{T} \xrightarrow{d} \chi^2(JK - p)$ (Qin and Lawless, 1994). As a result of Theorem 2.1 and K growing with the sample size, it will provide a consistent test of the moment restrictions. Any case where $\mathrm{E}[\rho(z,\beta)|x] \neq 0$ for all β should be detected by finding $\mathrm{E}[\rho(z,\beta) \otimes q^K(x)] \neq 0$ for large enough K. In the theory we will show that when normalized this test statistic is asymptotically normal.

The empirical likelihood approach to conditional moment restrictions can be generalized. Let $s(v)$ be a concave function with domain $\mathscr{V}$ that is an open interval containing 0. We normalize this function so that $s_1 = s_2 = -1$, where $s_j(v) = \partial^j s(v)/\partial v^j$ and $s_j = s_j(0)$. The estimator is given by

$$\hat{\beta} = \underset{\beta\in\mathscr{B}}{\operatorname{argmin}} \sup_{\lambda\in\hat{\Lambda}(\beta)} \sum_{i=1}^{n} s(\lambda' g_i(\beta)), \quad \hat{\pi}_i = s_1\,(\hat{\lambda}' g_i(\hat{\beta})) \Big/ \sum_{j=1}^{n} s_1(\hat{\lambda}' g_j(\hat{\beta})),$$

where $\hat{\Lambda}(\beta)=\{\lambda:\lambda' g_i(\beta)\in\mathscr{V},\ i=1,\dots,n\}$. An estimator of the asymptotic variance can be formed as for EL, with the $\hat{\pi}_i$ given here replacing the EL $\hat{\pi}_i$ in the formula.

This generalized empirical likelihood (GEL) estimator includes EL as a special case, where $s(v)=\ln(1-v)$. It also includes the exponential tilting estimator of Imbens et al. (1998) and Kitamura and Stutzer (1997), where $s(v)=-\exp(v)$. As shown by Newey and Smith (2002), it includes the continuous updating estimator of Hansen et al. (1996), for $s(v)=-(1+v)^2/2$.

Each of these estimators will be asymptotically equivalent to the empirical likelihood estimator. Thus, each will be asymptotically efficient with conditional moment restrictions, and their asymptotic variance can be estimated as for empirical likelihood. As shown by Newey and Smith (2002), there is some theoretical preference for the empirical likelihood estimator based on its first-order bias that (unlike the others) does not grow with K, and based on its higher-order efficiency after bias correction. An exception is the case where $\rho(z_i,\beta_0)$ has zero third conditional moments, where all have the same first-order bias. For this symmetric case Donald et al. (2002) have derived the higher-order mean-square error of these estimators, and find that comparison between them depends on conditional kurtosis of $\rho(z_i,\beta_0)$ given x_i, and that the continuous updating estimator has smaller mean-square error than a bias corrected GMM estimator. Of course, the small sample behavior of these estimators may be different than their asymptotic approximations.

For each of these estimators there is a corresponding test of conditional moment restrictions based on their objective function. It takes the form

$$\hat{T}=2\left\{\max_{\lambda\in\hat{\Lambda}(\hat{\beta})}\sum_{i=1}^{n}s(\lambda' g_i(\hat{\beta}))-ns(0)\right\},$$

as in Smith (1997). For empirical likelihood, where $s(v)=\ln(1-v)$, this statistic is the empirical log-likelihood ratio. For $s(v)$ quadratic, it is equal to the GMM overidentification statistic where $\hat{g}(\beta)=\sum_{i=1}^{n} g_i(\beta)/n$, $\hat{\Omega}(\beta)=\sum_{i=1}^{n} g_i(\beta)g_i(\beta)'/n$,

$$\hat{T}=n\hat{g}(\hat{\beta})'\hat{\Omega}(\hat{\beta})^{-1}\hat{g}(\hat{\beta}).$$

This statistic is like that of Hansen (1982), where the efficient estimator $\hat{\beta}$ is used in the middle matrix in the test statistic. For the exponential tilting estimator where $s(v)=-\exp(v)$ the statistic is

$$\hat{T}=2\left\{n-\min_{\lambda\in\hat{\Lambda}(\hat{\beta})}\sum_{i=1}^{n}\exp(\lambda' g_i(\hat{\beta}))\right\}.$$

These statistics are mutually asymptotically equivalent. The distribution of each can be approximated by $\chi^2(JK-p)$, even as K grows with n, at the rates we specify.

4. GMM and IV

The GMM and IV estimators are alternatives to EL that are simpler to compute and are of long standing interest in econometrics. Also, the low bias of EL surely has some cost in terms of variance, so that GMM or IV may be preferred in some settings. For these reasons we briefly discuss these estimators here.

GMM and IV are obtained as the minimum to a quadratic form,

$$\hat{\beta} = \operatorname*{argmin}_{\beta \in B} \hat{g}(\beta)' \hat{W} \hat{g}(\beta),$$

where $\hat{W}$ is a positive semi-definite matrix. For GMM the matrix $\hat{W}$ is

$$\hat{W} = \hat{\Omega}(\tilde{\beta})^{-1},$$

where $\tilde{\beta}$ is some preliminary estimator of β_0 (such as IV). For IV the matrix $\hat{W}$ is

$$\hat{W} = \tilde{\Sigma}^{-1} \otimes \hat{A}^{-}, \quad \hat{A} = \sum_{i=1}^{n} q^K(x_i) q^K(x_i)'/n,$$

where $\tilde{\Sigma}$ is a positive definite matrix and $\hat{A}^{-}$ is any generalized inverse of $\hat{A}$ (i.e. satisfies $\hat{A}\hat{A}^{-}\hat{A} = \hat{A}$). The IV weighting matrix will not lead to an efficient estimator in general, but will in the homoskedastic case where $\Sigma(x) = \mathrm{E}[\rho(z_i, \beta_0)\rho(z_i, \beta_0)'|x]$ is constant and $\tilde{\Sigma}$ is a consistent estimator of $\Sigma = \Sigma(x)$. In at least some settings the IV estimator has better finite sample properties than GMM, e.g. see Arellano and Bond (1991).

For IV the generalized inverse is important because it allows for perfect multicollinearity among the approximating functions, even asymptotically. This is one of the attractive features of the estimation theory for IV that differs from the theory for GMM and EL. For IV it is possible to avoid conditions for guaranteeing positive definiteness of the second moment matrix of the approximating functions in large samples. Consequently, no restrictions need be put on the marginal distribution of x.

For GMM and IV the asymptotic variance of $\sqrt{n}(\hat{\beta} - \beta_0)$ can be estimated by

$$\hat{V} = (\hat{G}' \hat{W} \hat{G})^{-1}, \quad \hat{G} = \partial \hat{g}(\hat{\beta})/\partial \beta.$$

Also, the matrix $\hat{W}$ can be updated and the corresponding asymptotic variance estimator be constructed. For GMM this variance estimator would replace $\hat{W}$ with $\hat{\Omega}(\hat{\beta})^{-1}$. For IV $\hat{W}$ is replaced by $\hat{\Sigma}^{-1} \otimes \hat{A}^{-}$ where $\hat{\Sigma} = \sum_{i=1}^{n} \rho(z_i, \hat{\beta})\rho(z_i, \hat{\beta})'/n$. For IV these variance estimators will only be consistent under homoskedasticity.

For GMM and IV the test of conditional moment restrictions is the overidentifying test statistic. This test statistic is

$$\hat{T} = n\hat{g}(\hat{\beta})' \hat{W} \hat{g}(\hat{\beta}).$$

As for the asymptotic variance estimator, $\hat{W}$ can be replaced by an update. For GMM, this test is similar to the GEL test in the quadratic $s(v)$ case. For IV the matrix $\hat{W}$ is only appropriate under homoskedasticity, so that the asymptotic distribution approximation will only be correct under this assumption. These test statistics will be discussed in more detail in Section 6.

5. Large sample properties of the estimators

We proceed by giving conditions for consistency, asymptotic normality, and asymptotic efficiency of IV, GMM, and generalized EL. We follow this outline because the conditions tend to increase in strength as we consider each of these estimators in turn. We begin with the following condition for consistency.

Assumption 3. The data are i.i.d. and (a) β_0 is unique value of β in $\mathscr{B}$ satisfying $\mathrm{E}[\rho(z,\beta)|x]=0$; (b) $\mathscr{B}$ is compact; (c) $\mathrm{E}[\sup_{\beta\in\mathscr{B}}\|\rho(z,\beta)\|^2|x]$ is bounded; (d) there is $\delta(z)$ and $\alpha>0$ such that for all $\tilde{\beta},\beta\in\mathscr{B}$, $\|\rho(z,\tilde{\beta})-\rho(z,\beta)\|\leqslant\delta(z)\|\tilde{\beta}-\beta\|^{\alpha}$ and $\mathrm{E}[\delta(z)^2]<\infty$.

This condition imposes the minimal identification condition, that β_0 is the unique value where the conditional moment restrictions are satisfied. In particular, the existence of a known K such that the unconditional moment restrictions $\mathrm{E}[\rho(z,\beta)\otimes q^K(x)]=0$ serve to identify β_0 is not required. We are able to use this weak assumption because K is growing with n, and so β_0 will be identified by Lemma 2.1. This hypothesis also imposes a bounded second conditional moment and Lipschitz condition, that is used to apply the uniform convergence result of Newey (1991).

With these conditions in place we obtain the following consistency result:

Theorem 5.1. *If Assumptions* 1 *and* 3 *are satisfied*, $\tilde{\Sigma}\xrightarrow{p}\Sigma$, Σ *is positive definite*, $K\to\infty$, *and* $K/n\to 0$ *then the IV estimator satisfies* $\hat{\beta}\xrightarrow{p}\beta_0$.

The only rate condition imposed here is the minimal one $K/n\to 0$, that is needed for the variance of series estimators to vanish in large samples. This result is a parametric version of Theorem 5.1 of Newey and Powell (1989) for the vector $\rho(z,\beta)$ case.

An additional condition is needed for asymptotic normality. Let $\rho_\beta(z,\beta)=\partial\rho(z,\beta)/\partial\beta$, $D(x)=\mathrm{E}[\rho_\beta(z,\beta_0)|x]$, and $\rho_{j\beta\beta}(z,\beta)=\partial^2\rho_j(z,\beta)/\partial\beta\partial\beta'$, $(j=1,\ldots,J)$.

Assumption 4. (a) $\beta_0\in int(\mathscr{B})$; (b) $\rho(z,\beta)$ is twice continuously differentiable in a neighborhood $\mathscr{N}$ of β_0, $\mathrm{E}[\sup_{\beta\in\mathscr{N}}\|\rho_\beta(z,\beta)\|^2|x]$ and $\mathrm{E}[\|\rho_{j\beta\beta}(z,\beta_0)\|^2|x]$ are bounded, $(j=1,\ldots,J)$; (c) $\mathrm{E}[D(x)'D(x)]$ is nonsingular.

These are quite standard regularity conditions. Part (c) is a local identification condition that is essential for asymptotic normality. The other conditions are familiar smoothness conditions, although the assumption of twice differentiability of the residual vector is stronger than is usually assumed. This assumption is useful in showing $\sqrt{n}$-consistency when the number of moments is growing with the sample size.

Our IV asymptotic normality result is the following:

Theorem 5.2. *If Assumptions* 1, 3, *and* 4 *are satisfied*, $\tilde{\Sigma}\xrightarrow{p}\Sigma$, Σ *is positive definite*, $K\to\infty$, *and* $K^2/n\to 0$ *then the IV estimator satisfies* $\sqrt{n}(\hat{\beta}-\beta_0)\xrightarrow{d}\mathrm{N}(0,V)$ *for*

$$V=(\mathrm{E}[D(x)'\Sigma^{-1}D(x)])^{-1}\mathrm{E}[D(x)'\Sigma^{-1}\Sigma(x)\Sigma^{-1}D(x)](\mathrm{E}[D(x)'\Sigma^{-1}D(x)])^{-1}.$$

Also, *if* $\Sigma(x)=\Sigma$ *then* $\hat{V}\xrightarrow{p}V=(\mathrm{E}[D(x)'\Sigma^{-1}D(x)])^{-1}$.

The slower rate of growth for K here as compared with the consistency result is essential when there is endogeneity. As shown in Donald and Newey (2001), there is a bias term for $\sqrt{n}(\hat{\beta}-\beta_0)$ that is of order $K/\sqrt{n}$, that must converge to zero for the asymptotic distribution to be centered at zero. This condition for asymptotic normality of IV is the same as in Newey (1990) for a one-step estimator. This result adds to the previous literature by giving conditions for asymptotic normality of IV under the weakest possible identification condition, that the conditional moment restrictions are uniquely satisfied at the truth. Also, it shows efficiency of IV under homoskedasticity for the fully iterated estimator, while the results of Newey (1990) only apply to a one-step estimator.

Neither of these results requires Assumption 2. This is one of the virtues of the theory for IV, that no conditions need be imposed on the marginal distribution of x. For GMM, the theory here requires the stronger conditions of Assumption 2. Indeed, we can show consistency under conditions like those of IV, with the addition of Assumption 2 and the following condition:

Assumption 5. (a) $\Sigma(x)=\mathrm{E}[\rho(z,\beta_0)\rho(z,\beta_0)'|x]$ has smallest eigenvalue bounded away from zero; (b) for a neighborhood $\mathcal{N}$ of β_0, $\mathrm{E}[\sup_{\beta\in\mathcal{N}}\|\rho(z,\beta)\|^4|x]$ is bounded, and for all $\beta\in\mathcal{N}$, $\|\rho(z,\beta)-\rho(z,\beta_0)\|\leqslant\delta(z)\|\beta-\beta_0\|$ and $\mathrm{E}[\delta(z)^2|x]$ is bounded.

This condition is useful in obtaining a convergence rate for the sample second moment matrix $\hat{\Omega}(\tilde{\beta})$ and for guaranteeing that it is bounded away from singularity.

Theorem 5.3. *If Assumptions* 1–3 *and* 5 *are satisfied*, $\tilde{\beta}=\beta_0+\mathrm{O_p}(1/\sqrt{n})$, $K\to\infty$, *and* $\zeta(K)^2K/n\to 0$ *then the GMM estimator satisfies* $\hat{\beta}\xrightarrow{\mathrm{p}}\beta_0$.

As previously discussed, $\zeta(K)\leqslant C\sqrt{K}$ for splines and $\zeta(K)\leqslant CK$ for power series, so that $K^2/n\to 0$ suffices for splines and $K^3/n\to 0$ for power series. These conditions are stronger than for IV. It is difficult to weaken them because they are used to control the singularity of the second moment matrix $\hat{\Omega}(\tilde{\beta})$. This seems to be the first result showing consistency of the two-step GMM estimator when the number of moment restrictions grow with the sample size. Previous results, such as those of Newey (1993) and Hahn (1997) only apply to a one-step estimator.

Asymptotic normality and asymptotic efficiency of the GMM estimator holds under the additional condition of Assumption 4.

Theorem 5.4. *If Assumptions* 1–5 *are satisfied*, $K\to\infty$, $\tilde{\beta}=\beta_0+\mathrm{O_p}(1/\sqrt{n})$, *and* $\zeta(K)^2K/n\to 0$ *then the GMM estimator satisfies*

$$\sqrt{n}(\hat{\beta}-\beta_0)\xrightarrow{\mathrm{d}}\mathrm{N}(0,V),\qquad \hat{V}\xrightarrow{\mathrm{p}}V=(\mathrm{E}[D(x)'\Sigma(x)^{-1}D(x)])^{-1}.$$

This result gives conditions for the GMM estimator to attain the semiparametric asymptotic variance bound V. This seems to be the first asymptotic normality result for the fully iterated two-step nonlinear GMM estimator with growing numbers of moments. Newey (1993) and Hahn (1997) only consider linearization around an initial asymptotically normal estimator and Koenker and Machado (1999) linear models with stronger restrictions on the growth rate of K.

The following condition is useful for generalized empirical likelihood:

Assumption 6. (a) $s(v)$ is twice continuously differentiable with Lipschitz second derivative in a neighborhood of 0; (b) there is $\gamma > 2$ with $\mathrm{E}[\sup_{\beta\in\mathscr{B}} \|\rho(z_i,\beta)\|^{\gamma}] < \infty$ and $\zeta(K)^2K/n^{1-2/\gamma} \to 0$.

Part (b) of this assumption imposes a slightly stronger restriction on the growth rate of K than the condition $\zeta(K)^2K/n \to 0$ used in GMM estimation, that is less strong the more moments of $\rho(z_i,\beta)$ there are.

With these conditions we obtain the following consistency result:

Theorem 5.5. *If Assumptions* 1–3, 5, *and* 6 *are satisfied and* $K \to \infty$ *then the GEL estimator satisfies* $\hat{\beta}\xrightarrow{\mathrm{p}}\beta_0$.

The restrictions on the growth rate of K may be stronger than are needed. Indeed, as shown in Newey and Smith (2002), the bias of the empirical likelihood estimator should be of smaller order for large K than that of the two-step GMM estimator. Consequently even $K^2/n \to 0$ may not be needed for $\sqrt{n}$-consistency. We leave the pursuit of a weaker bound on the growth rate of K to future research.

We also have the following asymptotic normality result:

Theorem 5.6. *If Assumptions* 1–6 *are satisfied*, $K \to \infty$, *and* $\zeta(K)^2K^2/n \to 0$, *then the GEL estimator satisfies*

$$\sqrt{n}(\hat{\beta}-\beta_0)\xrightarrow{\mathrm{d}}\mathrm{N}(0,V), \qquad \hat{V}\xrightarrow{\mathrm{p}}V, \qquad V = (\mathrm{E}[D(x)'\Sigma(x)^{-1}D(x)])^{-1}.$$

This result gives conditions for generalized EL estimators to attain Chamberlain's (1987) semiparametric efficiency bound. As previously discussed, $\zeta(K) \leqslant C\sqrt{K}$ for splines and $\zeta(K) \leqslant CK$ for power series, so that the rate conditions correspond to $K^3/n \to 0$ for splines and $K^4/n \to 0$ for power series. These seem to be the first results of any kind on the asymptotic properties of GEL estimators when the number of moment restrictions can grow with the sample size.

6. Consistent conditional moment tests

We consider tests based on GEL, GMM, and IV. The GEL test has the form given above,

$$\hat{T}_{\mathrm{GEL}} = 2\left\{\max_{\lambda\in\hat{\Lambda}(\hat{\beta})}\sum_{i=1}^{n} s(\lambda' g_i(\hat{\beta})) - ns(0)\right\}.$$

The GMM test statistic is

$$\hat{T}_{\mathrm{GMM}} = n\hat{g}(\hat{\beta})'\hat{\Omega}(\tilde{\beta})^{-1}\hat{g}(\hat{\beta}),$$

where $\tilde{\beta}$ is some $\sqrt{n}$-consistent estimator. The IV test has the form

$$\hat{T}_{\mathrm{IV}} = n\hat{g}(\hat{\beta})'(\tilde{\Sigma}^{-1} \otimes \hat{A}^{-})\hat{g}(\hat{\beta}), \quad \hat{A} = \sum_{i=1}^{n} q^K(x_i)q^K(x_i)'/n,$$

where $\tilde{\Sigma}$ is a consistent estimator of $Var(\rho(z_i, \beta_0))$. Unlike GEL and GMM, the distribution results we give for this test statistic will require homoskedasticity.

The relationship of these tests with some in the literature is clarified by some special cases of the IV test. Consider the case where $J = 1$, so

$$\hat{T}_{\mathrm{IV}} = \hat{g}(\hat{\beta})'\hat{A}^{-1}\hat{g}(\hat{\beta})/\tilde{\Sigma}.$$

This test statistic can be interpreted as $(\hat{\sigma}^2/\tilde{\Sigma})nR^2$, where R^2 is the constant unadjusted r-squared from a regression of $\rho(z_i, \hat{\beta})$ on $q^K(x_i)$ and $\hat{\sigma}^2 = \sum_{i=1}^{n} \rho(z_i, \hat{\beta})^2/n$. For instance, consider the case where for large enough K

$$\rho(z_i, \beta) = y_i - [Sq^K(x_i)]'\beta,$$

and S is a selection matrix that picks out the same variables for each K. In other words, $\rho(z, \beta)$ is a residual for a linear regression where the right-hand side variables are included in $q^K(x)$. Also, suppose that $\tilde{\Sigma}$ is the sum of squared residuals from a regression of y_i on $q^K(x_i)$ divided by $n - K$. Then $\hat{T}$ is $K - p$ times the F-statistic for the null hypothesis that all but the regressors $Sq^K(x_i)$ have zero coefficients. This statistic was previously considered in Eubank and Spiegelman (1990) and Hong and White (1995). We obtain the asymptotic distribution of this statistic under weaker conditions on the growth rate of K than previously given in some cases. Our results also generalize previous results by allowing for endogeneity in right-hand side variables in testing conditional moment restrictions.

For fixed K the asymptotic distribution of all of the statistics is known to be $\chi^2(JK - p)$, under the null hypothesis that the conditional moment restrictions are satisfied. Here we show that this approximation will continue to hold as K grows with the sample size. For this purpose we use the asymptotic normal approximation to the chi-square for large degrees of freedom. From the fact that a random variable $Y_m \sim \chi^2(m)$ has the same distribution as the sum of m i.i.d. random variables with mean 1 and variance 2, the central limit theorem gives

$$\frac{\chi^2(JK - p) - (JK - p)}{\sqrt{2(JK - p)}} \xrightarrow{d} \mathrm{N}(0, 1),$$

as $K \to \infty$. We will give conditions for

$$\frac{\hat{T} - (JK - p)}{\sqrt{2(JK - p)}} \xrightarrow{d} \mathrm{N}(0, 1). \tag{6.1}$$

It follows from these two results that for $q_{\alpha,m}$ the $1-\alpha$ quantile of the $\chi^2(m)$ distribution,

$$\Pr(\hat{T} \geqslant q_{\alpha,JK-p}) = \Pr\left(\frac{\hat{T} - (JK - p)}{\sqrt{2(JK - p)}} \geqslant \frac{q_{\alpha,JK-p} - (JK - p)}{\sqrt{2(JK - p)}}\right) \to \alpha.$$

Thus, Eq. (6.1) will imply that the $\chi^2(JK-p)$ approximation to the distribution of $\hat{T}$ will be asymptotically correct even with K growing.

One could also use directly the asymptotic normality result of Eq. (6.1), choosing a rejection region of the form

$$\frac{\hat{T}-(JK-p)}{\sqrt{2(JK-p)}} \geqslant z_{1-\alpha},$$

where $z_{1-\alpha}$ is the $1-\alpha$ quantile of the standard normal distribution. Either this rejection region or the one $\hat{T} \geqslant q_{\alpha,JK-p}$ have asymptotic level α. We have a preference for the chi-squared approximation because it is correct for fixed K, and because the asymptotic normality approximation does not depend on the value of p used in Eq. (6.1) or on the estimator used in forming the test, as we show below. A formal justification of this preference would depend on the use of higher order asymptotic approximations, which we reserve to future work.

Before showing asymptotic normality of the statistics we first present two preliminary results. These results are of interest because they apply more generally, to other moment restriction settings as well as those considered here. Also, they help to highlight the differences between the testing theory given here and that already in the literature. For each of them, let $g(z_i,\beta)$ be some $m\times 1$ vector of functions, that is not necessarily of the Kronecker product form we have considered here. Let

$$\hat{g}(\beta)=\sum_{i=1}^{n} g(z_i,\beta)/n,$$

$$\Omega = \mathrm{E}[g(z_i,\beta_0)g(z_i,\beta_0)'], \qquad G=\mathrm{E}[\partial g(z,\beta_0)/\partial\beta].$$

Also, let $\hat{\beta}$ and $\hat{\Omega}$ denote estimators of β_0 and Ω respectively (that need not be GEL, GMM, or IV). The first result is:

Lemma 6.1. *If* $\mathrm{E}[g(z_i,\beta_0)]=0$, $\sqrt{n}(\hat{\beta}-\beta_0)=\mathrm{O_p}(1)$, $\|\hat{\Omega}-\Omega\|=\mathrm{o_p}(1/\sqrt{m})$, *the smallest eigenvalue of* Ω *is bounded away from zero*, $g(z,\beta)$ *is differentiable in a neighborhood of* β_0, $\|\partial\hat{g}(\bar{\beta})/\partial\beta - G\|\xrightarrow{p}0$ *for any* $\bar{\beta}$ *with* $\|\bar{\beta}-\beta_0\|\leqslant\|\hat{\beta}-\beta_0\|$, $G'\Omega^{-1}G$ *is bounded, and* $m\to\infty$ *then for any constant a*

$$\frac{n\hat{g}(\hat{\beta})'\hat{\Omega}^{-1}\hat{g}(\hat{\beta}) - n\hat{g}(\beta_0)'\Omega^{-1}\hat{g}(\beta_0)}{\sqrt{2(m-a)}} \xrightarrow{p} 0. \tag{6.2}$$

Although this result is similar to a combination of Lemmas 4 and 5 of de Jong and Bierens (1994), it is different in some useful ways. Since $\hat{g}(\hat{\beta})'\hat{\Omega}^{-1}\hat{g}(\hat{\beta})$ is invariant to nonsingular linear transformations of $g(z_i,\beta_0)$, the restriction that the smallest eigenvalue of Ω is bounded away from zero is a normalization, e.g. that can be obtained by replacing $g(z,\beta)$ by $\Omega^{-1/2}g(z,\beta)$. This normalization loads all of the rate and size restrictions onto the condition $\|\hat{\Omega}-\Omega\|=\mathrm{o_p}(1/\sqrt{m})$. This normalization actually leads to weaker rate conditions than can be obtained in other ways, in some cases. In addition, this result is stated with general conditions, allowing it to be applied to any GMM overidentifying tests where the number of moment restrictions is growing with the sample size.

The only condition that this result imposes on the estimated parameters is that they are $\sqrt{n}$-consistent. Thus, the limiting distribution of the quadratic form $n\hat{g}(\hat{\beta})'\hat{\Omega}^{-1}\hat{g}(\hat{\beta})$ will be invariant to the choice of estimator. This occurs because the growth of the number of moment restrictions m overwhelms the effect of the estimated parameters. In addition the condition on convergence of the moment Jacobian is generally weaker than the one on convergence of $\hat{\Omega}$, while boundedness of $G'\Omega^{-1}G$ is easily seen to hold quite generally. Note that $(G'\Omega^{-1}G)^{-1}$ is the asymptotic variance matrix of the optimal GMM estimator, which will generally be bounded below by the semiparametric efficiency bound, for whatever model the moments come from, so that $G'\Omega^{-1}G$ is bounded above. Thus, boundedness of $G'\Omega^{-1}G$ above is satisfied in great generality.

The next result is essentially Theorem 1 of de Jong and Bierens (1994), for a general GMM setting:

Lemma 6.2. *If* $\mathrm{E}[g(z_i,\beta_0)]=0$, $m\to\infty$, *and* $\mathrm{E}[\{g(z_i,\beta_0)'\Omega^{-1}g(z_i,\beta_0)\}^2]/(m\sqrt{n})\to 0$, *then for any constant* a,

$$\frac{n\hat{g}(\beta_0)'\Omega^{-1}\hat{g}(\beta_0)-(m-a)}{\sqrt{2(m-a)}}\xrightarrow{\mathrm{d}}\mathrm{N}(0,1). \tag{6.3}$$

These basic results lead directly to the asymptotic distribution results for the GEL, GMM, and IV test statistics. Indeed, most of the conditions of Lemmas 6.1 and 6.2 were shown in the course of deriving the asymptotic normality results of Section 5. We give the limiting distribution result for the GMM and IV statistics in the next theorem.

Theorem 6.3. *If Assumptions* 1–4 *are satisfied* $K\to\infty$, $\tilde{\beta}=\beta_0+\mathrm{O_p}(1/\sqrt{n})$, *and* $\zeta(K)^2K^2/n\to 0$ *then*

$$\frac{\hat{T}_{\mathrm{GMM}}-(JK-p)}{\sqrt{2(JK-p)}}\xrightarrow{\mathrm{d}}\mathrm{N}(0,1).$$

If in addition $\tilde{\Sigma}=\Sigma+\mathrm{O_p}(1/\sqrt{n})$ *and* $Var(\rho(z,\beta_0)|x)$ *is constant then*

$$\frac{(\hat{T}_{\mathrm{IV}}-\hat{T}_{\mathrm{GMM}})}{\sqrt{2(JK-p)}}\xrightarrow{\mathrm{p}}0.$$

This result is new in providing an asymptotic distribution result for the GMM (and IV) overidentification test statistics of conditional moment restrictions. These results allow for endogeneity of the variables, which is important in many settings. Also, the rate conditions are weaker than several in the literature. For Fourier series and splines $\zeta(K)^2K^2/n\to 0$ is equivalent to $K^3/n\to 0$. This is the same rate given for Fourier series in de Jong and Bierens (1994) and Hong and White (1995), but is better than the $K^5/n\to 0$ for splines given in Hong and White (1995). Also, for multivariate power series it suffices that $K^4/n\to 0$. This condition is new, power series having not been considered before, except with a Gaussian disturbance in Eubank and Spiegelman (1990).

Table 1
Rate restrictions; density of x bounded positive

$q_{kK}(x)$	EL	GMM	IV[a]	EL Test	GMM Test
Splines	$K^3/n \to 0$	$K^2/n \to 0$	$K^2/n \to 0$	$K^4/n \to 0$	$K^3/n \to 0$
Power series	$K^4/n \to 0$	$K^3/n \to 0$	$K^2/n \to 0$	$K^5/n \to 0$	$K^4/n \to 0$
Fourier series	$K^3/n \to 0$	$K^2/n \to 0$	$K^2/n \to 0.$	$K^4/n \to 0$	$K^3/n \to 0$

[a]IV does not require density of x bounded positive.

We can state a corresponding result for the generalized empirical likelihood statistic $\hat{T}_{\mathrm{GEL}}$. We will allow any $\sqrt{n}$-consistent estimator $\hat{\beta}$ to be used in forming the test.

Theorem 6.4. *If Assumptions* 1–6 *are satisfied,* $K \to \infty$, $\hat{\beta} = \beta_0 + \mathrm{O_p}(1/\sqrt{n})$, *and* $\zeta(K)^2K^3/n \to 0$ *then* $(\hat{T}_{\mathrm{GEL}} - \hat{T}_{\mathrm{GMM}})/\sqrt{2(JK-p)} \xrightarrow{\mathrm{p}} 0$.

This gives a new asymptotic distribution result for empirical likelihood and its generalizations. The rate conditions here are stronger than for the GMM and IV tests, but still weaker than the spline results of Hong and White (1995).

A full analysis of the power properties of these tests is beyond the scope of this paper. Here we just note that GMM tests of overidentifying restrictions should perform similarly to other tests. The following general result provides an illustration. Here let β_a denote the limit of $\hat{\beta}$ under misspecification, $\Omega_a = \mathrm{E}[g(z_i,\beta_a)g(z_i,\beta_a)']$, and $G_a = \mathrm{E}[\partial g(z_i,\beta_a)/\partial\beta]$.

Lemma 6.5. *If* $\hat{\beta} \xrightarrow{\mathrm{p}} \beta_a$, $\|\hat{\Omega} - \Omega_a\| \xrightarrow{\mathrm{p}} 0$, *the smallest eigenvalue of* Ω_a *is bounded away from zero,* $g(z,\beta)$ *is differentiable in a neighborhood of* β_a, $\|\partial\hat{g}(\bar{\beta})/\partial\beta - G_a\| \xrightarrow{\mathrm{p}} 0$ *for any* $\bar{\beta}$ *with* $\|\bar{\beta} - \beta_a\| \leqslant \|\hat{\beta} - \beta_a\|$, $G_a'\Omega_a^{-1}G_a$ *is bounded,* $m \to \infty$, $m/n \to 0$, $\mathrm{E}[g(z_i,\beta_a)]'\Omega_a^{-1}\mathrm{E}[g(z_i,\beta_a)] \to \Delta$ *then for any constant* a

$$\left(\frac{\sqrt{m}}{n}\right)\frac{n\hat{g}(\hat{\beta})'\hat{\Omega}^{-1}\hat{g}(\hat{\beta}) - (m-a)}{\sqrt{2(m-a)}} \xrightarrow{\mathrm{p}} \Delta/\sqrt{2}.$$

This result implies that the statistic that is asymptotically normal under the null hypothesis grows at rate $n/\sqrt{m}$ under the alternative. This is the same rate of growth obtained by de Jong and Bierens (1994) and Hong and White (1995).

7. Conclusion

We have derived limits on the rate of growth for the number of moment restrictions that lead to asymptotic efficiency of GMM and GEL. We summarize primitive conditions for these results in Table 1, under the condition that the density of x is bounded away from zero (except for IV).

It is possible to weaken the condition that the density of x is bounded away from zero, at the expense of stronger restrictions on the growth of K. As shown in Newey (1988b), if the density is bounded away from zero over some interval, not the whole support, then for power series the smallest eigenvalue of $E[q^K(x)q^K(x)']$ is bounded below by K^{-CK}/C for some constant C. Then $\tilde{q}^K(K)=K^{CK}q^K(x)$ will satisfy Assumption 2 with $\zeta(K)=K^{CK}$ for some constant C. The resulting rate restriction for power series is

$$\text{Power series; density bounded positive on } \textit{subset} \text{ of support}:$$

$$K\ln(K)/\ln(n) \to 0.$$

This condition shows a trade-off between the strength of the assumptions on the distribution of x and the restrictions on the growth rate for K. Here the density is only required to be bounded away from zero over an arbitrarily small interval, but a very strong restriction is imposed on the growth rate for number of terms.

Strong restrictions on the growth rate of K are also needed for Assumption 2 when Fourier series are modified by also including power series and/or restricting the domain of the power series. As discussed in Gallant and Souza (1991), the smallest eigenvalue goes to zero very quickly for such series, when the density of x is bounded away from zero on its support. For the case where the density is bounded away from zero on its support we have

$$\text{Fourier flexible form}: K/\ln(n) \to 0.$$

It would be useful to know how to choose K in practice. The rate results we have presented are useful preliminaries for the development of sample based methods for choosing K. Donald et al. (2002) give criteria that can be used for choosing the number of moment restrictions.

Appendix A. Proofs

Throughout appendix, C will denote a generic positive constant that may be different in different uses. Also, with probability approaching one will be abbreviated as w.p.a.1, positive semi-definite as p.s.d., positive definite as p.d., $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$, and $A^{1/2}$ will denote the minimum and maximum eigenvalues, and square root, respectively, of a symmetric matrix A. Let $\sum_i$ denote $\sum_{i=1}^n$. Also, let CS, M, and T refer to the Cauchy–Schwartz, Markov, and triangle inequalities, respectively.

Proof of Lemma 2.1. The first conclusion follows by iterated expectations. To show the second conclusion, $\rho=\rho(z,\beta_0)$ and Γ_K be such that $E[\|E[\rho|x]-\Gamma_K q^K(x)\|^2]\to 0$. It then follows that $E[q^K(x)'\Gamma_K'\rho]=E[q^K(x)'\Gamma_K' E[\rho|x]]\to E[\|E[\rho|x]\|^2]>0$, implying that $E[\rho\otimes q^K(x)]\neq 0$ for all K large enough. □

The following is a version of the standard consistency result that will be useful in the proofs:

Lemma A.1. *Suppose that* (i) $R(\beta)$ *has a unique minimum at* $\beta_0 \in B$; (ii) B *is compact*; (iii) $R(\beta)$ *is continuous*; *and* (iv) $\sup_{\beta \in B} |\hat{R}(\beta) - R(\beta)| \xrightarrow{p} 0$. *Then for any* $\tilde{\beta} \in B$, *if* $\hat{R}(\tilde{\beta}) \xrightarrow{p} R(\beta_0)$ *then* $\tilde{\beta} \xrightarrow{p} \beta_0$.

Proof. By (iv) and the triangle inequality, $R(\tilde{\beta}) \xrightarrow{p} R(\beta_0)$. Also, for any open set $\mathcal{N}$ containing β_0, by (i)–(iii) $\inf_{\beta \in B \setminus \mathcal{N}} R(\beta) > R(\beta_0)$, so that $\tilde{\beta} \in \mathcal{N}$ *w.p.a.1*. □

Lemma A.2. *If Assumption* 2 *is satisfied then it can be assumed without loss of generality that* $\tilde{q}^K(x_i) = q^K(x_i)$ *and that* $\mathrm{E}[q^K(x_i)q^K(x_i)'] = I_K$.

Proof. All of the estimators are invariant to a nonsingular linear transformation of $q^K(x)$. In particular, the estimator is the same with $\tilde{q}^K(x)$ as with $q^K(x)$, giving the first conclusion. Now, let $B = \mathrm{E}[q^K(x_i)q^K(x_i)']$ and $\bar{q}^K(x) = B^{-1/2}q^K(x)$. By the smallest eigenvalue of B bounded below, the largest eigenvalue of B^{-1} is bounded above. Then $\|\bar{q}^K(x)\| = \sqrt{q^K(x)'B^{-1}q^K(x)} \leqslant C\|q^K(x)\| \leqslant C\zeta(K)$. Furthermore, $\mathrm{E}[\bar{q}^K(x_i)\bar{q}^K(x_i)'] = I_K$ holds, giving the conclusion. □

The next result is helpful to introduce some notation. Let $q_i = q^K(x_i)$.

Lemma A.3. *If Assumption* 1 *is satisfied*, (i) $\hat{\beta} \xrightarrow{p} \tilde{\beta}$; (ii) $a_i(\beta) = a(z_i, \beta)$ *and* $b_i(\beta) = b(z_i, \beta)$ *are* $r \times 1$ *vectors of functions that are continuous at* $\tilde{\beta}$ *with probability one and there is a neighborhood* N *of* $\tilde{\beta}$ *such that* $\mathrm{E}[\sup_{\beta \in \mathcal{N}} \|a_i(\beta)\|^2] < \infty$ *and* $\mathrm{E}[\sup_{\beta \in \mathcal{N}} \|b_i(\beta)\|^2] < \infty$, $\mathrm{E}[\|a_i(\tilde{\beta})\|^2 | x]$ *and* $\mathrm{E}[\|b_i(\tilde{\beta})\|^2 | x]$ *are bounded*; (iii) $U_i = U(x_i)$ *is* $r \times r$ *p.d. matrix that is bounded and has smallest eigenvalue bounded away from zero*; (iv) $K \to \infty$, *and* $K/n \to 0$, *then*

$$\sum_i a_i(\hat{\beta})' \otimes q_i' \left(\sum_i U_i \otimes q_i q_i' \right)^- \sum_i b_i(\hat{\beta}) \otimes q_i / n \xrightarrow{p} \mathrm{E}[\mathrm{E}[a_i(\tilde{\beta})' | x_i] U_i^{-1} \mathrm{E}[b_i(\tilde{\beta}) | x_i]].$$

Proof. Let $F_i = U_i^{1/2}$ be a symmetric square root of U_i, $P_i = F_i \otimes q_i'$, $P = [P_1', \ldots, P_n']'$, $A_i(\beta) = F_i^{-1} a_i(\beta)$, $A(\beta) = [A_1(\beta)', \ldots, A_n(\beta)']'$, $\hat{A} = A(\hat{\beta})$, $A = A(\tilde{\beta})$, $B_i(\beta) = F_i^{-1} b_i(\beta)$, $B(\beta) = [B_1(\beta)', \ldots, B_n(\beta)']'$, $\hat{B} = B(\hat{\beta})$, and $B = B(\tilde{\beta})$. Note that $\sum_i U_i \otimes q_i q_i' = P'P$, and that

$$\sum_i a_i(\hat{\beta})' \otimes q_i' \left(\sum_i U_i \otimes q_i q_i' \right)^- \sum_i b_i(\hat{\beta}) \otimes q_i = \hat{A}' Q \hat{B},$$

$$Q = P(P'P)^- P'.$$

It follows by Lemma 4.3 of Newey and McFadden (1994), with $a(z, \theta)$ there equal to $\Delta(z, \beta) = [b(z, \beta) - b(z, \tilde{\beta})]' U(x)^{-1} [b(z, \beta) - b(z, \tilde{\beta})]$, and by Q idempotent that

$$\hat{T}_B \stackrel{\text{def}}{=} (\hat{B} - B)' Q (\hat{B} - B)/n \leqslant \|\hat{B} - B\|^2/n$$

$$\leqslant \sum_i \Delta(z_i, \hat{\beta})/n \xrightarrow{p} \mathrm{E}[\Delta(z_i, \tilde{\beta})] = 0.$$

Also, the same result holds $\hat{T}_A$ defined in the analogous way. For $X=(x_1,\ldots,x_n)$, let $a_i=a_i(\tilde{\beta})$, $\bar{a}_i=\mathrm{E}[a_i|x_i]$, and note that $\bar{A}\stackrel{\text{def}}{=}\mathrm{E}[A|X]=(\bar{a}_1'F_1,\ldots,\bar{a}_n'F_n)'$. Note that by i.i.d. observations,

$$\mathrm{E}[(A-\bar{A})(A-\bar{A})|X]=diag(F_1^{-1}Var(a_1|x_1)F_1^{-1},\ldots,F_n^{-1}Var(a_n|x_n)F_n^{-1})\leqslant CI.$$

By iterated expectations and $\mathrm{tr}(Q)\leqslant CK$, so for $\tilde{T}_A\stackrel{\text{def}}{=}(A-\bar{A})'Q(A-\bar{A})/n$,

$$\mathrm{E}[\tilde{T}_A]=\mathrm{E}[\mathrm{tr}(Q\mathrm{E}[(A-\bar{A})(A-\bar{A})'|X]Q)]/n\leqslant C\mathrm{E}[\mathrm{tr}(Q)]/n\leqslant CK/n\to 0.$$

Then $\tilde{T}_A\xrightarrow{\text{p}}0$ by M. Also, the same result holds for $\tilde{T}_B$ defined in the analogous way. By Assumption 1 there exists Γ_K such that $\mathrm{E}[\|U_i^{-1}\bar{a}_i-\Gamma_K q_i\|^2]\to 0$. Then for $\tilde{\gamma}_K=vec(\Gamma_K')$, by M,

$$\begin{aligned}\|\bar{A}-P\tilde{\gamma}_K\|^2/n&=\sum_i\|F_i^{-1}\bar{a}_i-(F_i\otimes q_i')\tilde{\gamma}_K\|^2/n\\&=\sum_i\|F_i\|^2\|U_i^{-1}\bar{a}_i-(I\otimes q_i')\tilde{\gamma}_K\|^2/n\\&=\sum_i\|F_i\|^2\|U_i^{-1}\bar{a}_i-\Gamma_K q_i\|^2/n\leqslant C\sum_i\|U_i^{-1}\bar{a}_i-\Gamma_K q_i\|^2/n\xrightarrow{\text{p}}0.\end{aligned}$$

It follows by $QP=P$ and $I-Q$ idempotent that

$$\bar{T}_A\stackrel{\text{def}}{=}\bar{A}'(I-Q)\bar{A}/n=(\bar{A}-P\tilde{\gamma}_K)'(I-Q)(\bar{A}-P\tilde{\gamma}_K)/n\leqslant\|\bar{A}-P\tilde{\gamma}_K\|^2/n\xrightarrow{\text{p}}0,$$

with the same result holding for the analogous term $\bar{T}_B$.

Next, note that by CS,

$$\begin{aligned}T_A&=(\hat{A}-\bar{A})'Q(\hat{A}-\bar{A})=(\hat{A}-A+A-\bar{A})'Q(\hat{A}-A+A-\bar{A})\\&\leqslant\hat{T}_A+\tilde{T}_A+2\sqrt{\hat{T}_A}\sqrt{\tilde{T}_A}\xrightarrow{\text{p}}0.\end{aligned}$$

Also, by M, $\bar{A}'\bar{A}/n=\mathrm{O_p}(1)$. The analogous results also hold for B replacing A. Then by the CS and T,

$$\begin{aligned}|\hat{A}'Q\hat{B}/n-\bar{A}'\bar{B}/n|&=|(\hat{A}-\bar{A})'Q(\hat{B}-\bar{B})+(\hat{A}-\bar{A})'Q\bar{B}\\&\quad+\bar{A}'Q(\hat{B}-\bar{B})-\bar{A}'(I-Q)\bar{B}|/n\\&\leqslant\sqrt{T_A}\sqrt{T_B}+\sqrt{T_A}\sqrt{\bar{B}'\bar{B}/n}+\sqrt{\bar{A}'\bar{A}/n}\sqrt{T_B}+\sqrt{\bar{T}_A}\sqrt{\bar{T}_B}\xrightarrow{\text{p}}0.\end{aligned}$$

Noting that $\bar{A}'\bar{B}/n = \sum_i \bar{a}_i' U_i^{-1}\bar{b}_i/n$, the conclusion follows by Khintchine's law of large numbers. □

Lemma A.4. *If Assumption* 1 *is satisfied,* ε_i *and* Y_i *are* $r \times 1$ *random vectors with* $E[\varepsilon_i|x_i] = 0$, $E[\|\varepsilon_i\|^2|x_i] \leqslant C$, *and* $E[\|Y_i\|^2|x_i] \leqslant C$, $U_i = U(x_i)$ *is a* $r \times r$ *p.d. matrix that is bounded and has smallest eigenvalue bounded away from zero,* $K \to \infty$, *and* $K^2/n \to 0$ *then*

$$\sum_i Y_i' \otimes q_i' \left(\sum_i U_i \otimes q_i q_i'\right)^{-} \sum_i \varepsilon_i \otimes q_i/\sqrt{n} - \sum_i E[Y_i|x_i]' U_i^{-1}\varepsilon_i/\sqrt{n} \xrightarrow{p} 0.$$

Proof. Let F_i and P be as specified in the proof of Lemma A.3, $A_i = F_i^{-1}Y_i$, $\bar{A}_i = E[A_i|x_i] = F_i^{-1}E[Y_i|x_i]$, $A = (A_1', \ldots, A_n')'$, $\bar{A} = (\bar{A}_1', \ldots, \bar{A}_n')'$, $B_i = F_i^{-1}\varepsilon_i$, and $B = (B_1', \ldots, B_n')'$. Then, similarly to Lemma A.3, by $E[b_i|x_i] = 0$,

$$\sum_i Y_i' \otimes q_i' \left(\sum_i U_i \otimes q_i q_i'\right)^{-} \sum_i \varepsilon_i \otimes q_i/\sqrt{n} - \sum_i E[Y_i|x_i]' U_i^{-1}\varepsilon_i/\sqrt{n}$$

$$= A'QB/\sqrt{n} - \bar{A}'B/\sqrt{n} = (A - \bar{A})'QB/\sqrt{n} - \bar{A}'(I - Q)B/\sqrt{n}.$$

It follows as in the proof of Lemma A.3 that $(A - \bar{A})'Q(A - \bar{A}) = O_p(K)$ and $B'QB = O_p(K)$ so that

$$|(A - \bar{A})'QB/\sqrt{n}| \leqslant \sqrt{(A - \bar{A})'Q(A - \bar{A})}\sqrt{B'QB}/\sqrt{n} = O_p(K/\sqrt{n}) \xrightarrow{p} 0.$$

Also, as in Lemma A.3, $E[\bar{A}'(I - Q)\bar{A}]/n \to 0$, so by iterated expectations

$$E[\|\bar{A}'(I - Q)B/\sqrt{n}\|^2] = E[\bar{A}'(I - Q)E[BB'|x](I - Q)\bar{A}]/n$$

$$\leqslant CE[\bar{A}'(I - Q)\bar{A}]/n \to 0.$$

the conclusion then follows M and T. □

For the statement of the next result, let

$$\hat{R}(\beta) = \hat{g}(\beta)'[\tilde{\Sigma}^{-1} \otimes \hat{A}^{-}]\hat{g}(\beta),$$

$$R(\beta) = E[E[\rho(z, \beta)|x]'\Sigma^{-1}E[\rho(z, \beta)|x]]. \qquad (A.1)$$

Lemma A.5. *If Assumptions* 1 *and* 3 *are satisfied,* $\tilde{\Sigma} \xrightarrow{p} \Sigma$, Σ *is positive definite,* $K \to \infty$, *and* $K/n \to 0$, *then* $R(\beta)$ *has a unique minimum at* β_0, $R(\beta)$ *is continuous on* B *and* $\sup_{\beta \in B} |\hat{R}(\beta) - R(\beta)| \xrightarrow{p} 0$.

Proof. By Σ p.d., for any $\beta \neq \beta_0$ it follows by $E[\rho(z, \beta)|x] \neq 0$ that

$$R(\beta) \geqslant CE[E[\rho(z, \beta)|x]'E[\rho(z, \beta)|x]] > 0 = R(\beta_0).$$

To show continuity of $R(\beta)$, note that by Σ^{-1} p.d. and CS,

$$\begin{aligned}|R(\tilde{\beta})-R(\beta)| &\leqslant \mathrm{E}[\mathrm{E}[\rho(z,\tilde{\beta})-\rho(z,\beta)|x]'\Sigma^{-1}\mathrm{E}[\rho(z,\tilde{\beta})-\rho(z,\beta)|x]] \\ &\leqslant C\mathrm{E}[\|\mathrm{E}[\rho(z,\tilde{\beta})-\rho(z,\beta)|x]\|^2] \leqslant C\mathrm{E}[\|\rho(z,\tilde{\beta})-\rho(z,\beta)\|^2] \\ &\leqslant C\mathrm{E}[\delta(z)^2]\|\tilde{\beta}-\beta\|^{2\alpha}.\end{aligned}$$

By Corollary 2.2 of Newey (1991) it suffices to show that (i) $\hat{R}(\beta)\xrightarrow{\text{p}}R(\beta)$ for each β in B and (ii) that there is $\hat{D}=\mathrm{O_p}(1)$ with $|\hat{R}(\tilde{\beta})-\hat{R}(\beta)|\leqslant\hat{D}\|\tilde{\beta}-\beta\|^{\alpha}$. To show (i), apply Lemma A.1 with $a(z,\beta)=a(z)=\rho_j(z,\beta)$ and $b(z,\beta)=b(z)=\rho_k(z,\beta)$, with β fixed, to obtain, for $\rho_j=(\rho_j(z_1,\beta),\ldots,\rho_j(z_n,\beta))'$ and $\bar{\rho}_{ji}=\mathrm{E}[\rho_j(z_i,\beta)|x_i]$,

$$\rho_j'Q\rho_k/n\xrightarrow{\text{p}}\mathrm{E}[\bar{\rho}_{ji}\bar{\rho}_{ki}].$$

Then by the continuous mapping theorem, $\tilde{\Sigma}^{-1}\xrightarrow{\text{p}}\Sigma^{-1}$ and

$$\hat{R}(\beta)=\sum_{j,k=1}^{J}(\hat{\Sigma}^{-1})_{jk}(\rho_j'Q\rho_k/n)\xrightarrow{\text{p}}\sum_{j,k=1}^{J}(\Sigma^{-1})_{jk}\mathrm{E}[\bar{\rho}_{ji}\bar{\rho}_{ki}]=R(\beta).$$

To show (ii), let $\rho_j(\beta)=(\rho_j(z_1,\beta),\ldots,\rho_j(z_n,\beta))'$ and $\tilde{\rho}_j=\rho_j(\tilde{\beta})$. Note that by Assumption 3 and M, $\hat{D}_\delta=[\sum_{i=1}^{n}\delta(z_i)^2/n]^{1/2}=\mathrm{O_p}(1)$. Also,

$$\|\tilde{\rho}_j-\rho_j\|/\sqrt{n}=\left[\sum_{i=1}^{n}(\tilde{\rho}_{ji}-\rho_{ji})^2/n\right]^{1/2}\leqslant\hat{D}_\delta\|\tilde{\beta}-\beta\|^{\alpha}.$$

Furthermore, for any fixed $\bar{\beta}\in B$, note that by M and Assumption 3, $\bar{D}=[\sum_{i=1}^{n}\rho(z_i,\bar{\beta})^2/n]^{1/2}=\mathrm{O_p}(1)$. Then

$$\begin{aligned}\sup_{\beta\in B}\|\rho_j(\beta)\|/\sqrt{n} &\leqslant \sup_{\beta\in B}\|\rho_j(\beta)-\rho_j(\bar{\beta})\|/\sqrt{n}+\bar{D} \\ &\leqslant \hat{D}_\delta\sup_{\beta\in B}\|\beta-\bar{\beta}\|^{\alpha}+\bar{D}\leqslant C\hat{D}_\delta+\bar{D}=\mathrm{O_p}(1).\end{aligned}$$

Then by the T and CS, and by Q idempotent, for $\hat{D}_\Sigma=\max_{j,k}|(\hat{\Sigma}^{-1})_{jk}|=\mathrm{O_p}(1)$

$$\begin{aligned}|\hat{R}(\tilde{\beta})-\hat{R}(\beta)| &\leqslant \hat{D}_\Sigma\sum_{j,k=1}^{J}|\tilde{\rho}_j'Q\tilde{\rho}_k-\rho_j'Q\rho_k|/n \\ &\leqslant \hat{D}_\Sigma\sum_{j,k=1}^{J}[|(\tilde{\rho}_j-\rho_j)Q\tilde{\rho}_k|+|\rho_jQ(\tilde{\rho}_k-\rho_k)|]/n \\ &\leqslant \hat{D}\|\tilde{\beta}-\beta\|^{\alpha},\ \hat{D}=\hat{D}_\Sigma 2J^2\hat{D}_\delta[C\hat{D}_\delta+\bar{D}]=\mathrm{O_p}(1),\end{aligned}$$

giving (ii). □

Proof of Theorem 5.1. By Lemma A.5 and compactness of B, all the hypotheses of Theorem 2.1 of Newey and McFadden (1994) are satisfied, the conclusion of which gives the result. □

Proof of Theorem 5.2. By $\beta_0 \in int(B)$ and Theorem 2.1, $\hat{\beta} \in int(B)$ w.p.a.1. Therefore, w.p.a.1 the first order conditions are satisfied, i.e.

$$\partial\hat{R}(\hat{\beta})/\partial\beta = [\partial\hat{g}(\hat{\beta})/\partial\beta]'\hat{W}\hat{g}(\hat{\beta}) = 0.$$

By a mean-value expansion in β, there is $\bar{\beta}$ on the line joining $\hat{\beta}$ and β_0 such that

$$[\partial^2\hat{R}(\bar{\beta})/\partial\beta\partial\beta'](\hat{\beta} - \beta_0) + \partial\hat{R}(\beta_0)/\partial\beta = 0. \tag{A.2}$$

By the chain rule $\partial^2\hat{R}(\bar{\beta})/\partial\beta\partial\beta' = [\partial\hat{g}(\bar{\beta})/\partial\beta]'\hat{W}[\partial\hat{g}(\bar{\beta})/\partial\beta] + \bar{F}$, where $\bar{F}$ is the $p \times p$ matrix with rth column $[\partial^2\hat{g}(\bar{\beta})/\partial\beta\partial\beta_r]\hat{W}\hat{g}(\bar{\beta})$. Note also that by Assumption 3, $\mathrm{E}[\|\rho(z,\beta_0)\|^2|x]$ is bounded, $\rho(z,\beta)$ is continuous at β_0, and

$$\begin{aligned}\sup_{\beta\in B}\|\rho(z,\beta)\|^2 &\leqslant C\left[\sup_{\beta\in B}\|\rho(z,\beta)-\rho(z,\beta_0)\|^2 + \|\rho(z,\beta_0)\|^2\right]\\ &\leqslant C[\delta(z)^2 + \|\rho(z,\beta_0)\|^2].\end{aligned}$$

Then by Assumptions 1, 3, and 4 the hypotheses of Lemma A.3 are satisfied for $\hat{\beta} = \bar{\beta}$, $\tilde{\beta} = \beta_0$, $a(z,\beta) = \partial^2\rho_j(z,\beta)/\partial\beta_r\partial\beta_s$, and $b(z,\beta) = \rho_k(z,\beta)$. Then by Lemma A.3 $a(\hat{\beta})'Qb(\hat{\beta})/n \xrightarrow{p} 0$. It then follows by $\tilde{\Sigma}^{-1} \xrightarrow{p} \Sigma^{-1}$ and the continuous mapping theorem that $\bar{F} \xrightarrow{p} 0$. Similarly the hypotheses of Lemma A.3 are satisfied for $\hat{\beta} = \bar{\beta}$, $\tilde{\beta} = \beta_0$, $a(z,\beta) = \partial\rho_j(z,\beta)/\partial\beta_r$ and $b(z,\beta) = \partial\rho_k(z,\beta)/\partial\beta_s$, so by Lemma A.3 $a(\hat{\beta})'Qb(\hat{\beta})/n \xrightarrow{p} \mathrm{E}[D_{jr}(x)D_{ks}(x)]$. It then follows by $\tilde{\Sigma}^{-1} \xrightarrow{p} \Sigma^{-1}$ and the continuous mapping theorem that $[\partial\hat{g}(\bar{\beta})/\partial\beta]'\hat{W}[\partial\hat{g}(\bar{\beta})/\partial\beta] \xrightarrow{p} \mathrm{E}[D(x)'\Sigma^{-1}D(x)]$, so that

$$\partial^2\hat{R}(\bar{\beta})/\partial\beta\partial\beta' \xrightarrow{p} \mathrm{E}[D(x)'\Sigma^{-1}D(x)]. \tag{A.3}$$

Next, for each r, $(r = 1,\ldots,p)$, let $Y_{ij} = \partial\rho_j(z_i,\beta_0)/\partial\beta_r$, $\varepsilon_{ik} = \rho_k(z_i,\beta_0)$, and $U_i = 1$. Then the hypotheses of Lemma A.4 are satisfied, so that for $Y_j = (Y_{1j},\ldots,Y_{nj})'$,

$$Y_j'Q\varepsilon_k/\sqrt{n} = \sum_i D_{jr}(x_i)\rho_k(z_i,\beta_0)/\sqrt{n} + o_p(1).$$

It follows from $\tilde{\Sigma}_{jk}^{-1} = O_p(1)$ that,

$$\sqrt{n}\partial\hat{R}(\beta_0)/\partial\beta_r = \sum_{j,k}\tilde{\Sigma}_{jk}^{-1}(Y_j'Q\varepsilon_k/\sqrt{n}) = \sum_{j,k}\tilde{\Sigma}_{jk}^{-1}\sum_i D_{jr}(x_i)\varepsilon_{ik}/\sqrt{n} + o_p(1).$$

Also, the Slutzky Lemma and consistency of $\tilde{\Sigma}_{jk}^{-1}$ give

$$\begin{aligned}\sum_{j,k}\tilde{\Sigma}_{jk}^{-1}\sum_i D_{jr}(x_i)\varepsilon_{ik}/\sqrt{n} &= \sum_{j,k}\Sigma_{jk}^{-1}\sum_i D_{jr}(x_i)\varepsilon_{ik}/\sqrt{n} + o_p(1)\\ &= \sum_i D_r(x_i)'\Sigma^{-1}\rho(z_i,\beta_0)/\sqrt{n} + o_p(1)\end{aligned}$$

where $D_r(x)$ is the rth column of $D(x)$. Note also that $D(x_i)$ is bounded, so that $Z_i = D(x_i)'\Sigma^{-1}\rho(z_i,\beta_0)$ has finite second moment, and that it has mean zero by iterated expectations. Therefore, by the Lindbergh–Levy central limit and Slutzky theorems,

$$\sqrt{n}\partial\hat{R}(\beta_0)/\partial\beta = \sum_{i=1}^{n} D(x_i)'\Sigma^{-1}\rho(z_i,\beta_0)/\sqrt{n} + o_p(1)$$

$$\xrightarrow{d} N(0, E[D(x)'\Sigma^{-1}\Sigma(x)\Sigma^{-1}D(x)]). \tag{A.4}$$

The remainder of the asymptotic normality proof follows from Eqs. (A.2)–(A.4) by standard arguments. Consistency of the asymptotic variance estimator follows similarly to the proof of $\partial\hat{g}(\bar{\beta})/\partial\beta'\hat{W}\partial\hat{g}(\bar{\beta})/\partial\beta \xrightarrow{p} E[D(x)'\Sigma^{-1}D(x)]$ given above, with $\hat{\beta}$ replacing $\bar{\beta}$. □

For the purposes of the next several results, let $\hat{\beta}$ be some $p \times 1$ random vector (not necessarily equal to the estimators we have considered), $\hat{g}_i = g_i(\hat{\beta})$, $g_i = g_i(\beta_0)$, $\Sigma_i = \Sigma(x_i)$, and

$$\hat{\Omega} = \sum_i \hat{g}_i\hat{g}_i'/n, \quad \tilde{\Omega} = \sum_i g_i g_i'/n, \quad \bar{\Omega} = \sum_i \Sigma_i \otimes q_i q_i'/n, \quad \Omega = E[g_i g_i'].$$

Lemma A.6. *If Assumption* 2 *and* 5(b) *are satisfied and* $\hat{\beta} = \beta_0 + O_p(\tau_n)$ *with* $\tau_n \to 0$ *then*

$$\|\hat{\Omega} - \tilde{\Omega}\| = O_p(\tau_n K), \quad \|\tilde{\Omega} - \bar{\Omega}\| = O_p(\zeta(K)\sqrt{K/n}),$$

$$\|\bar{\Omega} - \Omega\| = O_p(\zeta(K)\sqrt{K/n}).$$

If Assumption 5(a) *is also satisfied then* $1/C \leqslant \lambda_{\min}(\Omega) \leqslant \lambda_{\max}(\Omega) \leqslant C$, *and if* $\tau_n K + \zeta(K)\sqrt{K/n} \to 0$ *then w.p.a.1,* $1/C \leqslant \lambda_{\min}(\hat{\Omega}) \leqslant \lambda_{\max}(\hat{\Omega}) \leqslant C$, $1/C \leqslant \lambda_{\min}(\bar{\Omega}) \leqslant \lambda_{\max}(\bar{\Omega}) \leqslant C$.

Proof. By Lemma A.2, $E[\|q_i\|^4] \leqslant C\zeta(K)^2 E[\|q_i\|^2] = C\zeta(K)^2 K$. For $\delta_i = \delta(z_i)$, $\rho_i = \rho(z_i,\beta_0)$, and $\hat{\rho}_i = \rho(z_i,\hat{\beta})$ we have $\|\hat{\rho}_i - \rho_i\| \leqslant \delta_i\|\hat{\beta} - \beta_0\|$ for each $i = 1,\ldots,n$ w.p.a.1. Also, note that $M_i = \delta_i^2 + 2\delta_i(z_i)\|\rho_i\|$ has $E[M_i|x_i]$ bounded by CS, so that $E[M_i\|q_i\|^2] = E[E[M_i|x_i]\|q_i\|^2] \leqslant CE[\|q_i\|^2] \leqslant CK$. Then by the T, CS, and M, w.p.a.1

$$\|\hat{\Omega} - \tilde{\Omega}\| \leqslant \sum_i \|\hat{\rho}_i\hat{\rho}_i' - \rho_i\rho_i'\|\|q_i\|^2/n$$

$$\leqslant \sum_i (\|\hat{\rho}_i - \rho_i\|^2 + 2\|\hat{\rho}_i - \rho_i\|\|\rho_i\|)\|q_i\|^2/n$$

$$\leqslant C\|\hat{\beta} - \beta_0\| \sum_i M_i\|q_i\|^2/n = O_p(\tau_n E[M_i\|q_i\|^2]) = O_p(\tau_n K),$$

giving the first conclusion. Also,

$$\mathrm{E}[\|\tilde{\Omega}-\bar{\Omega}\|^2] = \mathrm{E}\left[\left\|\sum_i (\rho_i\rho_i' - \Sigma_i)\otimes q_iq_i'/n\right\|^2\right]$$
$$= tr\,\mathrm{E}[(\rho_i\rho_i' - \Sigma_i)^2 \otimes \{q_iq_i'\}^2/n]$$
$$\leqslant tr\,\mathrm{E}[\|\rho_i\|^2\rho_i\rho_i' \otimes \|q_i\|^2 q_iq_i'/n]$$
$$= \mathrm{E}[\mathrm{E}[\|\rho_i\|^4|x_i]\|q_i\|^4]/n \leqslant C\zeta(K)^2K/n,$$

so the second conclusion follows by M. The third conclusion follows by M and

$$\mathrm{E}[\|\bar{\Omega}-\Omega\|^2] = \mathrm{E}\left[\left\|\sum_i \Sigma_i\otimes q_iq_i'/n - \Omega\right\|^2\right] \leqslant tr\,\mathrm{E}[\Sigma_i^2\otimes\{q_iq_i'\}^2]/n$$
$$\leqslant C\mathrm{E}[\|q_i\|^4]/n = C\zeta(K)^2K/n.$$

For the fourth conclusion, note that $C^{-1}I_J \leqslant \Sigma_i \leqslant CI_J$, so that

$$C^{-1}I_{JK} = C^{-1}\mathrm{E}[I_J\otimes q_iq_i'] \leqslant \Omega \leqslant C\mathrm{E}[I\otimes q_iq_i'] = CI.$$

Hence, $C^{-1} \leqslant \lambda_{\min}(\Omega) \leqslant \lambda_{\max}(\Omega) \leqslant C$. Also, note that if $\tau_nK + \zeta(K)\sqrt{K/n} \to 0$ then $\|\hat{\Omega}-\bar{\Omega}\| \xrightarrow{p} 0$ and $\|\bar{\Omega}-\Omega\| \xrightarrow{p} 0$ by the first two conclusions, so $\|\hat{\Omega}-\Omega\| \xrightarrow{p} 0$ by M. Then, by $|\lambda(A)-\lambda(B)| \leqslant \|A-B\|$, where $\lambda(A)$ denotes the minimum or maximum eigenvalue, it follows that

$$|\lambda_{\min}(\hat{\Omega})-\lambda_{\min}(\Omega)| \xrightarrow{p} 0, \quad |\lambda_{\max}(\hat{\Omega})-\lambda_{\max}(\Omega)| \xrightarrow{p} 0,$$

giving the fourth conclusion. The other conclusions follow similarly. □

Let $\rho_\beta(z,\beta) = \partial\rho(z,\beta)/\partial\beta$, $D(x) = \mathrm{E}[\rho_\beta(z,\beta_0)|x]$, $D_i = D(x_i)$, and

$$\hat{G} = \sum_i \rho_\beta(z_i,\hat{\beta})\otimes q_i/n, \bar{G} = \sum_i D_i\otimes q_i/n, G = \mathrm{E}[D_i\otimes q_i].$$

Lemma A.7. *If Assumptions* 2 *and* 5(b) *are satisfied and* $\hat{\beta}=\beta_0+\mathrm{O_p}(\tau_n)$ *with* $\tau_n \to 0$ *then*

$$\|\hat{G}-\bar{G}\| = \mathrm{O_p}(\tau_n\sqrt{K}+\sqrt{K/n}), \quad \|\bar{G}-G\| = \mathrm{O_p}(\sqrt{K/n}).$$

Proof. Let $\rho_{\beta i} = \rho_\beta(z_i,\beta_0)$, $\delta_i = \delta(z_i)$, and $\tilde{G} = \sum_i \rho_\beta(z_i,\beta_0)\otimes q_i/n$. Then by Lemma A.2,

$$\mathrm{E}[\|\tilde{G}-\bar{G}\|^2] = \mathrm{E}\left[\left\|\sum_i(\rho_{\beta i}-D_i)\otimes q_i/n\right\|^2\right]$$
$$= tr\,\mathrm{E}[(\rho_{\beta i}-D_i)'(\rho_{\beta i}-D_i)\|q_i\|^2/n]$$
$$\leqslant \mathrm{E}[\mathrm{E}[\|\rho_{\beta i}\|^2|x_i]\|q_i\|^2]/n \leqslant CK/n.$$

Then $\|\tilde{G}-\bar{G}\| = O_p(\sqrt{K/n})$ by M. Also, by the T, CS, and M, w.p.a.1

$$\|\hat{G}-\tilde{G}\| \leqslant \sum_i \|\rho_\beta(z_i,\hat{\beta}) - \rho_{\beta i}\| \, \|q_i\|/n \leqslant \|\hat{\beta}-\beta_0\| \sum_i \delta_i \|q_i\|/n$$

$$= O_p(\tau_n E[E[\delta_i|x_i]\|q_i\|]) = O_p(\tau_n \{E[\|q_i\|^2]\}^{1/2}) = O_p(\tau_n\sqrt{K}).$$

The first conclusion follows by T. The second conclusion follows by M and D_i bounded,

$$E[\|\bar{G}-G\|^2] = E\left[\left\|\sum_i D_i \otimes q_i/n - G\right\|^2\right] \leqslant E[\|D_i\|^2\|q_i\|^2]/n$$

$$\leqslant CE[\|q_i\|^2]/n = CK/n. \qquad \square$$

Lemma A.8. *If Assumption* 2, 4, *and* 5 *are satisfied and* $\hat{\beta} = \beta_0 + O_p(\tau_n)$ *with* $\tau_n K + \zeta(K)\sqrt{K/n} \to 0$ *then*

$$\|\bar{G}'\bar{\Omega}^{-1}(\hat{\Omega}-\bar{\Omega})\| = O_p(\tau_n\sqrt{K} + \zeta(K)/\sqrt{n}).$$

Proof. By Lemma A.3, with $U(x)=\Sigma(x)$, $a(z,\beta)$ equal to the rth column of $D(x)$, and $b(z,\beta)$ equal to the sth column of $D(x)$, it follows that to obtain $\bar{G}'\bar{\Omega}^{-1}\bar{G} \xrightarrow{p} E[D(x)'\Sigma(x)^{-1}D(x)]$ and (hence) $\bar{G}'\bar{\Omega}^{-1}\bar{G} = O_p(1)$. Let $H_i = \bar{G}'\bar{\Omega}^{-1}(I_J \otimes q_i)$. Then by $CI_J \leqslant \Sigma_i$,

$$\sum_i \|H_i\|^2/n = tr\left(\sum_i H_i H_i'/n\right) \leqslant C\, tr(\bar{G}'\bar{\Omega}^{-1}\bar{G}) = O_p(1).$$

Next, let $M_i = \delta_i^2 + 2\delta_i\|\rho_i\|$ and $\hat{R}_n = \sum_i M_i \|H_i\|\|q_i\|/n$. Let $X = (x_1,\ldots,x_n)$. It is well known that if $E[\hat{R}_n|X] = O_p(\alpha_n)$ for some α_n then $\hat{R}_n = O_p(\alpha_n)$. By CS and M

$$E[\hat{R}_n|X] \leqslant C\sum_i \|H_i\|\|q_i\|/n \leqslant C\left(\sum_i \|H_i\|^2/n\right)^{1/2}\left(\sum_i \|q_i\|^2/n\right)^{1/2}$$

$$= O_p(\{E[\|q_i\|^2]\}^{1/2}) = O_p(\sqrt{K}),$$

so that $R_n = O_p(K^{1/2})$. Therefore, by T and CS,

$$\|\bar{G}'\bar{\Omega}^{-1}(\hat{\Omega}-\tilde{\Omega})\| \leqslant \sum_i \|H_i\|\|\hat{\rho}_i\hat{\rho}_i' - \rho_i\rho_i'\|\|q_i\|/n \leqslant \|\hat{\beta}-\beta_0\|\hat{R}_n = O_p(\tau_n\sqrt{K}).$$

It also follows similarly to the proof of Lemma A.5 that

$$E[\|\bar{G}'\bar{\Omega}^{-1}(\tilde{\Omega}-\bar{\Omega})\|^2|X] \leqslant \sum_i E[\|\rho_i\|^4|x_i]\|H_i\|^2\|q_i\|^2/n^2$$

$$\leqslant C[\zeta(K)^2/n]\sum_i \|H_i\|^2/n = O_p(\zeta(K)^2/n),$$

so that $\|\bar{G}'\bar{\Omega}^{-1}(\tilde{\Omega}-\bar{\Omega})\| = O_p(\zeta(K)/\sqrt{n})$. The conclusion follows by T. $\square$

Lemma A.9. *If Assumptions* 2 *and* 3 *are satisfied and* $\Sigma(x)$ *is bounded then* $\|\hat{g}(\beta_0)\| = O_p(\sqrt{K/n})$.

Proof. By Lemma A.2,

$$E[\|\hat{g}(\beta_0)\|^2] = E[\|\rho_i\|^2\|q_i\|^2]/n = E[tr(\Sigma_i)\|q_i\|^2]/n \leqslant CE[\|q_i\|^2]/n = CK/n,$$

so the conclusion follows by M. □

We now prove the GMM results. Here let $\hat{\Omega} = \hat{\Omega}(\tilde{\beta})$.

Proof of Theorem 5.3. Let $\hat{g} = \hat{g}(\hat{\beta})$ and $\bar{g} = \hat{g}(\beta_0)$. By Assumption 5 and $\tilde{\beta} = \hat{\beta} + O_p(1/\sqrt{n})$ all of the hypotheses of Lemma A.6 are satisfied for $\hat{\beta} = \tilde{\beta}$ and $\tau_n = 1/\sqrt{n}$. Let $\hat{W} = \hat{\Omega}^{-1}$. Then by Lemma A.6, $\lambda_{\max}(\hat{W}) = \lambda_{\min}(\hat{\Omega})^{-1} \leqslant C$ and $\lambda_{\min}(\hat{W}) = \lambda_{\max}(\hat{\Omega})^{-1} \geqslant C$ w.p.a.1. Also, define $\tilde{W} = I_J \otimes \hat{A}^{-}$. It follows from Lemma A.6 for the special case where $\Sigma_i = I_J$ (and hence $\tilde{W} = \bar{\Omega}$) that $\lambda_{\max}(\tilde{W}) \leqslant C$ w.p.a.1. Then by Lemma A.9 and the definition of $\hat{\beta}$, w.p.a.1

$$\hat{g}'\tilde{W}\hat{g} \leqslant C\|\hat{g}\|^2 \leqslant C\hat{g}'\hat{W}\hat{g} \leqslant C\bar{g}'\hat{W}\bar{g} \leqslant C\|\bar{g}\|^2 = O_p(K/n) \xrightarrow{p} 0.$$

Now, note that $\hat{g}'\tilde{W}\hat{g} = \hat{R}(\beta)$ as in Eq. (A.1), for $\tilde{\Sigma} = I_J$. Then, by Lemma A.5 all of the conditions of Lemma A.1 are satisfied, and its conclusion gives the result. □

Proof of Theorem 5.4. Let $\hat{S}(\beta) = \hat{g}(\beta)'\hat{\Omega}^{-1}\hat{g}(\beta)/2$. By consistency of $\hat{\beta}$, it follows by the first-order conditions and a mean-value expansion that w.p.a.1,

$$[\partial^2\hat{S}(\bar{\beta})/\partial\beta\partial\beta'](\hat{\beta} - \beta_0) + \partial\hat{S}(\beta_0)/\partial\beta = 0, \tag{A.5}$$

where $\bar{\beta}$ denotes a mean value and, by the chain rule, $\partial^2\hat{S}(\bar{\beta})/\partial\beta\partial\beta' = \hat{G}'\hat{\Omega}^{-1}\hat{G} + \bar{F}$, for $\hat{G} = \partial\hat{g}(\bar{\beta})/\partial\beta$ and $\bar{F}$ the $p \times p$ matrix with rth column $[\partial^2\hat{g}(\bar{\beta})/\partial\beta\partial\beta_r]\hat{\Omega}^{-1}\hat{g}(\bar{\beta})$. Note that by Assumptions 4 and 5, $E[D(x)'\Sigma(x)^{-1}D(x)] \geqslant CE[D(x)'D(x)]$, so that $V = \{E[D(x)'\Sigma(x)^{-1}D(x)]\}^{-1}$ exists. Now successively apply Lemma A.3 with $\tilde{\beta} = \beta_0$, $\hat{\beta} = \bar{\beta}$, $a(z,\beta) = \partial\rho(z,\beta)/\partial\beta_r$, $b(z,\beta) = \partial\rho(z,\beta)/\partial\beta_s$, and $U(x) = \Sigma(x)$, for $r,s = 1,\ldots,J$, to obtain $\hat{G}'\bar{\Omega}^{-1}\hat{G} \xrightarrow{p} V^{-1}$. Also, by Lemma A.6, $\lambda_{\max}(\bar{\Omega}^{-1}) = \lambda_{\min}(\bar{\Omega})^{-1} \leqslant C$ w.p.a.1. Thus, for $\hat{B} = \bar{\Omega}^{-1}\hat{G}$,

$$\|\hat{B}\|^2 = tr(\hat{B}'\hat{B}) \leqslant Ctr(\hat{G}'\bar{\Omega}^{-1}\hat{G}) = O_p(1).$$

Also, by Lemma A.6 $\lambda_{\max}(\hat{\Omega}^{-1}) \leqslant C$ w.p.a.1, so by Lemma A.6, T, and CS

$$\begin{aligned}\|\hat{G}'\hat{\Omega}^{-1}\hat{G} - \hat{G}'\bar{\Omega}^{-1}\hat{G}\| &= \|\hat{B}'\{\bar{\Omega} - \hat{\Omega} + (\bar{\Omega} - \hat{\Omega})\hat{\Omega}^{-1}(\bar{\Omega} - \hat{\Omega})\hat{B}\| \\ &\leqslant \|\hat{B}\|^2(\|\bar{\Omega} - \hat{\Omega}\| + C\|\bar{\Omega} - \hat{\Omega}\|^2) \xrightarrow{p} 0.\end{aligned}$$

Then by T,

$$\hat{G}'\hat{\Omega}^{-1}\hat{G} \xrightarrow{p} V^{-1}. \tag{A.6}$$

With analogous arguments, e.g. applying Lemma A.1 with $a(z,\beta)=\partial^2\rho(z,\beta)/\partial\beta_s\partial\beta_r$ and $b(z,\beta)=\rho(z,\beta)$, gives $\bar{F}\xrightarrow{p}0$. It then follows by the triangle inequality that

$$\partial^2\hat{S}(\bar{\beta})/\partial\beta\partial\beta'\xrightarrow{p}V^{-1}. \tag{A.7}$$

Next, consider $\partial\hat{S}(\beta_0)/\partial\beta=\tilde{G}'\hat{\Omega}^{-1}\hat{g}(\beta_0)$, where $\tilde{G}=\partial\hat{g}(\beta_0)/\partial\beta$. For $\bar{G}=\sum_i D_i\otimes q_i/n$, it follows by Lemma A.7 with $\tilde{\beta}=\beta_0$ and $\tau_n=0$ that $\|\tilde{G}-\bar{G}\|=o_p(\sqrt{K/n})$. By Lemmas A.6 and A.9, w.p.a.1 $\|\hat{\Omega}^{-1}\bar{g}\|^2\leqslant C\|\bar{g}\|=O_p(\sqrt{K/n})$.

$$\begin{aligned}\|(\tilde{G}'\hat{\Omega}^{-1}-\bar{G}'\bar{\Omega}^{-1})\bar{g}\| &\leqslant (\|\tilde{G}-\bar{G}\|+\|\bar{G}'\bar{\Omega}^{-1}(\hat{\Omega}-\bar{\Omega})\|)\|\hat{\Omega}^{-1}\bar{g}\|\\ &= O_p([\sqrt{K}+\zeta(K)]/\sqrt{n})O_p(\sqrt{K/n})=o_p(1/\sqrt{n}).\end{aligned}$$

Furthermore, by Lemma A.4, with Y_i equal to $D(x_i)$, $\varepsilon_i=\rho_i$, and $U_i=\Sigma(x_i)$,

$$\bar{G}'\bar{\Omega}^{-1}\bar{g}-\sum_i D(x_i)'\Sigma(x_i)^{-1}\rho_i/n=o_p(1/\sqrt{n}).$$

By the Lindbergh–Levy central limit theorem, $\sum_i D(x_i)'\,\Sigma(x_i)^{-1}\rho_i/\sqrt{n}\xrightarrow{d}N(0,V^{-1})$, so it follows by the triangle inequality and the Slutzky theorem that

$$\sqrt{n}\partial\hat{S}(\beta_0)/\partial\beta\xrightarrow{d}N(0,V^{-1}).$$

The first conclusion follows by Eq. (A.7) and the Slutzky and continuous mapping theorems in the usual way. Also, consistency of $\hat{V}$ follows exactly as in the proof of Eq. (A.6). □

The following results are useful for the proofs for GEL. Let $\hat{S}(\beta,\lambda)=\sum_i s(\lambda' g_i(\beta))/n$.

Lemma A.10. *If Assumption* 6 *is satisfied then for any* $\delta_n=o(n^{-1/\gamma}\zeta(K)^{-1})$ *and* $\Lambda_n=\{\lambda:\|\lambda\|\leqslant\delta_n\}$ *we have* $\max_{\beta\in\mathscr{B},\lambda\in\Lambda_n,i\leqslant n}|\lambda' g_i(\beta)|\xrightarrow{p}0$ *and w.p.a.1* $\Lambda_n\subset\hat{\Lambda}(\beta)$ *for all* $\beta\in\mathscr{B}$.

Proof. For $b_i=\sup_{\beta\in\mathscr{B}}\|\rho(z_i,\beta)\|$, it follows by M that $\max_{i\leqslant n}\delta_i=O_p(n^{1/\gamma})$, so

$$\max_{\beta\in\mathscr{B},\lambda\in\Lambda_n,i\leqslant n}|\lambda' g_i(\beta)|\leqslant\delta_n\zeta(K)\max_{i\leqslant n}b_i\xrightarrow{p}0,$$

giving the first conclusion. Also, by the first conclusion, w.p.a.1 $\lambda' g_i(\beta)\in\mathscr{V}$ for all $\beta\in\mathscr{B}$ and $\lambda\in\Lambda_n$, giving the second conclusion. □

Lemma A.11. *If Assumptions* 2, 5, *and* 6 *are satisfied*, $\tilde{\beta}=\beta_0+O_p(\tau_n)$, $\tau_n K\to 0$, *and* $\|\hat{g}(\tilde{\beta})\|=O_p(\sqrt{K/n})$ *then* $\sup_{\lambda\in\hat{\Lambda}(\tilde{\beta})}\hat{S}(\tilde{\beta},\lambda)\leqslant s_0+O_p(K/n)$, $\tilde{\lambda}=\operatorname{argmax}_{\lambda\in\hat{\Lambda}(\tilde{\beta})}\hat{S}(\tilde{\beta},\lambda)$ *exists w.p.a.1*, *and* $\|\tilde{\lambda}\|=O_p(\sqrt{K/n})$.

Proof. Choose $\delta_n=o(n^{-1/\gamma}\zeta(K)^{-1})$ and $\sqrt{K/n}=o(\delta_n)$, which is possible by $\zeta(K)^2K/n^{1-2/\gamma}\to 0$. Then for Λ_n as in the statement of Lemma A.10, it follows by the conclusion of Lemma A.10 that $\hat{S}(\tilde{\beta},\lambda)$ is twice continuously differentiable on Λ_n, w.p.a.1. Then $\bar{\lambda}=\arg\max_{\lambda\in\Lambda_n}\hat{S}(\tilde{\beta},\lambda)$ exists w.p.a.1. Let $\tilde{g}_i=g_i(\tilde{\beta})$, and $\tilde{g}=\hat{g}(\tilde{\beta})$. By assumption we have $\|\tilde{g}\| = O_p(\sqrt{K/n})$. Also, by Assumption 6(b) $\zeta(K)\sqrt{K/n} \to 0$, so by Lemma A.6, $\lambda_{\min}(\sum_i \tilde{g}_i\tilde{g}_i'/n) \geqslant C$ w.p.a.1. Furthermore, by Lemma A.10 and Assumption 6(a), for any $\dot{\lambda}$ on the line joining $\bar{\lambda}$ and 0, w.p.a.1 $\max_{i\leqslant n} s_2(\dot{\lambda}'\tilde{g}_i) \leqslant -C$. By a Taylor expansion around $\lambda = 0$ with Lagrange remainder, w.p.a.1.

$$s_0 = \hat{S}(\tilde{\beta},0) \leqslant \hat{S}(\tilde{\beta},\bar{\lambda}) = s_0 - \bar{\lambda}'\tilde{g} + \bar{\lambda}'\left[\sum_i s_2(\dot{\lambda}'g_i)\tilde{g}_i\tilde{g}_i'/n\right]\bar{\lambda}/2$$

$$\leqslant s_0 + \|\bar{\lambda}\|\|\tilde{g}\| - C\bar{\lambda}'\left(\sum_i \tilde{g}_i\tilde{g}_i'/n\right)\bar{\lambda} \leqslant s_0 + \|\bar{\lambda}\|\|\tilde{g}\| - C\|\bar{\lambda}\|^2.$$

Then subtracting $s_0 - C\|\bar{\lambda}\|^2$ from both sides and dividing by $\|\bar{\lambda}\|$, we obtain $C\|\bar{\lambda}\| \leqslant \|\tilde{g}\|$. It follows from this that $\|\bar{\lambda}\| = O_p(\sqrt{K/n})$, so that w.p.a.1 $\|\bar{\lambda}\| < \delta_n$, i.e. $\bar{\lambda} \in int(\Lambda_n)$. It then follows that $\partial\hat{S}(\tilde{\beta},\bar{\lambda})/\partial\lambda = 0$. Also, since $\Lambda_n \subset \hat{\Lambda}(\tilde{\beta})$, we also have $\bar{\lambda} \in \hat{\Lambda}(\beta_0)$. By concavity of $\hat{S}(\tilde{\beta},\lambda)$ and convexity of $\hat{\Lambda}(\tilde{\beta})$ it then follows that $\hat{S}(\tilde{\beta},\bar{\lambda}) = \max_{\lambda\in\hat{\Lambda}(\tilde{\beta})} \hat{S}(\tilde{\beta},\lambda)$, giving the second and third conclusions with $\tilde{\lambda} = \bar{\lambda}$. The last inequality of the above equation then gives $\hat{S}(\tilde{\beta},\tilde{\lambda}) \leqslant s_0 + \|\tilde{\lambda}\|\|\tilde{g}\| - C\|\tilde{\lambda}\|^2 = s_0 + O_p(K/n)$. □

Lemma A.12. *If Assumptions* 2, 5, *and* 6 *are satisfied and* $K\zeta(K)^2/n \to 0$ *then for any* $\delta_n = o(n^{-1/\gamma}\zeta(K)^{-1})$, $\Lambda_n = \{\lambda\colon \|\lambda\| \leqslant \delta_n\}$, $\tilde{\beta} \in \mathscr{B}$, *and* $\tilde{\lambda} \in \Lambda_n$, *it is the case that* $\lambda_{\max}(-\sum_{i=1}^n s_2(\tilde{\lambda}'\tilde{g}_i)\tilde{g}_i\tilde{g}_i'/n) \leqslant C$ *w.p.a.1.*

Proof. Let $\tilde{g}_i = g_i(\tilde{\beta})$ and $\tilde{\rho}_i = \rho(z_i,\tilde{\beta})$. By Lemma A.10 and Assumption 6(a), w.p.a.1 $\max_{i\leqslant n} - s_2(\tilde{\lambda}'\tilde{g}_i) \leqslant C$, so that $-\sum_{i=1}^n s_2(\tilde{\lambda}'\tilde{g}_i)\tilde{g}_i\tilde{g}_i'/n \leqslant C\sum_{i=1}^n \tilde{g}_i\tilde{g}_i'/n$ w.p.a.1. Also, for $b_i = \sup_{\beta\in\mathscr{B}}\|\rho(z_i,\beta)\|$, by CS, $\tilde{\rho}_i\tilde{\rho}_i' \leqslant Cb_i^2 I_J$, so that w.p.a.1 $\sum_{i=1}^n \tilde{g}_i\tilde{g}_i'/n \leqslant C\tilde{\Omega}$, $\tilde{\Omega} = \sum_{i=1}^n (I_J \otimes q_iq_i')b_i^2/n$. Then similarly to the proof of Lemma A.6 it follows that $\|\tilde{\Omega} - E[\tilde{\Omega}]\| \xrightarrow{p} 0$. Since

$$E[\tilde{\Omega}] = E[b_i^2(I_J \otimes q_iq_i')] = E[E[b_i^2|x_i](I_J \otimes q_iq_i')] \leqslant CI_{JK},$$

it follows similarly to the proof of Lemma A.6 that $\lambda_{\max}(\tilde{\Omega}) \leqslant C$, implying $\tilde{\Omega} \leqslant CI_{JK}$ w.p.a.1. The conclusion then follows by the inequalities previously shown in this proof. □

Lemma A.13. *If Assumptions* 2, 3, 5, *and* 6 *are satisfied then for any* $\bar{\lambda} \in \hat{\Lambda}(\hat{\beta})$ *it is the case that w.p.a.1,*

$$\hat{S}(\hat{\beta},\bar{\lambda}) \leqslant \sup_{\lambda\in\hat{\Lambda}(\hat{\beta})} \hat{S}(\hat{\beta},\lambda) \leqslant s_0 + O_p(K/n).$$

Proof. The first inequality is obvious. Also, by Lemma A.9 the hypotheses of Lemma A.11 are satisfied for $\tilde{\beta} = \beta_0$, so that $\sup_{\lambda\in\hat{\Lambda}(\beta_0)}\hat{S}(\beta_0,\lambda) \leqslant s_0 + O_p(K/n)$. Then by the definition of $\hat{\beta}$ we have $\sup_{\lambda\in\hat{\Lambda}(\hat{\beta})}\hat{S}(\hat{\beta},\lambda) \leqslant \sup_{\lambda\in\hat{\Lambda}(\beta_0)}\hat{S}(\beta_0,\lambda)$, giving the second inequality. □

Lemma A.14. *If Assumptions* 2, 3, 5, *and* 6 *are satisfied then* $\|\hat{g}(\hat{\beta})\| = O_p(\sqrt{K/n})$.

Proof. Let $\hat{g} = \hat{g}(\hat{\beta})$ and choose $\delta_n = o(n^{-1/\gamma}\zeta(K)^{-1})$, $\sqrt{K/n} = o(\delta_n)$, and Λ_n as in the statement of Lemma A.10. Let $\bar{\lambda} = -\delta_n \hat{g}/\|\hat{g}\|$, so that $\bar{\lambda}'\hat{g} = -\delta_n\|\hat{g}\|$. Then $\bar{\lambda} \in \Lambda_n$, so by an expansion and Lemma A.12,

$$\hat{S}(\hat{\beta}, \bar{\lambda}) = s_0 - \bar{\lambda}'\hat{g} + \bar{\lambda}'\left[\sum_i s_2(\dot{\lambda}'\hat{g}_i)\hat{g}_i\hat{g}_i'/n\right]\bar{\lambda}/2$$

$$\geqslant s_0 - \bar{\lambda}'\hat{g} - C\|\bar{\lambda}\|^2 \geqslant s_0 + \delta_n\|\hat{g}\| - C\delta_n^2. \tag{A.8}$$

Subtracting s_0 from both sides of this equation and applying Lemma A.13 gives $\delta_n\|\hat{g}\| - C\delta_n^2 \leqslant O_p(K/n)$. Also, by the choice of δ_n, $K/(n\delta_n) = \sqrt{K/n}(\sqrt{K/n}/\delta_n) = o(\sqrt{K/n}) = o(\delta_n)$. Subtracting and dividing by δ_n gives

$$\|\hat{g}\| \leqslant O_p(K/(n\delta_n)) + C\delta_n = O_p(\delta_n).$$

Now, for any $\varepsilon_n \to 0$ consider $\tilde{\lambda} = -\varepsilon_n\hat{g}$. Then $\|\tilde{\lambda}\| = o_p(\delta_n)$, so that $\tilde{\lambda} \in \Lambda_n$, and hence $\tilde{\lambda} \in \hat{\Lambda}(\hat{\beta})$ w.p.a.1. It follows by the second equality above that for n large enough,

$$\hat{S}(\hat{\beta}, \tilde{\lambda}) \geqslant s_0 - \tilde{\lambda}'\hat{g} - C\|\tilde{\lambda}\|^2 = s_0 + \|\hat{g}\|^2(\varepsilon_n - C\varepsilon_n^2) > s_0 + \|\hat{g}\|^2\varepsilon_n/2,$$

where the last inequality follows by $\varepsilon_n \to 0$, and hence $(\varepsilon_n - C\varepsilon_n^2) > \varepsilon_n/2$ for n large enough. Then by Lemma A.13, and subtracting s_0 from both sides we have $\|\hat{g}\|^2\varepsilon_n = O_p(K/n)$. Since ε_n is any sequence converging to zero, it follows that $\|\hat{g}\|^2 = O_p(K/n)$. □

Proof of Theorem 5.5. Define $\tilde{W} = I_J \otimes \hat{A}^-$. As in the proof of Theorem 5.3 we have $\lambda_{\max}(\tilde{W}) \leqslant C$ w.p.a.1. Then by Lemma A.14,

$$\hat{g}'\tilde{W}\hat{g} \leqslant C\|\hat{g}\|^2 = O_p(K/n) \xrightarrow{p} 0.$$

Now, note that $\hat{g}'\tilde{W}\hat{g} = \hat{R}(\beta)$ as in Eq. (A.1), for $\tilde{\Sigma} = I_J$. Then, by Lemma A.5 all of the conditions of Lemma A.1 are satisfied, and its conclusion gives the result. □

Lemma A.15. *If Assumptions* 1–4 *and* 6 *are satisfied then* $\hat{\beta} = \beta_0 + O_p(\sqrt{K/n})$.

Proof. By Theorem 5.5 $\hat{\beta} \xrightarrow{p} \beta_0$. By an expansion, $\hat{g} = \bar{g} + \tilde{G}(\hat{\beta} - \beta_0)$ for $\tilde{G} = \partial\hat{g}(\bar{\beta})/\partial\beta$ and $\bar{\beta}$ the mean-value. Let $\tilde{W} = I_J \otimes \hat{A}^{-1}$ and $\hat{R}(\beta) = \hat{g}(\beta)'\tilde{W}\hat{g}(\beta)$.

$$\hat{R}(\hat{\beta}) = \hat{R}(\beta_0) + 2\bar{g}'\tilde{W}\tilde{G}(\hat{\beta} - \beta_0) + (\hat{\beta} - \beta_0)'\tilde{G}'\tilde{W}\tilde{G}(\hat{\beta} - \beta_0).$$

By T and CS, for $\hat{D} = [(\hat{\beta} - \beta_0)'\tilde{G}'W\tilde{G}(\hat{\beta} - \beta_0)]^{1/2}$ and $\hat{F} = [\hat{R}(\hat{\beta}) + \hat{R}(\beta_0)]^{1/2}$,

$$\hat{D}^2 \leqslant \hat{F}^2 + \hat{R}(\beta_0)^{1/2}\hat{D} \leqslant \hat{F}^2 + \hat{F}\hat{D} \leqslant \hat{F}^2 + 2\hat{F}\hat{D}.$$

Subtracting $2\hat{F}\hat{D}$ from and adding $\hat{F}^2$ to both sides, and then taking square roots gives $|\hat{D} - \hat{F}| \leqslant \sqrt{2}\hat{F}$. By T, $|\hat{D} - \hat{F}| \geqslant \hat{D} - \hat{F}$, so that $\hat{D} \leqslant (\sqrt{2} + 1)\hat{F} = C\hat{F}$. By Lemma A.6,

$\lambda_{\max}(\tilde{W}) \leqslant C$, so $\tilde{W} \leqslant CI_{\mathrm{JK}}$ w.p.a.1. Then by Lemmas A.9 and A.14 and T, $\hat{F}^2 = \mathrm{O_p}(K/n)$. Also, by Lemma A.3, $\tilde{G}'W\tilde{G} \xrightarrow{p} \mathrm{E}[D(x)'D(x)]$ which is nonsingular, so that $\hat{D}^2 \geqslant C\|\hat{\beta}-\beta_0\|^2$ w.p.a.1. Thus,

$$C\|\hat{\beta}-\beta_0\|^2 \leqslant \hat{D}^2 \leqslant C\hat{F}^2 = \mathrm{O_p}(K/n). \qquad \square$$

Lemma A.16. *If Assumptions* 2, 5 *and* 6 *are satisfied and* $\|\tilde{\lambda}\| = \mathrm{O_p}(\tau_n)$, *then* $\breve{\Omega} = -\sum_i s_{vv}(\tilde{\lambda}'\hat{g}_i)\hat{g}_i\hat{g}_i/n$ *and* $\breve{G} = -\sum_i s_v(\tilde{\lambda}'\hat{g}_i)\partial g_i(\hat{\beta})/\partial\beta/n$ *satisfy*

$$\|\breve{\Omega}-\hat{\Omega}\| = \mathrm{O_p}(\tau_n\zeta(K)\sqrt{K}), \quad \|\breve{G}-\hat{G}\| = \mathrm{O_p}(\tau_n\sqrt{K}).$$

Furthermore, if $\tau_n\zeta(K)\sqrt{K} \to 0$ *then* $1/C \leqslant \lambda_{\min}(\breve{\Omega}) \leqslant \lambda_{\max}(\breve{\Omega}) \leqslant C$ *w.p.a.1.*

Proof. By Lemma A.6, $\lambda_{\max}(\hat{\Omega}) \leqslant C$ w.p.a.1. For $b_i = \sup_{\beta\in\mathcal{N}}\|\rho(z_i,\beta)\|$, by T, CS and M,

$$\|\breve{\Omega}-\hat{\Omega}\| \leqslant C\sum_i|\tilde{\lambda}'\hat{g}_i|b_i^2\|q_i\|^2/n \leqslant \sqrt{\tilde{\lambda}'\hat{\Omega}\tilde{\lambda}}\sqrt{\sum_i b_i^4\|q_i\|^4/n}$$

$$\leqslant \|\tilde{\lambda}\|\mathrm{O_p}(\{\mathrm{E}[\mathrm{E}[b_i^4|x_i]\|q_i\|^4]\}^{1/2}) = \mathrm{O_p}(\tau_n\zeta(K)\sqrt{K}).$$

Also, changing to $b_i = \sup_{\beta\in\mathcal{N}}\|\rho_\beta(z_i,\beta)\|$,

$$\|\breve{G}-\hat{G}\| \leqslant C\sum_i|\tilde{\lambda}'\hat{g}_i|b_i\|q_i\|/n \leqslant \sqrt{\tilde{\lambda}'\hat{\Omega}\tilde{\lambda}}\sqrt{\sum_i b_i^2\|q_i\|^2/n}$$

$$\leqslant \|\tilde{\lambda}\|\mathrm{O_p}(\{\mathrm{E}[\mathrm{E}[b_i^2|x_i]\|q_i\|^2]\}^{1/2}) = \mathrm{O_p}(\tau_n\sqrt{K}).$$

The last conclusion follows from Lemma A.6 and $\|\breve{\Omega}-\hat{\Omega}\| \xrightarrow{p} 0$, as in the proof of Lemma A.6. $\square$

Lemma A.17. *If Assumptions* 2, 5 *and* 6 *are satisfied and* $\|\tilde{\lambda}\| = \mathrm{O_p}(\tau_n)$ *then* $\|\bar{G}'\bar{\Omega}^{-1}(\breve{\Omega}-\hat{\Omega})\| = \mathrm{O_p}(\tau_n\zeta(K))$.

Proof. For $H_i = H_i = \bar{G}'\bar{\Omega}^{-1}(I_J\otimes q_i)$, let $\hat{R}_n = \sum_i \|H_i\|^2\delta_i^4\|q_i\|^2/n$. As in the proof of Lemma A.8 we have $\sum_i \|H_i\|^2/n = \mathrm{O_p}(1)$. Then $\mathrm{E}[\hat{R}_n|X] \leqslant C\zeta(K)^2\sum_i\|H_i\|^2/n = \mathrm{O_p}(\zeta(K)^2)$, so that $\hat{R}_n = \mathrm{O_p}(\zeta(K)^2)$. Then by T and CS,

$$\|\bar{G}'\bar{\Omega}^{-1}(\breve{\Omega}-\hat{\Omega})\| \leqslant \sum_i\|H_i\||\lambda'\hat{g}_i|\delta_i^2\|q_i\|/n$$

$$\leqslant \sqrt{\tilde{\lambda}'\hat{\Omega}\tilde{\lambda}}\left(\sum_i\|H_i\|^2\delta_i^4\|q_i\|^2/n\right)^{1/2} \leqslant \mathrm{O_p}(\tau_n\zeta(K)). \qquad \square$$

Proof of Theorem 5.6. By Lemmas A.14 and A.15 the hypotheses of Lemma A.11 are satisfied with $\tilde{\beta}=\hat{\beta}$ and $\tau_n=\sqrt{K/n}$, so that $\hat{\lambda}$ exists w.p.a.1 and $\|\hat{\lambda}\|=\mathrm{O_p}(\sqrt{K/n})$. By Assumption 6 there is δ_n such that $\delta_n=\mathrm{o}(n^{-1/\gamma}\zeta(K)^{-1})$ and $\sqrt{K/n}=\mathrm{o}(\delta_n)$. Then $\hat{\lambda}\in\Lambda_n=\{\lambda\colon \|\lambda\|\leqslant\delta_n\}$, so by Lemma A.10 $\max_{i\leqslant n}|\hat{\lambda}'\hat{g}_i|\xrightarrow{\mathrm{p}}0$. Also, by consistency of $\hat{\beta}$ it will be an element of *int*(B) w.p.a.1. It follows that w.p.a.1 $\hat{S}(\beta,\lambda)=\sum_i s(\lambda'\hat{g}_i(\beta))/n$ is twice continuously differentiable in a neighborhood of $(\hat{\beta},\hat{\lambda})$. Also, by Lemma A.6 with $\tau_n=\sqrt{K/n}$, w.p.a.1 $\hat{\Omega}=\sum_{i=1}^n\hat{g}_i\hat{g}_i'/n$ is nonsingular, and by continuity of $s_2(v)$, $s_2(\hat{\lambda}'\hat{g}_i)\leqslant -C$ for each i, so that $\partial^2\hat{S}(\hat{\beta},\hat{\lambda})/\partial\lambda\partial\lambda'=\sum_i s_2(\hat{\lambda}'\hat{g}_i)\hat{g}_i\hat{g}_i'/n$ is nonsingular w.p.a.1. Then by the first-order condition $\partial\hat{S}(\hat{\beta},\hat{\lambda})/\partial\lambda=0$ and the implicit function theorem (e.g. Theorem 9.28 of Rudin, 1976), for all β in a neighborhood of $\hat{\beta}$ there is $\hat{\lambda}(\beta)$ such that $\partial\hat{S}(\beta,\hat{\lambda}(\beta))/\partial\lambda=0$ and $\hat{\lambda}(\beta)$ is continuously differentiable in β. By concavity of $\hat{S}(\beta,\lambda)$ we have $\hat{S}(\beta,\hat{\lambda}(\beta))=\max_{\lambda\in\hat{\Lambda}(\beta)}\hat{S}(\beta,\lambda)$. Then the first-order conditions for $\hat{\beta}$ and the envelope theorem give

$$0=\partial\hat{S}(\beta,\hat{\lambda}(\beta))/\partial\beta|_{\beta=\hat{\beta}}=\partial\hat{S}(\hat{\beta},\hat{\lambda})/\partial\beta=\breve{G}'\hat{\lambda}=0,$$

$$\breve{G}=n^{-1}\sum_{i=1}^n s_1(\hat{\lambda}'\hat{g}_i)\partial g_i(\hat{\beta})/\partial\beta.$$

Expanding the first-order condition for $\hat{\lambda}$ around $\lambda=0$ gives

$$0=-\hat{g}-\breve{\Omega}\hat{\lambda}=0,\quad \breve{\Omega}=-\sum_i s_2(\bar{\lambda}'\hat{g}_i)\hat{g}_i\hat{g}_i'/n.$$

By Lemma A.16 with $\tau_n=\sqrt{K/n}$, $\lambda_{\min}(\breve{\Omega})\geqslant C$ w.p.a.1. Solving for $\hat{\lambda}$, plugging in the first-order condition for $\hat{\beta}$, and multiplying by $s_2(0)/s_1(0)$

$$\breve{G}'\breve{\Omega}^{-1}\hat{g}=0.$$

Expanding $\hat{g}$ around β_0 gives, for a mean value $\dot{\beta}$ and $\dot{G}=\partial\hat{g}(\dot{\beta})/\partial\beta$,

$$\breve{G}'\breve{\Omega}^{-1}\dot{G}(\hat{\beta}-\beta_0)+\breve{G}'\breve{\Omega}^{-1}\bar{g}=0.$$

It follows similarly to the proof of Theorem 5.4 that $\breve{G}'\breve{\Omega}^{-1}\dot{G}\xrightarrow{\mathrm{p}}V^{-1}$. Also, it follows from Lemmas A.8, A.17 and T that $\|\bar{G}'\bar{\Omega}^{-1}(\breve{\Omega}-\bar{\Omega})\|=\mathrm{o_p}(\sqrt{K/n})$ and $\breve{G}'\breve{\Omega}^{-1}\bar{g}=\bar{G}'\bar{\Omega}^{-1}\bar{g}+\mathrm{o_p}(n^{-1/2})$. As in the proof of Theorem 5.4 it follows that $\sqrt{n}\bar{G}'\bar{\Omega}^{-1}\bar{g}\xrightarrow{\mathrm{d}}\mathrm{N}(0,V^{-1})$. The remainder of the asymptotic normality result follows by standard arguments.

Next, consider the consistency of the variance estimator. Replacing $s_2(v)$ with $s_1(v)$ in $\breve{\Omega}$ in Lemma A.16 and using also Lemmas A.6 and A.7, with $\tau_n=\sqrt{K/n}$ we have

$$\|\breve{\Omega}-\Omega\|=\mathrm{O_p}(\zeta(K)\sqrt{K^2/n})\xrightarrow{\mathrm{p}}0,$$

$$\|\breve{G}-G\|=\mathrm{O_p}(\sqrt{K^2/n})\xrightarrow{\mathrm{p}}0.$$

It then follows as in the proof of Theorem 5.4 that $\breve{G}'\breve{\Omega}^{-1}\breve{G} \xrightarrow{p} V^{-1}$. Furthermore, by $s_1(0) = -1$ an expansion gives

$$\left|1 + \sum_i s_1(\hat{\lambda}'\hat{g}_i)/n\right|$$

$$= \left|\sum_i s_2(\tilde{\lambda}'\hat{g}_i)\hat{g}_i'\hat{\lambda}/n\right| \leqslant C\left(\|\hat{\lambda}\|\sum_i \|\hat{g}_i\|/n\right)$$

$$\leqslant \mathrm{O_p}(\sqrt{K/n})\zeta(K)\sum_i \sup_{\beta\in\mathscr{B}} \|\rho(z_i,\beta)\|/n = \mathrm{O_p}(\sqrt{K/n}\zeta(K)) \xrightarrow{p} 0.$$

Then by T,

$$\hat{G}'\hat{\Omega}^{-1}\hat{G} - G'\Omega^{-1}G$$

$$= -\breve{G}'\breve{\Omega}^{-1}\breve{G}\left[1 + n\Big/\sum_i s_1(\hat{\lambda}'\hat{g}_i)\right] + \breve{G}'\breve{\Omega}^{-1}\breve{G} - G'\Omega^{-1}G \xrightarrow{p} 0.$$

Then the conclusion follows by $G'\Omega^{-1}G \to V^{-1}$ and T. □

Proof of Lemma 6.1. Let $\bar{g} = \hat{g}(\beta_0)$ and $\hat{g} = \hat{g}(\hat{\beta})$. By a mean value expansion

$$\hat{g} = \bar{g} + \bar{G}(\hat{\beta} - \beta_0), \quad \bar{G} = \partial\hat{g}(\bar{\beta})/\partial\beta. \tag{A.9}$$

Note that $|\lambda_{\min}(\hat{\Omega}) - \lambda_{\min}(\Omega)| \leqslant \|\hat{\Omega} - \Omega\| \xrightarrow{p} 0$, so $\lambda_{\min}(\hat{\Omega}) \geqslant C$ w.p.a.1, and hence,

$$\|\hat{\Omega}^{-1}(\bar{G} - G)\|^2 = tr((\bar{G} - G)'\hat{\Omega}^{-1/2}\hat{\Omega}^{-1}\hat{\Omega}^{-1/2}(\bar{G} - G))$$

$$\leqslant C\,tr((\bar{G} - G)'\hat{\Omega}^{-1}(\bar{G} - G)) \leqslant C\|\bar{G} - G\|^2 \xrightarrow{p} 0.$$

It follows similarly that w.p.a.1, $\|\hat{\Omega}^{-1}(\hat{\Omega} - \Omega)\| \leqslant C\|\hat{\Omega} - \Omega\| \xrightarrow{p} 0$. Note also that by $G'\Omega^{-1}G$ bounded,

$$\|\Omega^{-1}G\|^2 = tr(G'\Omega^{-1/2}\Omega^{-1}\Omega^{-1/2}G) \leqslant tr(G'\Omega^{-1}G) \leqslant C.$$

It then follows by T and CS that

$$\|\hat{\Omega}^{-1}\bar{G} - \Omega^{-1}G\| \leqslant \|\hat{\Omega}^{-1}(\bar{G} - G)\| + \|\hat{\Omega}^{-1}(\hat{\Omega} - \Omega)\Omega^{-1}G\|$$

$$\leqslant C\|\bar{G} - G\| + \|\hat{\Omega} - \Omega\|\|\Omega^{-1}G\| \xrightarrow{p} 0.$$

It also follows by T that $\|\hat{\Omega}^{-1}\bar{G}\| = \mathrm{O_p}(1)$. Also, note that for $g_i = g(z_i, \beta_0)$,

$$\mathrm{E}[\bar{g}'\Omega^{-1}\bar{g}] = \mathrm{E}[g_i'\Omega^{-1}g_i]/n = \mathrm{E}[tr(\Omega^{-1}g_ig_i')]/n = tr(I_m)/n = m/n.$$

Then by M, $\|\Omega^{-1}\bar{g}\|^2 \leqslant C\sqrt{\bar{g}'\Omega^{-1}\bar{g}} = O_p(\sqrt{m/n})$. Therefore, by T and CS

$$\begin{aligned}\|\bar{G}'\hat{\Omega}^{-1}\bar{g} - G'\Omega^{-1}\bar{g}\| &= \|\bar{G}'\hat{\Omega}^{-1}(\hat{\Omega}-\Omega)\Omega^{-1}\bar{g}\| + \|(\bar{G}-G)'\Omega^{-1}\bar{g}\| \\ &\leqslant (\|\bar{G}'\hat{\Omega}^{-1}\|\|\hat{\Omega}-\Omega\| + \|\bar{G}-G\|)\|\Omega^{-1}\bar{g}\| \\ &= [O_p(1)o_p(1) + o_p(1)]O_p(\sqrt{m/n}) = o_p(\sqrt{m/n}).\end{aligned}$$

Also,

$$E[\|G'\Omega^{-1}\bar{g}\|^2] = E[tr(G'\Omega^{-1}\bar{g}\bar{g}'\Omega^{-1}G)] = tr(G'\Omega^{-1}G)/n \leqslant C/n,$$

so by M, $\|G'\Omega^{-1}\bar{g}\| = O_p(1/\sqrt{n}) = o_p(\sqrt{m/n})$. Then by T,

$$\|\bar{G}'\hat{\Omega}^{-1}g\| = o_p(\sqrt{m/n}).$$

Also, it follows by T and CS that,

$$\begin{aligned}\|\bar{G}'\hat{\Omega}^{-1}\bar{G} - G'\Omega^{-1}G\| &\leqslant (\|\bar{G}'\hat{\Omega}^{-1}\| + \|\Omega^{-1}G\|)\|\bar{G}-G\| \\ &\quad + \|\bar{G}'\hat{\Omega}^{-1}\|\|\hat{\Omega}-\Omega\|\|\Omega^{-1}G\|\end{aligned}$$

so that $\bar{G}'\hat{\Omega}^{-1}\bar{G} = O_p(1)$. Therefore, by substituting in Eq. (A.9), we have

$$\begin{aligned}&|n\hat{g}'\hat{\Omega}^{-1}\hat{g} - n\bar{g}'\hat{\Omega}^{-1}\bar{g}|/\sqrt{2(m-p)} \\ &\quad\leqslant [|2(\hat{\beta}-\beta_0)'\bar{G}'\hat{\Omega}^{-1}\bar{g}| + |(\hat{\beta}-\beta_0)'\bar{G}'\hat{\Omega}^{-1}\bar{G}(\hat{\beta}-\beta_0)|]n/\sqrt{2(m-p)} \\ &\quad= [O_p(1/\sqrt{n})o_p(\sqrt{m/n}) + O_p(1/n)]n/\sqrt{2(m-p)} \\ &\quad= o_p(\sqrt{m}/n)n/\sqrt{2(m-p)} \xrightarrow{p} 0.\end{aligned}$$

Furthermore, we have similarly to Eq. (A.10) that

$$\begin{aligned}&|n\bar{g}'\hat{\Omega}^{-1}\bar{g} - n\bar{g}'\Omega^{-1}\bar{g}|/\sqrt{2(m-p)} \\ &\quad= |n\bar{g}'(\hat{\Omega}^{-1} - \Omega^{-1})\bar{g}|/\sqrt{2(m-p)} \\ &\quad\leqslant |n\bar{g}'\Omega^{-1}(\hat{\Omega}-\Omega)\Omega^{-1}\bar{g}| + |n\bar{g}'\Omega^{-1}(\hat{\Omega}-\Omega)\hat{\Omega}^{-1}(\hat{\Omega}-\Omega)\Omega^{-1}\bar{g}|/\sqrt{2(m-p)} \\ &\quad\leqslant n\|\Omega^{-1}\bar{g}\|^2(\|\hat{\Omega}-\Omega\| + C\|\hat{\Omega}-\Omega\|^2)/\sqrt{2(m-p)} \\ &\quad= nO_p(m/n)o_p(1/\sqrt{m})/\sqrt{2(m-p)} \xrightarrow{p} 0.\end{aligned}$$

The conclusion then follows by T. □

Proof of Lemma 6.2. Let $T = [n\bar{g}'\Omega^{-1}\bar{g} - (m-a)]/\sqrt{2(m-a)}$, and note that

$$T = T_1T_2 + T_3 + T_4, \quad T_1 = \sqrt{\frac{m}{m-a}}, \quad T_2 = \sum_{i,j:i<j} \sqrt{\frac{2}{n^2m}} g_i'\Omega^{-1}g_j,$$

$$T_3 = \left(\sum_i g_i'\Omega^{-1}g_i/n - m\right) \Bigg/ \sqrt{2(m-a)}, \quad T_4 = a/\sqrt{2(m-a)}.$$

Note that $T_1 \to 1$ and $T_4 \to 0$. Also, $\mathrm{E}[T_3] = 0$ and $Var(T_3) \leqslant \mathrm{E}[(g_i'\Omega^{-1}g_i)^2]/(2n(m-a)) \to 0$, so by M, $T_3 \xrightarrow{p} 0$. Therefore, by the Slutzky Theorem it suffices to show that $T_2 \xrightarrow{d} \mathrm{N}(0,1)$. To do this we show that the hypotheses of Lemma 2 of Hall (1984), as cited by de Jong and Bierens (1994), are satisfied. Let $H_n(u,v) = \sqrt{2/(n^2m)}g(u,\beta_0)'\Omega^{-1} g(v,\beta_0)$ and

$$G_n(u,v) = \mathrm{E}[H_n(z_1,v)H_n(z_1,u)]$$
$$= \frac{2}{n^2m} g(u,\beta_0)'\Omega^{-1}\mathrm{E}[g_1g_1']\Omega^{-1}g(v,\beta_0) = \sqrt{\frac{2}{n^2m}} H_n(u,v).$$

Note that $\mathrm{E}[H_n(z_1,z_2)|z_1] = \sqrt{2/(n^2m)}g(z_1,\beta_0)'\Omega^{-1}\mathrm{E}[g_2] = 0$ and that

$$\mathrm{E}[H_n(z_1,z_2)^2] = 2\mathrm{E}[g_1'\Omega^{-1}g_2g_2'\Omega^{-1}g_1]/(n^2m)$$
$$= 2\mathrm{E}[g_1'\Omega^{-1}g_1]/(n^2m) = 2/n^2.$$

It follows by the Cauchy–Schwartz inequality that

$$n^{-1}\mathrm{E}[H_n(z_1,z_2)^4]/\{\mathrm{E}[H_n(z_1,z_2)^2]\}^2 = 4n^{-5}m^{-2}\mathrm{E}[(g_1'\Omega^{-1}g_2)^4]/(4/n^4)$$
$$\leqslant n^{-1}m^{-2}\mathrm{E}[(g_1'\Omega^{-1}g_1)^2(g_2'\Omega^{-1}g_2)^2]$$
$$\leqslant \{\mathrm{E}[(g_1'\Omega^{-1}g_1)^2]/(m\sqrt{n})\}^2 \to 0.$$

Also,

$$\mathrm{E}[G_n(z_1,z_2)^2]/\{\mathrm{E}[H_n(z_1,z_2)^2]\}^2 = [2/(n^2m)]/\mathrm{E}[H_n(z_1,z_2)^2]$$
$$= 1/m \to 0.$$

Therefore, the conclusion follows from the conclusion of Lemma 2 of Hall (1984). □

Proof of Theorem 6.3. To prove the first conclusion, let $\hat{\Omega}$ in Lemma 6.1 equal $\hat{\Omega}(\tilde{\beta})$. By Lemma A.6 and $\zeta(K)^2K^2/n \to 0$, we have $\lambda_{\min}(\Omega) \geqslant C$ and

$$\|\hat{\Omega} - \Omega\| = \mathrm{O_p}([K^{3/2}/\sqrt{n} + \zeta(K)K/\sqrt{n}]/K^{1/2}) = \mathrm{O_p}(1/\sqrt{m}), \tag{A.10}$$

for $m = JK$. Also, it follows by Lemma A.7 that $\|\partial\hat{g}(\bar{\beta})/\partial\beta - G\| \xrightarrow{p} 0$ for any $\bar{\beta} = \beta_0 + \mathrm{O_p}(1/\sqrt{n})$. It also follows as in the proof of Theorem 5.4 that $G'\Omega^{-1}G$ is bounded. Thus, all of the hypotheses of Lemma 6.1 are satisfied. Furthermore, note that

$$\mathrm{E}[\{g_i'\Omega^{-1}g_i\}^2] \leqslant C\mathrm{E}[\|g_i\|^4] \leqslant C\mathrm{E}[\|\rho(z_i,\beta_0)\|^4\|q_i\|^4] \leqslant C\mathrm{E}[\|q_i\|^4] \leqslant C\zeta(K)^2K.$$

Therefore, it follows that the hypotheses of Lemma 6.2 are satisfied, so that the first conclusion follows by Lemmas 6.1 and 6.2. To show the second conclusion, let $\hat{\Omega} = \tilde{\Sigma} \otimes \hat{A}$. It follows similarly to Lemma A.6 that under homoskedasticity Eq. (A.10) is also satisfied here. The rest of the hypotheses of Lemma 6.1 then hold as before, so the second conclusion follows by the conclusion of Lemma 6.1. □

Proof of Theorem 6.4. Note that for $\hat{g} = \hat{g}(\hat{\beta})$ and $\bar{g} = \hat{g}(\beta_0)$,

$$\|\hat{g} - \bar{g}\| \leqslant \sum_i \|\rho(z_i, \hat{\beta}) - \rho(z_i, \beta_0)\| \|q_i\|/n$$

$$\leqslant \left(\sum_i \delta_i^2/n\right)^{1/2} \left(\sum_i \|q_i\|^2/n\right)^{1/2} \|\hat{\beta} - \beta_0\| = \mathrm{O_p}(\sqrt{K/n}).$$

It then follows by the triangle inequality and Lemma A.6 that $\|\hat{g}\| = \mathrm{O_p}(\sqrt{K/n})$. Then by Lemma A.14 we have $\|\hat{\lambda}\| = \mathrm{O_p}(\sqrt{K/n})$. It then follows as in the proof of Theorem 5.6 that $\hat{\lambda} \in \hat{\Lambda}_n(\hat{\beta})$ w.p.a.1. Therefore, the first-order conditions for $\hat{\lambda}$ are satisfied w.p.a.1. Expanding around $\lambda = 0$ then gives

$$0 = \partial \hat{S}(\hat{\beta}, \hat{\lambda})/\partial \lambda = \sum_{i=1}^n s_1(\hat{\lambda}' \hat{g}_i)\hat{g}_i/n = s_1(0)\hat{g} + s_2(0)\breve{\Omega}\hat{\lambda},$$

$$\breve{\Omega} = \sum_{i=1}^n [s_2(\bar{\lambda}' \hat{g}_i)/s_2(0)]\hat{g}_i\hat{g}_i'/n.$$

It follows as in Lemma A.6 that $\lambda_{\min}(\breve{\Omega}) \geqslant C$ w.p.a.1, so inverting gives

$$\hat{\lambda} = -[s_1(0)/s_2(0)]\breve{\Omega}^{-1}\hat{g}.$$

Furthermore, expanding $\hat{S}(\hat{\beta}, \hat{\lambda})$ around $\lambda = 0$ and plugging in gives

$$\hat{T}_{\mathrm{GEL}} = -2n[s_2(0)/s_1(0)^2]\{s_1(0)\hat{g}'\hat{\lambda} + s_2(0)\hat{\lambda}'\ddot{\Omega}\hat{\lambda}/2\}$$

$$= n\hat{g}'(2\breve{\Omega}^{-1} - \breve{\Omega}^{-1}\ddot{\Omega}\breve{\Omega}^{-1})\hat{g},$$

where $\ddot{\Omega} = \sum_{i=1}^n [s_2(\ddot{\lambda}' \hat{g}_i)/s_2(0)]\hat{g}_i\hat{g}_i'/n$ and $\ddot{\lambda}$ lies on the line joining $\hat{\lambda}$ and 0.

Now, note that $(2\breve{\Omega}^{-1} - \breve{\Omega}^{-1}\ddot{\Omega}\breve{\Omega}^{-1})^{-1} = \breve{\Omega}(2\breve{\Omega} - \ddot{\Omega})^{-1}\breve{\Omega}$, so all of the conditions of Lemma 6.1 will be satisfied if it can be shown that $\|\breve{\Omega}(2\breve{\Omega} - \ddot{\Omega})^{-1}\breve{\Omega} - \Omega\| = \mathrm{o_p}(1/\sqrt{K})$. By Lemmas A.6 with $\tau_n = 1/\sqrt{n}$ and (A.16) with $\tau_n = \sqrt{K/n}$ we have

$$\|\breve{\Omega} - \Omega\| = \mathrm{O_p}(\zeta(K)\sqrt{K^2/n}/\sqrt{K}) = \mathrm{o_p}(1/\sqrt{K}), \quad \|\ddot{\Omega} - \Omega\| = \mathrm{o_p}(1/\sqrt{K}).$$

It follows that for $\hat{B} = 2\breve{\Omega} - \ddot{\Omega}$ we have $\|\hat{B} - \Omega\| \xrightarrow{\mathrm{p}} 0$, so that $\lambda_{\max}(\hat{B}^{-1}) \leqslant C$ w.p.a.1. Then we have

$$\|\breve{\Omega}\hat{B}^{-1}\breve{\Omega} - \Omega\hat{B}^{-1}\Omega\| \leqslant \|(\breve{\Omega} - \Omega)\hat{B}^{-1}(\breve{\Omega} - \Omega)\| + 2\|\Omega\hat{B}^{-1}(\breve{\Omega} - \Omega)\|$$

$$\leqslant C(\|\breve{\Omega} - \Omega\|^2 + \|\breve{\Omega} - \Omega\|) = \mathrm{o_p}(1/\sqrt{K}).$$

We also have, by $\lambda_{\max}(\Omega) \leqslant C$,

$$\|\Omega\hat{B}^{-1}\Omega - \Omega\| = \|\Omega\hat{B}^{-1}(\Omega - \hat{B})\| \leqslant \|\Omega\hat{B}^{-1}(\Omega - \hat{B})\| = \mathrm{o_p}(1/\sqrt{K}). \qquad \square$$

Proof of Lemma 6.5. Note that for $\hat{g} = \hat{g}(\hat{\beta})$,

$$\left(\frac{\sqrt{m}}{n}\right) \frac{n\hat{g}'\hat{\Omega}^{-1}\hat{g} - (m-a)}{\sqrt{2(m-a)}} = T_1\hat{g}'\hat{\Omega}^{-1}\hat{g} + T_2, \quad T_1 = \sqrt{\frac{m}{2(m-a)}} \to \frac{1}{\sqrt{2}},$$

$$T_2 = T_1\left(-\frac{m-a}{n}\right) \to 0.$$

Therefore it suffices to show that $\hat{g}'\hat{\Omega}^{-1}\hat{g} \xrightarrow{\mathrm{p}} \Delta$. Let $g_{ai} = g(z_i, \beta_a)$, $\bar{g} = \sum_i g_{ai}/n$, and $g_a = \mathrm{E}[g_{ia}]$. Note that $Var(g_i) \leqslant \Omega$, so that

$$\mathrm{E}[(\bar{g} - g_a)'\Omega^{-1}(\bar{g} - g_a)] \leqslant \mathrm{E}[(\bar{g} - g_a)'Var(g_{ai})^{-1}(\bar{g} - g_a)] \leqslant Cm/n \to 0.$$

Then by T and CS,

$$\begin{aligned}|\bar{g}'\Omega_a^{-1}\bar{g} - g_a'\Omega_a^{-1}g_a| &\leqslant |(\bar{g} - g_a)'\Omega_a^{-1}(\bar{g} - g_a)| + 2|g_a'\Omega_a^{-1}(\bar{g} - g_a)| \\ &\leqslant \mathrm{o_p}(1) + 2\sqrt{g_a'\Omega_a^{-1}g_a}\sqrt{(\bar{g} - g_a)'\Omega_a^{-1}(\bar{g} - g_a)}.\end{aligned}$$

Thus, by $g_a'\Omega_a^{-1}g_a \to \Delta$, we have $\bar{g}'\Omega_a^{-1}\bar{g} = \mathrm{O_p}(1)$. Also, $\hat{g} = \bar{g} + \tilde{G}(\hat{\beta} - \beta_a)$. Note that $(\hat{\beta} - \beta_a)'G_a'\Omega_a^{-1}G_a(\hat{\beta} - \beta_a) \xrightarrow{\mathrm{p}} 0$ and

$$\begin{aligned}&|(\hat{g} - \bar{g})'\Omega_a^{-1}(\hat{g} - \bar{g}) - (\hat{\beta} - \beta_a)'G_a'\Omega_a^{-1}G_a(\hat{\beta} - \beta_a)| \\ &\quad \leqslant |(\hat{\beta} - \beta_a)'(\tilde{G} - G_a)'\Omega_a^{-1}(\tilde{G} - G_a)(\hat{\beta} - \beta_a)| \\ &\qquad + 2|(\hat{\beta} - \beta_a)'G_a'\Omega_a^{-1}(\tilde{G} - G)(\hat{\beta} - \beta_a)| \\ &\quad \leqslant \|\tilde{G} - G_a\|^2\|\hat{\beta} - \beta_a\|^2 \\ &\qquad + 2\sqrt{(\hat{\beta} - \beta_a)'G_a'\Omega_a^{-1}G_a(\hat{\beta} - \beta_a)}\|\tilde{G} - G_a\|\|\hat{\beta} - \beta_a\| \xrightarrow{\mathrm{p}} 0.\end{aligned}$$

Therefore, $(\hat{g} - \bar{g})'\Omega_a^{-1}(\hat{g} - \bar{g}) = \mathrm{O_p}(1)$. It now follows that

$$\begin{aligned}|\hat{g}'\Omega_a^{-1}\hat{g} - \bar{g}'\Omega_a^{-1}\bar{g}| &\leqslant |(\hat{g} - \bar{g})'\Omega_a^{-1}(\hat{g} - \bar{g})| + 2|\bar{g}'\Omega_a^{-1}(\hat{g} - \bar{g})| \\ &\leqslant \mathrm{o_p}(1) + 2\sqrt{\bar{g}'\Omega_a^{-1}\bar{g}}\sqrt{(\hat{g} - \bar{g})'\Omega_a^{-1}(\hat{g} - \bar{g})} \xrightarrow{\mathrm{p}} 0,\end{aligned}$$

so that $\hat{g}'\Omega_a^{-1}\hat{g} = \mathrm{O_p}(1)$. We then have

$$\begin{aligned}|\hat{g}'(\hat{\Omega}^{-1} - \Omega_a^{-1})\hat{g}| &\leqslant |\hat{g}'\Omega_a^{-1}(\hat{\Omega} - \Omega_a)\Omega_a^{-1}\hat{g}| + |\hat{g}'\Omega_a^{-1}(\hat{\Omega} - \Omega_a)\hat{\Omega}^{-1}(\hat{\Omega} - \Omega)\Omega^{-1}\hat{g}| \\ &\leqslant \|\Omega_a^{-1}\hat{g}\|^2(\|\hat{\Omega} - \Omega_a\| + C\|\hat{\Omega} - \Omega_a\|^2).\end{aligned}$$

It then follows by the triangle inequality that $\hat{g}'\hat{\Omega}^{-1}\hat{g} \xrightarrow{\mathrm{p}} \Delta$. $\square$

References

Amemiya, T., 1974. The nonlinear two-stage least squares estimator. Journal of Econometrics 2, 105–110.

Amemiya, T., 1977. The maximum likelihood and the nonlinear three-stage least squares estimator in the general nonlinear simultaneous equations model. Econometrica 45, 955–968.

Andrews, D.W.K., 1991. Asymptotic normality of series estimators for nonparametric and semiparametric models. Econometrica 59, 307–345.

Arellano, M., Bond, S., 1991. Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. Review of Economic Studies 58, 277–297.

Bierens, H.J., 1982. Consistent model specification tests. Journal of Econometrics 20, 105–134.

Brown, B.W., Newey, W.K., 1998. Efficient estimation of asymptotic variances. Working paper, Rice University.

Burman, P., Chen, K.W., 1989. Nonparametric estimation of a regression function. Annals of Statistics 17, 1567–1596.

Chamberlain, G., 1987. Asymptotic efficiency in estimation with conditional moment restrictions. Journal of Econometrics 34, 305–334.

de Jong, R.M., Bierens, H., 1994. On the limit behavior of a chi-square type test if the number of conditional moments tested approaches infinity. Econometric Theory 9, 70–90.

Donald, S.G., Newey, W.K., 2001. Choosing the number of instruments. Econometrica 69, 1161–1191.

Donald, S.G., Imbens, G.W., Newey, W.K., 2002. Choosing the number of moments in conditional moment restriction models, preprint.

Eubank, R.L., Spiegelman, S.P., 1990. Testing the goodness of fit of a linear model via nonparametric regression techniques. Journal of the American Statistical Association 87, 387–392.

Gallant, A.R., Souza, G., 1991. On the asymptotic normality of Fourier flexible form estimates. Journal of Econometrics 50, 329–353.

Hahn, J., 1997. Efficient estimation of panel data models with sequential moment restrictions. Journal of Econometrics 79, 1–21.

Hall, P., 1984. Central limit theorem for integrated squared error of multivariate nonparametric density estimators. Journal of Multivariate Analysis 14, 1–16.

Hansen, L.P., 1982. Large sample properties of generalized method of moments estimators. Econometrica 50, 1029–1054.

Hansen, L.P., Heaton, J., Yaron, A., 1996. Finite-sample properties of some alternative GMM estimators. Journal of Business and Economic Statistics 14, 262–280.

Hong, Y.-M., White, H., 1995. Consistent specification testing via nonparametric series regression. Econometrica 63, 1133–1160.

Imbens, G., 1997. One step estimators for over-identified generalized method of moments models. Review of Economic Studies 64, 359–383.

Imbens, G.W., Spady, R.H., Johnson, P., 1998. Information theoretic approaches to inference in moment condition models. Econometrica 66, 333–357.

Kitamura, Y., Stutzer, M., 1997. An information-theoretic alternative to generalized method of moments estimation. Econometrica 65, 861–874.

Kitamura, Y., Tripathi, G., Ahn, H., 2001. Empirical likelihood based inference in conditional moment restriction models. Working paper, Department of Economics, University of Wisconsin.

Koenker, R., Machado, J.A.F., 1999. GMM inference when the number of moment conditions is large. Journal of Econometrics 93, 327–344.

LeBlanc, M., Crowley, J., 1995. Semiparametric regression functionals. Journal of the American Statistical Association 90, 95–105.

Newey, W.K., 1985. Maximum likelihood specification testing and conditional moment tests. Econometrica 53, 1047–1070.

Newey, W.K., 1988a. Two-step series estimators of sample selection models. Working paper, MIT Press, Cambridge, MA.

Newey, W.K., 1988b. Adaptive estimation of regression models via moment restrictions. Journal of Econometrics 38, 301–339.

Newey, W.K., 1990. Efficient instrumental variables estimation of nonlinear models. Econometrica 58, 809–837.

Newey, W.K., 1991. Uniform convergence in probability and stochastic equicontinuity. Econometrica 59, 1161–1167.

Newey, W.K., 1993. Efficient estimation of models with conditional moment restrictions. In: Maddala, G.S., Rao, C.R., Vinod, H.D. (Eds.), Handbook of Statistics, Vol. 11: Econometrics, North-Holland, Amsterdam.

Newey, W.K., 1997. Convergence rates and asymptotic normality for series estimators. Journal of Econometrics 79, 147–168.

Newey, W.K., Chipty, T., 2000. Efficient estimation of semiparametric models via moment restrictions. Working paper, MIT Press, Cambridge, MA.

Newey, W.K., McFadden, D., 1994. Large Sample Estimation and Hypothesis Testing. In: Engle, McFadden, D. (Eds.), Handbook of Econometrics, Vol. 4. North-Holland, Amsterdam, pp. 2113–2245.

Newey, W.K., Smith, R.J., 2002. Higher order properties of GMM and generalized empirical likelihood estimators. Working paper, MIT Press, Cambridge, MA.

Newey, W.K., Powell, J.L., 1989. Nonparametric instrumental variables estimation. Econometrica, forthcoming.

Owen, A., 1988. Empirical likelihood ratio confidence regions for a single functional. Biometrika 75, 237–249.

Qin, J., Lawless, J., 1994. Empirical likelihood and general estimating equations. Annals of Statistics 22, 300–325.

Rudin, W., 1976. Principles of Mathematical Analysis. McGraw-Hill, New York.

Schumaker, L., 1981. Spline Functions: Basic Theory. Wiley, New York.

Smith, R.J., 1997. Alternative semi-parametric likelihood approaches to generalized method of moments estimation. Economic Journal 107, 503–519.

Stone, C., 1985. Additive regression and other nonparametric models. Annals of Statistics 13, 689–705.