Econometric Theory, **8**, 1992, 203–222. Printed in the United States of America.

SEMIPARAMETRIC GENERALIZED LEAST SQUARES IN THE MULTIVARIATE NONLINEAR REGRESSION MODEL

MIGUEL A. DELGADO
Indiana University

Asymptotically efficient estimates for the multiple equations nonlinear regression model are obtained in the presence of heteroskedasticity of unknown form. The proposed estimator is a generalized least squares based on nonparametric nearest neighbor estimates of the conditional variance matrices. Some Monte Carlo experiments are reported.

1. INTRODUCTION

Carroll [4] and Robinson [12] have shown that asymptotic efficiency, in the linear regression model under heteroskedasticity of unknown form, is achieved to the first order by generalized least-squares (GLS) estimates based on smooth nonparametric estimates of the conditional variances. Carroll [4] used nonparametric kernel estimates in the simple linear regression model. Robinson [12] employed nearest neighbor (NN) estimates, extending Carroll's result to the multiple linear regression model, and relaxing Carroll's assumptions to moment conditions. In this paper, we consider the multiple equations nonlinear regression model. We show that the GLS based on NN estimates of the conditional variance matrices is asymptotically efficient.

In the next section we describe the model, Section 3 presents the estimation procedure, Section 4 the asymptotic theory, Section 5 contains some Monte Carlo experiments, and Section 6 the conclusions and final remarks. The proofs are in an appendix.

2. THE MODEL

Let (X, Y) be a pair of random variables such that X is $\mathbb{R}^r$-valued and Y is $\mathbb{R}^m$ valued. The conditional distribution $P^Y(\cdot \mid X)$ of Y given X is non-

This research is derived from my doctoral dissertation at the London School of Economics. Financial support from the Economic and Social Research Council (ESRC) reference number: R00023411 is gratefully acknowledged.

I am grateful to Peter Robinson and the editor and two anonymous referees of this journal for their valuable comments. I have also benefited by discussions with Anil Bera, Javier Hidalgo, Lola Prada, Getulio Silveira, and Pravin Trivedi.

 0266-4666/92 $5.00 + .00

degenerate for all X at which it is defined and has conditional mean of known form

$$E(Y|X = x) = f(\theta^0, x) \text{ a.s.}, \tag{1}$$

where $f: \mathbb{R}^p \times \mathbb{R}^r \to \mathbb{R}^m$ and θ^0 is an unknown $p \times 1$ vector of parameters. The higher conditional moments of Y are of unknown functional form and, therefore, $\text{var}(Y|X = x) = \Omega(x)$ is a matrix of unknown functions of α.

Given a random sample $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ from the distribution of (X, Y), the GLS estimate of θ^0 associated with a given sequence of matrices $\{\Sigma_i, i = 1, \ldots, n\}$ is the vector $\theta_n(\Sigma)$ which minimizes, in θ, the quadratic form

$$Q_n(\theta, \Sigma) = n^{-1} \sum_{i=1}^{n} (Y_i - f(\theta, X_i))^T \Sigma_i^{-1} (Y_i - f(\theta, X_i)). \tag{2}$$

Then $\theta_n(I_m)$, which uses $\Sigma_i = I_m$ in (2) (I_m is the $m \times m$ identity matrix) is the multivariate least-squares estimate; $\theta_n(S_n)$, which uses $\Sigma_i = S_n$ in (2), where

$$S_n = n^{-1} \sum_{i=1}^{n} (Y_i - f(\theta_n(I_m), X_i))(Y_i - f(\theta_n(I_m), X_i))^T \tag{3}$$

is the minimum distance estimator (see [9]); and $\theta_n(\Omega)$, which uses $\Sigma_i = \Omega(X_i)$ in (2), is the unfeasible GLS.

Under mild regularity conditions (see [5]), the semiparametric efficiency bound of any estimate of θ^0, when (1) is the only available information on $P^Y(\cdot|X)$, is

$$V_0^{-1} = \{E(\mathring{f}(\theta^0, X)\Omega(X)^{-1}\mathring{f}(\theta^0, X)^T)\}^{-1}, \tag{4}$$

where $\mathring{f}(\theta, \cdot) = \partial f(\theta, \cdot)/\partial \theta$. The asymptotic bound (4) is achieved by $n^{1/2}(\theta_n(\Omega) - \theta^0)$. When $\Omega(X)$ is independent of X, the feasible $\theta_n(S_n)$ is asymptotically efficient (see [9]). Under conditional heteroskedasticity, the asymptotic covariance matrix of $n^{1/2}(\theta_n(S_n) - \theta^0)$ is

$$M_0^{-1} E(\mathring{f}(\theta^0, X) E(\Omega(X))^{-1} \Omega(X) E(\Omega(X))^{-1} \mathring{f}(\theta^0, X)^T) M_0^{-1}, \tag{5}$$

where $M_0 = E(\mathring{f}(\theta^0, X) E(\Omega(X))^{-1} \mathring{f}(\theta^0, X)^T$. The asymptotic covariance matrix (5) may exceed that of $n^{1/2}(\theta_n(I_m) - \theta^0)$, given by

$$\{E(\mathring{f}(\theta^0, X)\mathring{f}(\theta^0, X)^T)\}^{-1} E(\mathring{f}(\theta^0, X)\Omega(X)\mathring{f}(\theta^0, X)^T) \times \{E(\mathring{f}(\theta^0, X)\mathring{f}(\theta^0, X)^T)\}^{-1}. \tag{6}$$

The matrices (5) and (6) are equal under some particular circumstances, for example, when the conditional means have the same functional form for all equations or when the cross-equation covariances are equal to zero (see [7] and [18]).

The goal of this paper is to obtain estimates of θ^0 asymptotically equivalent, to the first order, to $\theta_n(\Omega)$, and consistent estimates of V_0.

3. THE ESTIMATION PROCEDURE

The unknown conditional variance matrices $\Omega(X_j)$ are estimated by

$$\hat{\Omega}_j = \sum_{i=1}^{n} (Y_i - f(\tilde{\theta}_n, X_i))(Y_i - f(\tilde{\theta}_n, X_i))^T W_i(X_j), \tag{7}$$

where $\tilde{\theta}_n$ is some preliminary root-n-consistent estimate (e.g., $\theta_n(I_m)$) and $\{W_i(X_j),\ i = 1,\ldots,n\}$ are NN weights.

We use a modification of Stone's [16] weights proposed by Robinson [12]. These weights are technically convenient in semiparametric estimation (see [10] and [12]). The reader is referred to [12] and [16] for a motivation of these weights.

Let $\Omega(X_i)$ be known to depend on $d \le r$ elements of $X_i, X_{1i},\ldots,X_{di}$. (None of these elements is a constant.) The weights depend on the distances from X_i to $X_1,\ldots,X_n$ according to the Euclidean metric. Before applying the Euclidean metric, the regressors are scaled by their sample standard deviation. That is, we consider the distance function

$$\rho_n(X_i, X_j) = \left\{ \sum_{m=1}^{d} \frac{(X_{mi} - X_{mj})^2}{s_m^2} \right\}^{1/2},$$

$$s_m^2 = (n-1)^{-1} \sum_{i=1}^{n} \left(X_{mi} - \left(n^{-1} \sum_{i=1}^{n} X_{mi} \right) \right)^2, \qquad m = 1,\ldots,d. \tag{8}$$

Let us define

$$\nu_i(X_j) = 1(i \ne j) + \#(\ell : \ell \ne i \ne j, \rho_n(X_\ell, X_j) < \rho_n(X_i, X_j)),$$

$$\lambda_i(X_j) = 1 + \#(\ell : \ell \ne i \ne j, \rho_n(X_\ell, X_j) = \rho_n(X_i, X_j))$$

where $1(\cdot)$ is the usual indicator function. Then

$$W_i(X_j) = \frac{c_{n\nu_i(X_j)} + \cdots + c_{n\nu_i(X_j)+\lambda_i(X_j)-1}}{\lambda_i(X_j)}, \tag{9}$$

where $c_{n0} = 0$ and $c_n = (c_{n1},\ldots,c_{nn})$ is a vector of probability weights (i.e., $c_{ni} \ge 0$ all $i \ge 1$ and $\sum_{i=1}^{n} c_{ni} = 1$). Thus, $(W_{n1}(x),\ldots,W_{nn}(x))$ is also a vector of probability weights. We assume that there exists a sequence of numbers $k_n = k$ such that

W1. $c_{ni} = 0$ for all $i > k$ and $\overline{\lim}_{n\to\infty} k \max_i c_{ni} < \infty$.
W2. $kn^{-1} \to 0$ and $n^{-1/\upsilon} k \to \infty$ as $n \to \infty$ for $\upsilon > 1$.

Examples of k-NN weights satisfying W1 are given in [16]. The simplest weight is $c_{ni} = k^{-1}$ when $i \le k$ and $c_{ni} = 0$ when $i > k$ (uniform weight).

Note that in the absence of ties $W_i(X_j) = 1(0 < \nu_i(X_j) \le k)/k$. The rate of convergence in W2 is related to conditions, in next section, on the tails of the distributions of regressors and disturbances.

Failure to identify the regressors on which $\Omega(\cdot)$ depends leads to inconsistent weights and therefore inefficient GLS estimates. In order to be on the safe side, one can construct the weights using all the regressors.

We estimate θ^0 by $\theta_n(\hat{\Omega})$, which uses $\Sigma_i = \hat{\Omega}_i$ in (2). Under the conditions stated in next section, $\theta_n(\hat{\Omega})$ achieves the semiparametric efficiency bound, and V_0 is consistently estimated by

$$\hat{V}_n = n^{-1} \sum_{i=1}^{n} \mathring{f}(\theta_n(\hat{\Omega}), X_i) \hat{\Omega}_i^{-1} \mathring{f}(\theta_n(\hat{\Omega}), X_i)^T. \tag{10}$$

This paper does not satisfactorily solve the problem of choosing the nonparametric weights in an optimal way. Consistent nonparametric regression estimates based on smoothing parameters which optimize some criterion or cross-validation function are in abundant supply, see, for example, [8] for a survey. It has not been proved that semiparametric estimates based on these cross-validated estimates are optimal with respect to any criterion. The cross-validation function in these cases should be designed in terms of the parameters to be estimated in the semiparametric model (see [15]). This problem seems much harder than ours and, possibly, stronger conditions are required. We can relax the assumption that the nonparametric weights are not data dependent as suggested in [10].

N1. k is (possibly) random and with probability approaching one $k \in \{k_1(n), \ldots, k_{\mathcal{L}}(n)\}$, where $k_1(n), \ldots, k_{\mathcal{L}}(n)$ are nonrandom sequences.

This says little to the practitioner about how to choose the number of nearest neighbors. The practitioner has to verify whether or not the $k_i(n)$, computed by any method, satisfy N1 and W2.

4. ASYMPTOTICS

Let us introduce the following extra notation:

$$\ddot{f}(\theta, x) = \frac{\partial^2 f(\theta, x)}{\partial \theta} \partial \theta^T; \qquad \bar{\lambda}(A) = \sup_a \frac{a^T A a}{a^T a}; \qquad \underline{\lambda}(A) = \inf_a \frac{a^T A a}{a^T a};$$

for any (possibly nonsquare) matrix B, $\|B\| = \bar{\lambda}((B^T B)^{1/2})$.

We require the following conditions on the joint distribution of (X, Y):

D1. θ^0 is an interior point of the parameter space Θ, which is a compact subset in $\mathbb{R}^p$.

D2. For any $\epsilon > 0$, there exists a $\delta > 0$ such that

$$\inf_{\{\theta \in \Theta : \|\theta - \theta^0\| \ge \epsilon\}} E\{(f(\theta, X) - f(\theta^0, X))^T \Omega(X)^{-1} (f(\theta, X) - f(\theta^0, X))\} \ge \delta.$$

D3. For the same v as in W2,

(i) $f(\theta,x)$ is continuous in Θ for each x and

$$\|f(\theta,x) - f(\theta^0,x)\| \le m_0(x) \qquad \text{where} \quad E|m_0(X)|^{4v/(2v-1)} < \infty.$$

(ii) In a neighborhood of θ^0, $N(\theta^0)$, $f(\theta,x)$ is twice continuously differentiable for each x and

$$\|\mathring{f}(\theta,x)\| \le m_1(x) \qquad \text{where} \quad E|m_1(X)|^{4v/(2v-1)} < \infty,$$

$$\|\ddot{f}(\theta,x)\| \le m_2(x) \qquad \text{where} \quad E|m_2(X)|^{4v/(4v-3)} < \infty.$$

D4. For the same v as in W2, $E\|Y - f(\theta^0,X)\|^{4v} < \infty$.

D5. $\Pr\{\underline{\lambda}(\Omega(X)) \ge \epsilon\} = 1 \qquad$ for some $\quad \epsilon > 0$.

D6. V_0 is positive definite (p.d.).

Assumptions D1, D2, and D3(i) ensure asymptotic global identifiability of θ^0 and are used in the proof of consistency. Similar assumptions are required for consistency in different nonlinear econometric models, see, for example, [1] and references therein. The main difference is the moment condition required in D3(i), which is the same as that assumed in [12] for the regressors. The smoothness conditions D3(ii) are also typically required for the asymptotic normality of nonlinear econometric models, see, for example, [1]. Again the main differences are the moment conditions required for the functions dominating the derivatives of the regression function, which are the same as that assumed in [12] for the regressors. Sufficient conditions for D3 are that X admits a density with bounded support and $f(\theta,x)$ is continuous in Θ for each x and twice continuously differentiable in a neighborhood of θ^0 for each x. This sufficient condition is very strong but very easy to check. Also note that D3 is redundant in models linear in parameters where it suffices to assume moment conditions on the regressors (see [12]). The moment conditions in D3 imply that the functions dominating the regression function and its first derivative have between 2 and 4 moments. The function dominating the second derivative has also between 2 and 4 moments.

The moment condition D4 implies, via Jensen's inequality, that $E\|\Omega(X)\|^{2v} < \infty$. Conditions W2 and D4 indicate a tradeoff between the higher moments of the disturbances and the number of nearest neighbors required. This is the main motivation of not assuming fixed moments in D3 and D4 and fixed rates of convergence in W2. In particular, with infinite moments for the disturbances, only two moments in D3 are required and we can use an arbitrarily small number of nearest neighbors provided, of course, that the weights converge to zero as the sample size increases.

Assumption D5 implies that the conditional variance matrices are p.d. a.s.

THEOREM. *Under* W1, W2, D1–D6,

(i) $n^{1/2}(\theta_n(\hat{\Omega}) - \theta^0) \underset{d}{\rightarrow} N(0, V_0^{-1})$,

(ii) $\hat{V}_n = V_0 + o_p(1)$. ■

The proof of the theorem is in the appendix. The main technical difficulty in the proof is to obtain asymptotic normality using the weakest possible moment conditions and slowest possible rates of convergence for k, dealing with nonparametric estimates which are not root-n-consistent. Our proof uses the same general strategy developed in [12], but introduces some alternative arguments which lead to a proof that is fairly short and easy to follow.

Our semiparametric estimator is included in a general class considered by Andrews [2]. Most of the Andrew's conditions for asymptotic normality are verifiable from the proof of our theorem and the assumed conditions. However, the condition in [2] relating to stochastic equicontinuity is not directly verifiable from our results.

Since V_0^{-1} is dominated by either (5) or (6), the theorem implies that $\theta_n(\hat{\Omega})$ is better than $\theta_n(I_m)$, $\theta_n(S_n)$ or any GLS estimate based on an incorrect parameterization of the residual variances, in the sense of asymptotically smaller confidence ellipsoids (or concentration ellipsoids). In particular, $\theta_n(\hat{\Omega})$ possesses a concentration ellipsoid contained entirely within that of the asymptotic distribution of $\theta_n(I_m)$ or $\theta_n(S_n)$. Though consistent standard errors of $\theta_n(I_m)$ or $\theta_n(S_n)$ can be computed (see, e.g., [6] and [17]), the resulting confidence intervals are asymptotically inefficient with respect to that based on $\theta_n(\hat{\Omega})$ and $\hat{V}_n^{-1}$.

The theorem immediately follows assuming the nonparametric weights are data dependent in the way described in N1.

COROLLARY. *The theorem also follows under* N1. ■

The validity of our theorem for approximating the distribution of $\theta_n(\hat{\Omega})$ in small samples and a given k can be legitimately questioned. In the absence of a more refined asymptotic theory, we offer, in the next section, some Monte Carlo experiments for illustrating the performance of the estimate in finite samples.

5. MONTE CARLO

The Monte Carlo experiments try to compare the performance of $\theta_n(I_m)$, $\theta_n(S_n)$, $\theta_n(\hat{\Omega})$, and $\theta_n(\text{diag}\,(\hat{\Omega}))$ (which uses $\Sigma_i = \text{diag}(\hat{\Omega}_i)$ in (2)) and their corresponding asymptotic covariance estimates in the following situations:

- Different sample sizes.
- Different number of nearest neighbors.
- Different heteroskedasticity severity.
- Different number of regressors.

We have chosen the following design

$$Y_{1i} = \exp\{X_i^T\theta_1^0\} + \exp\{-\gamma(\alpha^0 + X_i^T\delta^0)\}U_i\{E(\exp\{-2\gamma(\alpha^0 + X_1^T\delta^0)\})\}^{-1/2} \qquad (11)$$

$$Y_{2i} = \exp\{X_i^T\theta_2^0\} + |\alpha^0 + X_i^T\delta^0|^{2\gamma}(\rho U_i + (1-\rho^2)^{1/2}V_i)\{E|\alpha^0 + X_1^T\delta^0|^{4\gamma}\}^{-1/2}. \quad \textbf{(12)}$$

The X_i are i.i.d. generated according to a uniform (0,1); U_i and V_i are mutually independent generated as i.i.d. standard normals and independent of X_i. At each Monte Carlo iteration new U_i, V_i, and X_i are generated. The parameters in $\theta^0 = (\theta_1^{0T}, \theta_2^{0T})^T$ all take the value 0.5 and $\alpha^0 = 1$, $\delta_j^0 = 1, j = 1, \ldots, r$. The parameter ρ is the correlation coefficient between the disturbances of each equation. The random variates were generated using NAG-13. The expectations in (11) and (12) have been computed by numerical integration using NAG-13. The nonparametric estimates in (7) have been computed using all the r regressors entering in the conditional mean and $\tilde{\theta}_n = \theta_n(I_2)$. All the estimates of θ^0 have been computed without introducing cross-equational restrictions; that is, the information $\theta_1^0 = \theta_2^0$ is not used.

We are using two different heteroskedasticity models for each equation and the severity of the heteroskedasticity is controlled by the parameter γ. These heteroskedasticity models are popular in the econometric literature. Note that the regression functions for each equation are the same. Therefore, $\theta_n(\text{diag}(\hat{\Omega}))$, $\theta_n(\hat{\Omega})$, $\theta_n(I_2)$, $\theta_n(S_n)$, and $\theta_n(\Omega)$ are asymptotically equivalent (see [7] and [18]) when $\gamma = 0$. Moreover, $\theta_n(\text{diag}(\hat{\Omega}))$, $\theta_n(\hat{\Omega})$ are also asymptotically equivalent when $\rho = 0$. $\theta_n(\hat{\Omega})$ is the sole feasible and asymptotically efficient estimate only when $\rho \neq 0$ and $\gamma \neq 0$.

The asymptotic variance of the parameter estimates for each equation are the same. This permits us to report only the average of the summary statistics for each equation. The exponential linear model has been used in other numerical works and it seems of practical relevance. By dividing the disturbances by the expectation of the conditional variances in each equation we cause $E(\text{diag}(\Omega(X))) = I_2$. Then the residual variances are of the same magnitude for different values of γ. At the same time the severity of the heteroskedasticity, measured by the population coefficient of variation of the conditional variances for each equation (CV_1 and CV_2), can vary arbitrarily.

For the different experiments we always offer Monte Carlo summary statistics and its corresponding asymptotic values. The population CV_α is also reported as well as the value of $E(\omega_{\alpha\beta}(X))$ $\alpha \neq \beta$ where $\omega_{\alpha\beta}(X)$ is the $\alpha\beta$-th component of $\Omega(X)$. All the asymptotic values were computed by numerical integration using NAG-13. The programs were written in FORTRAN-77 and run on a VAX machine. The nonparametric estimates use uniform k-NN weights.

Table 1 reports results for two small sample sizes, $r = 1$, $\rho = 0.9$ and γ taking different values. In the absence of heteroskedasticity, the behavior of the semiparametric estimates improves as the number of nearest neighbors increases. This is because as k approaches n, the semiparametric estimate approaches $\theta_n(I_2)$ or $\theta_n(S_n)$. In the presence of heteroskedasticity, the choice of the smoothing parameter does affect the relative efficiency of the semi-

TABLE 1. Monte Carlo results for $r = 1$ and $\rho = 0.9$

			$\gamma = 0$; $CV_1 = CV_2 = 1$, $E(\omega_{12}(X)) = 0.9$					
			$n = 50$ 10,000 Replications			$n = 100$ 2000 Replications		
		asy var	var	eff	asy^var	var	eff	asy^var
$\theta_n(I_2)$	θ_1	1.392	0.029	1.000	1.414	0.014	1.000	1.387
	θ_2	1.392	0.029	1.000	1.420	0.014	1.000	1.392
$\theta_n(S_n)$	θ_1	1.392	0.029	1.000	1.413	0.014	1.000	1.387
	θ_2	1.392	0.029	1.000	1.419	0.014	1.000	1.392
$\theta_n(\hat{\Omega})$								
$k = \lfloor n^{1/2} \rfloor$	θ_1	1.392	0.107	0.273	1.232	0.022	0.644	1.074
	θ_2	1.392	0.157	0.187	3.051	0.021	0.646	1.080
$k = \lfloor n^{2/3} \rfloor$	θ_1	1.392	0.036	0.804	1.273	0.015	0.908	1.277
	θ_2	1.392	0.036	0.802	1.279	0.015	0.906	1.285
$k = \lfloor n^{4/5} \rfloor$	θ_1	1.392	0.031	0.932	1.386	0.014	0.976	1.362
	θ_2	1.392	0.031	0.935	1.390	0.014	0.977	1.367
$k = \lfloor n^{5/6} \rfloor$	θ_1	1.392	0.031	0.951	1.408	0.014	0.982	1.374
	θ_2	1.392	0.031	0.954	1.407	0.014	0.986	1.379
$\theta_n(\text{diag}(\hat{\Omega}))$								
$k = \lfloor n^{1/2} \rfloor$	θ_1	1.392	0.053	0.547	1.333	0.018	0.781	1.180
	θ_2	1.392	0.054	0.542	1.317	0.018	0.775	1.187
$k = \lfloor n^{2/3} \rfloor$	θ_1	1.392	0.033	0.875	1.329	0.015	0.943	1.316
	θ_2	1.392	0.033	0.880	1.333	0.015	0.945	1.324
$k = \lfloor n^{4/5} \rfloor$	θ_1	1.392	0.030	0.960	1.403	0.014	0.989	1.374
	θ_2	1.392	0.030	0.962	1.407	0.014	0.990	1.380
$k = \lfloor n^{5/6} \rfloor$	θ_1	1.392	0.030	0.970	1.418	0.014	0.989	1.382
	θ_2	1.392	0.030	0.974	1.423	0.014	0.993	1.387
$\theta_n(\Omega)$	θ_1	1.216	0.025	2.013	1.256	0.012	2.037	1.235
	θ_2	0.386	0.008	1.889	0.408	0.004	1.857	0.395
$\theta_n(I_2)$	θ_1	2.403	0.050	1.000	2.419	0.024	1.000	2.383
	θ_2	0.720	0.015	1.000	0.747	0.007	1.000	0.726

continued

parametric estimates. Interestingly, the best results are generally obtained for $k = \lfloor n^{4/5} \rfloor$ ($\lfloor a \rfloor$ is the greatest integer less than or equal to a). The asymptotic mean square error of uniform k-NN nonparametric estimates is minimized by $k = Cn^{4/5}$, where C is a given constant (see discussion in [8] Section 3.2). It is also interesting to note that when $\gamma = 1$, inefficient esti-

Table 1 continued

		$\gamma = 1$; $CV_1 = 0.69$, $CV_2 = 0.56$, $E(\omega_{12}(X)) = 0.726$						
			$n = 50$ 10,000 Replications			$n = 100$ 2000 Replications		
		asy var	var	eff	asy^var	var	eff	asy^var
$\theta_n(S_n)$	θ_1	2.403	0.049	1.017	2.391	0.024	1.008	2.371
	θ_2	0.720	0.015	1.022	0.737	0.007	1.012	0.722
$\theta_n(\hat{\Omega})$								
$k = \lfloor n^{1/2} \rfloor$	θ_1	1.216	0.055	0.910	0.869	0.019	1.253	0.947
	θ_2	0.386	0.017	0.880	0.316	0.006	1.181	0.320
$k = \lfloor n^{2/3} \rfloor$	θ_1	1.216	0.034	1.485	1.184	0.014	1.719	1.200
	θ_2	0.386	0.011	1.389	0.462	0.005	1.560	0.429
$k = \lfloor n^{4/5} \rfloor$	θ_1	1.216	0.033	1.534	1.435	0.014	1.748	1.428
	θ_2	0.386	0.010	1.476	0.663	0.004	1.614	0.594
$k = \lfloor n^{5/6} \rfloor$	θ_1	1.216	0.035	1.442	1.505	0.015	1.650	1.489
	θ_2	0.386	0.011	0.438	0.763	0.005	1.579	0.669
$\theta_n(\text{diag}(\hat{\Omega}))$								
$k = \lfloor n^{1/2} \rfloor$	θ_1	1.997	0.068	0.729	1.597	0.026	0.920	1.665
	θ_2	0.634	0.020	0.773	0.564	0.008	0.891	0.557
$k = \lfloor n^{2/3} \rfloor$	θ_1	1.997	0.034	1.024	1.782	0.022	1.114	1.830
	θ_2	0.634	0.011	1.000	0.686	0.007	1.072	0.651
$k = \lfloor n^{4/5} \rfloor$	θ_1	1.997	0.045	1.104	1.796	0.014	1.155	1.831
	θ_2	0.634	0.014	1.059	0.882	0.004	1.097	0.760
$k = \lfloor n^{5/6} \rfloor$	θ_1	1.997	0.046	1.094	1.767	0.021	1.139	1.801
	θ_2	0.634	0.014	1.058	0.889	0.007	1.094	0.808
$\theta_n(\Omega)$	θ_1	0.450	0.009	7.067	0.458	0.004	7.182	0.455
	θ_2	0.063	0.001	5.300	0.068	7E−5	5.129	0.065
$\theta_n(I_2)$	θ_1	3.124	0.064	1.000	3.110	0.031	1.000	3.084
	θ_2	0.337	0.007	1.000	0.359	0.003	1.000	0.345

continued

mates are obtained choosing $k = \lfloor n^{1/2} \rfloor$. The asymptotic variance estimates are in general very imprecise for the sample sizes considered. In fact the Eicker-White estimates of the asymptotic variances for $\theta_n(I_2)$ or $\theta_n(S_n)$ are less biased than the semiparametric estimates. In practice, an Eicker-White type of estimate can be also used for the semiparametric GLS, as suggested

Table 1 continued

		$\gamma = 2$; $CV_1 = 1.17$, $CV_2 = 1.04$, $E(\omega_{12}(X)) = 0.426$						
			$n = 50$ 10,000 Replications			$n = 100$ 2000 Replications		
		asy var	var	eff	asy^var	var	eff	asy^var
$\theta_n(S_n)$	θ_1	3.124	0.064	1.008	3.091	0.031	1.004	3.076
	θ_2	0.337	0.007	1.021	0.355	0.003	1.012	0.343
$\theta_n(\hat{\Omega})$								
$k = \lfloor n^{1/2} \rfloor$	θ_1	0.450	0.023	2.802	0.395	0.007	4.104	0.398
	θ_2	0.063	0.004	2.029	0.074	0.001	3.003	0.064
$k = \lfloor n^{2/3} \rfloor$	θ_1	0.450	0.016	4.048	0.651	0.006	5.363	0.598
	θ_2	0.063	0.002	2.836	0.133	9E−5	3.602	0.103
$k = \lfloor n^{4/5} \rfloor$	θ_1	0.450	0.021	3.008	1.052	0.007	4.250	0.980
	θ_2	0.063	0.003	2.626	0.279	0.001	3.194	0.212
$k = \lfloor n^{5/6} \rfloor$	θ_1	0.450	0.027	2.327	1.213	0.009	3.357	1.131
	θ_2	0.063	0.003	2.426	0.374	0.001	2.955	0.277
$\theta_n(\text{diag}(\hat{\Omega}))$								
$k = \lfloor n^{1/2} \rfloor$	θ_1	1.429	0.047	1.358	1.107	0.019	1.636	1.190
	θ_2	0.202	0.006	1.128	0.203	0.003	1.309	0.188
$k = \lfloor n^{2/3} \rfloor$	θ_1	1.429	0.036	1.782	1.304	0.016	1.972	1.352
	θ_2	0.202	0.005	1.398	0.269	0.002	1.535	0.235
$k = \lfloor n^{4/5} \rfloor$	θ_1	1.429	0.037	1.712	1.443	0.016	1.909	1.470
	θ_2	0.202	0.005	1.401	0.386	0.002	1.497	0.326
$k = \lfloor n^{5/6} \rfloor$	θ_1	1.429	0.041	1.571	1.486	0.017	1.784	1.502
	θ_2	0.202	0.005	1.366	0.461	0.002	1.466	0.376

var is the Monte Carlo variance. eff is the ratio of the MSE of $\theta_n(I_2)$ to the MSE of the estimate. asy var is the asymptotic variance of $n^{1/2}(\theta_{ni}(*) - \theta_i^0)$, $i = 1, 2$. asy^var is the estimate of asy var, the Eicker-White heteroskedasticity robust estimate is used for $\theta_n(I_2)$ and $\theta_n(S_n)$.

in [13], in order to avoid the effect of bad choices of the smoothing parameter. It is expected that with large sample sizes, available in cross-sectional data sets, better estimates of the asymptotic variances are possible. Note that the Monte Carlo variances are fairly close to the asymptotic ones (multiply var by n). As expected, the multivariate GLS perform better than the univariate GLS. However, the asymptotic variances estimates of the multivariate GLS are more biased than those of the univariate GLS.

Table 2. Monte Carlo results for $r = 5$ and $\rho = 0.9$

		$\gamma = 0$, $CV_1 = CV_2 = 1$, $E(\omega_{12}(X)) = 0.9$				$\gamma = 1$, $CV_1 = 0.708$, $CV_2 = 1.704$, $E(\omega_{12}(X)) = 0.546$			
			$n = 100$, 2000 Replications				$n = 100$, 2000 Replications		
		asy var	var	eff	asy^var	asy var	var	eff	asy^var
$\theta_n(I_r)$	θ_1	0.681	0.007	1.000	0.669	0.898	0.009	1.000	0.824
	θ_2	0.681	0.007	1.000	0.669	0.302	0.003	1.000	0.336
$\theta_n(S_n)$	θ_1	0.681	0.007	1.000	0.530	0.898	0.009	0.999	0.823
	θ_2	0.681	0.007	1.000	0.531	0.302	0.003	1.000	0.336
$\theta_n(\hat{\Omega})$									
$k = \lfloor n^{1/2} \rfloor$	θ_1	0.681	0.010	0.698	0.627	0.197	0.007	1.293	0.404
	θ_2	0.681	0.010	0.690	0.626	0.028	0.002	2.036	0.149
$k = \lfloor n^{2/3} \rfloor$	θ_1	0.681	0.008	0.884	0.626	0.197	0.006	1.482	0.548
	θ_2	0.681	0.008	0.878	0.626	0.028	0.001	2.307	0.241
$k = \lfloor n^{4/5} \rfloor$	θ_1	0.681	0.007	0.952	0.665	0.197	0.007	1.374	0.633
	θ_2	0.681	0.007	0.953	0.665	0.028	0.002	2.061	0.349
$k = \lfloor n^{5/6} \rfloor$	θ_1	0.681	0.007	0.964	0.671	0.197	0.007	1.323	0.649
	θ_2	0.681	0.007	0.967	0.671	0.028	0.002	1.924	0.388
$\theta_n(\text{diag}(\hat{\Omega}))$									
$k = \lfloor n^{1/2} \rfloor$	θ_1	0.681	0.009	0.793	0.586	0.197	0.009	0.981	0.564
	θ_2	0.681	0.009	0.791	0.586	0.028	0.002	1.676	0.211
$k = \lfloor n^{2/3} \rfloor$	θ_1	0.681	0.008	0.923	0.650	0.197	0.008	1.117	0.643
	θ_2	0.681	0.008	0.916	0.650	0.028	0.002	1.870	0.294
$k = \lfloor n^{4/5} \rfloor$	θ_1	0.681	0.007	0.972	0.675	0.197	0.008	1.124	0.678
	θ_2	0.681	0.007	0.970	0.675	0.028	0.002	1.792	0.391
$k = \lfloor n^{5/6} \rfloor$	θ_1	0.681	0.007	0.979	0.679	0.197	0.008	1.114	0.683
	θ_2	0.681	0.007	0.979	0.678	0.028	0.002	1.731	0.424

var is the Monte Carlo variance. eff is the ratio of the MSE of $\theta_n(I_1)$ to the MSE of the estimate. asy^var is the asymptotic variance of $n^{1/2}(\theta_{ni}(\cdot) - \theta_i^0)$, $i = 1, 2$. asy^var is the estimate of asy var, using (10) for the semiparametric estimates and the Eicker-White heteroskedasticity robust estimate for $\theta_n(I_r)$ and $\theta_n(S_n)$. The rows corresponding to var, eff and asy^var report the averages of the Monte Carlo statistics for the 5 parameters in θ_1^0 and θ_2^0, respectively.

In Table 2 more regressors and parameters are introduced. The first-order asymptotic properties of the k-NN estimates or the semiparametric estimates do not depend on the number of regressors. The point estimates behaviors are quite the same as in Table 1. However, the asymptotic variance estimates are even more biased.

In all the experiments the Monte Carlo biases are smaller than the variances for all the estimates considered. In particular, the Monte Carlo mean square errors are equal to the Monte Carlo variances up to the fourth decimal point, in all cases.

6. FINAL REMARKS

This paper has provided regularity conditions for adaptation in the multiple equations nonlinear regression model under heteroskedasticity of unknown form. Our theorem can be directly applied to reduced form estimation of linear simultaneous equation systems and panel data models where the number of periods is fixed. It is expected from [11] and [14] that iterating the GLS, by computing new nonparametric conditional variance estimates at each iteration, will improve the higher order efficiency of the resulting GLS estimates.

The Monte Carlo experiments are encouraging. However, the choice of the smoothing parameter strongly affects the performance of the semiparametric estimates. A judicious choice of the number of nearest neighbors seems imperative. In practice, it is recommended that one provide results for several values of the smoothing parameter. Though good point estimates are generally possible, despite the choice of the number of nearest neighbors, the estimates of the asymptotic covariance matrices are usually poor. Computation of the standard errors by bootstrapping techniques may give better results. It seems interesting to investigate the goodness of the normal approximation in finite samples as well as the behavior of statistics (e.g., the t-ratio) computed from the semiparametric estimates.

An important topic for further research is the design of an automatic mechanism for choosing the smoothing parameter (a possibility is pointed out in [15]), and providing some efficient algorithm for its implementation.

REFERENCES

1. Amemiya, T. *Advanced Econometrics*. Cambridge, MA: Cambridge University Press, 1985.
2. Andrews, D.W.K. Asymptotics for semiparametric econometric models I: Estimation and testing. Preprint 1990.
3. Bellman, R. *Introduction to Matrix Analysis*. New York: McGraw-Hill, 1970.
4. Carroll, R.J. Adapting for heteroscedasticity in linear models. *Annals of Statistics* 10 (1982):1233–1244.
5. Chamberlain, G. Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of Econometrics* 34 (1987):305–334.

6. Eicker, F. Asymptotic normality and consistency of the least squares estimators for families of linear regressions. *The Annals of Mathematical Statistics* 34 (1963):447–456.
7. Gallant, A.R. Seemingly unrelated nonlinear regression. *Journal of Econometrics* 3 (1975):35–50.
8. Härdle, W. *Applied Nonparametric Regression.* Cambridge: Cambridge University Press, 1990.
9. Malinvaud, E. *Statistical Methods of Econometrics.* Amsterdam: North-Holland, 1980.
10. Newey, W.K. Efficient instrumental variable estimation of nonlinear models. *Econometrica* 58 (1990): 809–837.
11. Phillips, P.C.B. The iterated minimum distance estimator and the quasi-maximum likelihood estimator. *Econometrica* 44 (1975):449–460.
12. Robinson, P.M. Asymptotically efficient estimation in the presence of heteroskedasticity of unknown form. *Econometrica* 55 (1987):875–891.
13. Robinson, P.M. Adaptive estimation of heteroskedastic econometric models. *Revista de Econometria* (1987):5–28.
14. Robinson, P.M. The stochastic difference between econometric statistics. *Econometrica* 56 (1988):531–548.
15. Robinson, P.M. Automatic frequency-domain inference on semiparametric and nonparametric models. Forthcoming in *Econometrica* (1991).
16. Stone, C.J. Consistent nonparametric regression (with discussion). *Annals of Statistics* 5 (1977):595–645.
17. White, H. A heteroscedasticity-consistent covariance matrix estimator and a direct test for heteroscedasticity. *Econometrica* 48 (1980):817–838.
18. Zellner, A. An efficient method of estimating seemingly unrelated regression and test for aggregation bias. *Journal of the American Statistical Association* 57 (1962):348–368.

APPENDIX

The following notation will be used through the appendix:

$$U_i = Y_i - f(\theta^0, X_i); \quad \mathring{f}_i(\theta) = \mathring{f}(\theta, X_i); \quad \mathring{f}_i = \mathring{f}_i(\theta^0); \quad \ddot{f}_i(\theta) = \ddot{f}(\theta, X_i);$$

$$\ddot{f}_i = \ddot{f}_i(\theta^0); \quad h_i(\theta) = f(\theta^0, X_i) - f(\theta, X_i); \quad \mu_i^{[t]} = E\{\, \|U_i\|^t \,|\, X_i\};$$

$$W_{ij} = W_j(X_i); \quad \tilde{\Omega}_i = \sum_j U_j U_j^T W_{ij}; \quad \text{and} \quad \bar{\Omega}_i = \sum_j \Omega_j W_{ij}.$$

Proof of the Theorem. This proof uses some auxiliary lemmas proved in the lemmata. In view of D2, the consistency of $\theta_n(\hat{\Omega})$ follows using standard consistency arguments, after proving that

$$\sup_{\Theta} \| Q_n(\theta, \hat{\Omega}) - E(h_1(\theta)^T \Omega(X_1)^{-1} h_1(\theta)) - m \| = o_p(1), \tag{A.1}$$

which is proved from

$$\sup_{\Theta} \| n^{-1} \sum_i h_i(\theta)^T \Omega(X_i)^{-1} h_i(\theta) - E(h_1(\theta)^T \Omega(X_1)^{-1} h_1(\theta)) \| = o_p(1), \tag{A.2}$$

$$\sup_{\Theta} \| n^{-1} \sum_i h_i(\theta)^T \Omega_i^{-1} U_i \| = o_p(1), \tag{A.3}$$

$$\|n^{-1}\sum_i U_i^T\Omega_i^{-1}U_i - m\| = o_p(1), \tag{A.4}$$

$$\sup_\Theta\|n^{-1}\sum_i h_i(\theta)^T(\hat{\Omega}_i^{-1} - \Omega_i^{-1})h_i(\theta)\| = o_p(1), \tag{A.5}$$

$$\sup_\Theta\|n^{-1}\sum_i h_i(\theta)^T(\hat{\Omega}_i^{-1} - \Omega_i^{-1})U_i\| = o_p(1), \tag{A.6}$$

$$\|n^{-1}\sum_i U_i^T(\hat{\Omega}_i^{-1} - \Omega_i^{-1})U_i\| = o_p(1). \tag{A.7}$$

A.2 and A.3 follow by the uniform weak law of large numbers (WLLN) and A.4 by the WLLN, A.5–A.7 applying Lemma 12. Therefore, after applying a mean value theorem argument, it suffices to prove that

$$n^{-1}\sum_i \mathring{f}_i(\bar{\theta}_n)^T\Omega_i^{-1}\mathring{f}_i(\bar{\theta}_n) = V_0 + o_p(1), \tag{A.8}$$

$$n^{-1}\sum_i \ddot{f}_i(\bar{\theta}_n)^T\Omega_i^{-1}U_i = o_p(1), \tag{A.9}$$

$$n^{-1}\sum_i \mathring{f}_i(\bar{\theta}_n)^T(\hat{\Omega}_i^{-1} - \Omega_i^{-1})\mathring{f}_i(\bar{\theta}_n) = o_p(1), \tag{A.10}$$

$$n^{-1}\sum_i \ddot{f}_i(\bar{\theta}_n)^T(\hat{\Omega}_i^{-1} - \Omega_i^{-1})U_i = o_p(1), \tag{A.11}$$

$$n^{-1/2}\sum_i \mathring{f}_i^T\Omega_i^{-1}U_i \underset{d}{\rightarrow} N(0, V_0), \tag{A.12}$$

$$n^{-1/2}\sum_i \mathring{f}_i^T(\hat{\Omega}_i^{-1} - \Omega_i^{-1})U_i = o_p(1), \tag{A.13}$$

for any $\bar{\theta}_n = \theta^0 + o_p(1)$. A.8 and A.9 are proved applying the uniform WLLN, A.10 and A.11 applying Lemma 12, A.12 by the Lindeberg–Levi central limit theorem, and A.13 by

$$\|n^{-1/2}\sum_i \mathring{f}_i^T(\hat{\Omega}_i^{-1} - \tilde{\Omega}_i^{-1})U_i\| = o_p(1), \tag{A.14}$$

$$\|n^{-1/2}\sum_i \mathring{f}_i^T(\tilde{\Omega}_i^{-1} - \bar{\Omega}_i^{-1})U_i\| = o_p(1), \tag{A.15}$$

$$\|n^{-1/2}\sum_i \mathring{f}_i^T(\bar{\Omega}_i^{-1} - \Omega_i^{-1})U_i\| = o_p(1). \tag{A.16}$$

Establishing A.14–A.16 is the problematic part of the proof. The proof of A.14 is based on the fact that the preliminary estimate of θ^0 is root-n-consistent. The proof of A.16 uses that $(\bar{\Omega}_i^{-1} - \Omega_i^{-1})$ and U_j are, conditionally on $(X_i, i \geq 1)$, independent. The proof of A.16 exploits the relationship between the higher moments of U_i and the rate of convergence of the nonparametric weights as suggested in [12]. We proceed to justify formally A.14–A.16. The left side of A.14 is, by Schwarz's inequality and D3, bounded by

$$\{\min_i \underline{\lambda}(\tilde{\Omega}_i)\min_i \underline{\lambda}(\hat{\Omega}_i)\}^{-1}\left\{\sum_i \|\hat{\Omega}_i - \tilde{\Omega}_i\|^2\right\}^{1/2}\left\{n^{-1}\sum_i m_1(X_i)^2\|U_i\|^2\right\}^{1/2}$$
$$= O_p(k^{-1/2}),$$

by Lemmas 5, 10, and 11. In order to prove A.15 note that

$$\begin{aligned}
\tilde{\Omega}_i^{-1} - \bar{\Omega}_i^{-1} &= \bar{\Omega}_i^{-1}(\bar{\Omega}_i - \tilde{\Omega}_i)\tilde{\Omega}_i^{-1} \\
&= \bar{\Omega}_i^{-1}(\bar{\Omega}_i - \tilde{\Omega}_i)\bar{\Omega}_i^{-1} + \bar{\Omega}_i^{-1}(\bar{\Omega}_i - \tilde{\Omega}_i)\bar{\Omega}_i^{-1}(\bar{\Omega}_i - \tilde{\Omega}_i)\tilde{\Omega}_i^{-1} \\
&= \bar{\Omega}_i^{-1}(\bar{\Omega}_i - \tilde{\Omega}_i)\bar{\Omega}_i^{-1} + \bar{\Omega}_i^{-1}(\bar{\Omega}_i - \tilde{\Omega}_i)\bar{\Omega}_i^{-1}(\bar{\Omega}_i - \tilde{\Omega}_i)\bar{\Omega}_i^{-1} \\
&\quad + \bar{\Omega}_i^{-1}(\bar{\Omega}_i - \tilde{\Omega}_i)\bar{\Omega}_i^{-1}(\bar{\Omega}_i - \tilde{\Omega}_i)\bar{\Omega}_i^{-1}(\bar{\Omega}_i - \tilde{\Omega}_i)\tilde{\Omega}_i^{-1} \\
&\qquad \cdots \\
&= \sum_{q=1}^{d} (\bar{\Omega}_i^{-1/2})^T\{\bar{\Omega}_i^{-1/2}(\bar{\Omega}_i - \tilde{\Omega}_i)(\bar{\Omega}_i^{-1/2})^T\}^q\bar{\Omega}_i^{-1/2} \\
&\quad + \{\bar{\Omega}_i^{-1}(\bar{\Omega}_i - \tilde{\Omega}_i)\}^{d+1}\tilde{\Omega}_i^{-1},
\end{aligned}$$

where $\bar{\Omega}_i^{-1} = (\bar{\Omega}_i^{-1/2})^T\bar{\Omega}_i^{-1/2}$. Then setting $d = \lfloor v \rfloor$, where $\lfloor a \rfloor$ is the greatest integer less than or equal to a (the "floor" of a),

$$\|n^{-1/2}\sum_i \mathring{f}_i^T(\tilde{\Omega}_i^{-1} - \bar{\Omega}_i^{-1})U_i\| \le \sum_{q=1}^{d}\left\|\sum_i D_{qi}\right\| + \|T\|$$

where

$$D_{qi} = n^{-1/2}\mathring{f}_i^T(\bar{\Omega}_i^{-1/2})^T\{\bar{\Omega}_i^{-1/2}(\bar{\Omega}_i - \tilde{\Omega}_i)(\bar{\Omega}_i^{-1/2})^T\}^q\bar{\Omega}_i^{-1/2}U_i,$$
$$T = n^{-1/2}\sum_i \mathring{f}_i^I\{\bar{\Omega}_i^{-1}(\bar{\Omega}_i - \tilde{\Omega}_i)\}^{d+1}\tilde{\Omega}_i^{-1}U_i.$$

Therefore, A.15 is proved from $\|T\| = o_p(1)$ and $E\|\sum_i D_{qi}\|^2 = o(1)$. By Schwarz's inequality

$$\|T\| \le \{\min_i \underline{\lambda}(\bar{\Omega}_i)^{d+1}\min_i \underline{\lambda}(\tilde{\Omega}_i)\}^{-1}\max_i\|\tilde{\Omega}_i - \bar{\Omega}_i\|^{d+1-v}\left\{\sum_i \|\tilde{\Omega}_i - \bar{\Omega}_i\|^{2v}\right\}^{1/2}$$
$$\times\left\{n^{-1}\sum_i m_i(X_i)^2\|U_i\|^2\right\}^{1/2}$$
$$= o_p(1),$$

by Lemmas 6, 9, and 10 and $\sum_i\|\tilde{\Omega}_i - \bar{\Omega}_i\|^{2v} = O_p(nk^{-v})$ by Markov's inequality and Lemma 5 (taking $s = v$ and applying Lemma 1). Then A.15 is proved by

$$E\left\|\sum_i D_{qi}\right\|^2 = \sum_i E(D_{qi}^T D_{qi}) + \sum\sum_{j\neq i} E(D_{qi}^T D_{qj}) = o(1).$$

$$\sum_i E(D_{qi}^T D_{qi}) \le KE\{\|U_1\|^2 m_1(X_1)^2\|\tilde{\Omega}_1 - \bar{\Omega}_1\|^{2q}\} = O((nk^{-(v+1)})^{q/v})$$

by Lemmas 8 and 13 (where K, henceforth, is a generic constant).

Let us define $A_i = \bar{\Omega}_i^{-1/2}(\bar{\Omega}_i - \tilde{\Omega}_i)(\bar{\Omega}_i^{-1/2})^T$, $B_{ij} = \bar{\Omega}_i^{-1/2}(U_jU_j^T - \Omega_j)(\bar{\Omega}_i^{-1/2})^T W_{ij}$, and $C_{ij} = \bar{\Omega}_i^{-1/2}\{\sum_{a\neq j}(U_aU_a^T - \Omega_a)W_{ia}\}(\bar{\Omega}_i^{-1/2})^T$. Since A_i^q are, conditional on $(X_i, i \geq 1)$, independent of U_i, and C_{ij} are, conditional on $(X_i, i \geq 1)$, independent of both U_i and U_j,

$$
\begin{aligned}
&n^{-1}\sum\sum_{i\neq j} E\{U_i^T(\bar{\Omega}_i^{-1/2})^T C_{ij}^q \bar{\Omega}_i^{-1/2}\mathring{f}_i\mathring{f}_j^T(\bar{\Omega}_j^{-1/2})^T A_j^q \bar{\Omega}_j^{-1/2}U_j\} \\
&\quad = n^{-1}\sum\sum_{i\neq j} E\{U_i^T(\bar{\Omega}_i^{-1/2})^T A_i^q \bar{\Omega}_i^{-1/2}\mathring{f}_i\mathring{f}_j^T(\bar{\Omega}_j^{-1/2})^T C_{ji}^q \bar{\Omega}_j^{-1/2}U_j\} \\
&\quad = n^{-1}\sum\sum_{i\neq j} E\{U_i^T(\bar{\Omega}_i^{-1/2})^T C_{ij}^q \bar{\Omega}_i^{-1/2}\mathring{f}_i\mathring{f}_j^T(\bar{\Omega}_j^{-1/2})^T C_{ji}^q \bar{\Omega}_j^{-1/2}U_j\} = 0.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
&\sum\sum_{i\neq j} E(D_{qi}^T D_{qj}) \\
&\quad = n^{-1}\sum\sum_{i\neq j} E\{U_i^T(\bar{\Omega}_i^{-1/2})^T(A_i^q - C_{ij}^q)\bar{\Omega}_i^{-1/2}\mathring{f}_i\mathring{f}_j^T(\bar{\Omega}_j^{-1/2})^T(A_j^q - C_{ji}^q)\bar{\Omega}_j^{-1/2}U_j\} \\
&\quad \leq Kn^{-1}\sum\sum_{i\neq j} E\{\|U_i\|\,\|A_i^q - C_{ij}^q\|\,\|\mathring{f}_i\|\,\|\mathring{f}_j\|\,\|A_j^q - C_{ji}^q\|\,\|U_j\|\} \\
&\quad \leq K\sum_{\ell_1=1}^{q}\sum_{\ell_2=1}^{q}\binom{q}{\ell_1}\binom{q}{\ell_2} n^{-1}\sum\sum_{i\neq j} E\{\|U_i\|\,\|B_{ij}\|^{\ell_2}\|C_{ij}\|^{q-\ell_2}\|\mathring{f}_i\| \\
&\qquad\qquad \times \|\mathring{f}_j\|\,\|B_{ji}\|^{\ell_1}\|C_{ji}\|^{q-\ell_1}\|U_j\|\},
\end{aligned}
$$

by Lemma 8 and rearranging terms in $\|A_i^q - C_{ij}^q\|$ and $\|A_j^q - C_{ji}^q\|$. Thus, applying Lemma 8, D3, $E\{\|U_i\|\,\|B_{ji}\|^a\} \leq K\mu_i^{[2a+1]}$, $a \geq 1$, and Hölder's inequality,

$$
\begin{aligned}
&n^{-1}\sum\sum_{i\neq j} E\{\|U_i\|\,\|B_{ij}\|^{\ell_2}\|C_{ij}\|^{q-\ell_2}\|\mathring{f}_i\|\,\|\mathring{f}_j\|\,\|B_{ji}\|^{\ell_1}\|C_{ji}\|^{q-\ell_1}\|U_j\|\} \\
&\quad \leq Kn^{-1}\sum\sum_{i\neq j} E\Bigg\{m_1(X_i)m_1(X_j)\mu_i^{[2\ell_1+1]}\mu_j^{[2\ell_2+1]}W_{ji}^{\ell_1}W_{ij}^{\ell_2} \\
&\qquad\qquad \times \Bigg\|\sum_{a\neq i}(U_aU_a^T - \Omega_a)W_{ja}\Bigg\|^{q-\ell_1}\Bigg\|\sum_{a\neq j}(U_aU_a^T - \Omega_a)W_{ia}\Bigg\|^{q-\ell_2}\Bigg\} \\
&\quad \leq K\Bigg\{n^{-1}\sum\sum_{i\neq j} E\Bigg(m_1(X_i)^2\mu_i^{[2\ell_1+1]2/(2\ell_1+1)}\mu_j^{[2\ell_2+1]4\ell_2/(2\ell_2+1)} \\
&\qquad\qquad \times W_{ij}^{2\ell_2}\Bigg\|\sum_{a\neq j}(U_aU_a^T - \Omega_a)W_{ia}\Bigg\|^{2(q-\ell_2)}\Bigg)\Bigg\}^{1/2} \\
&\qquad \times \Bigg\{n^{-1}\sum\sum_{i\neq j} E\Bigg(m_1(X_j)^2\mu_j^{[2\ell_2+1]2/(2\ell_2+1)}\mu_i^{[2\ell_1+1]4\ell_1/(2\ell_1+1)} \\
&\qquad\qquad \times W_{ji}^{2\ell_1}\Bigg\|\sum_{a\neq i}(U_aU_a^T - \Omega_a)W_{ja}\Bigg\|^{2(q-\ell_1)}\Bigg)\Bigg\}^{1/2}.
\end{aligned}
$$

Hence, since $\mu_i^{[a]c/a} \le \mu_i^{[c]}$ for $c \ge a \ge 1$, it suffices to prove that

$$n^{-1}\sum\sum_{i\neq j} E(m_1(X_i)^2\mu_i^{[2]}\mu_j^{[4q]}W_{ij}^{2q}) = o(1), \tag{A.17}$$

$$n^{-1}\sum\sum_{i\neq j} E\Bigg(m_1(X_i)^2\mu_i^{[2]}\mu_j^{[4\ell]}W_{ij}^{2\ell} \times \Bigg\| \sum_{a\neq j}(U_aU_a^T - \Omega_a)W_{ia}\Bigg\|^{2(q-\ell)}\Bigg) = o(1) \quad \text{for} \quad 1 \le \ell < q. \tag{A.18}$$

(A.17) follows by Lemma 14 and the left-hand side of (A.18) is, by Hölder's inequality, bounded by

$$\Bigg\{E\Bigg(m_1(X_1)^2\mu_1^{[2]}\Bigg\|\sum_{a\neq j}(U_aU_a^T-\Omega_a)W_{1a}\Bigg\|^{2q}\Bigg)\Bigg\}^{(q-\ell)/q} \times \Bigg\{\Bigg[E\Bigg(m_1(X_1)^2\mu_1^{[2]}\sum_j \mu_j^{[4\ell]}W_{1j}^{2\ell}\Bigg)\Bigg]^{\ell/q}[E(m_1(X_1)^2\mu_1^{[2]})]^{(q-\ell)/q}\Bigg\}^{\ell/q}$$
$$= o(1),$$

by Lemmas 13 and 14. Finally, A.16 follows by Markov's inequality, since by D.5 and Lemma 8

$$\begin{aligned}
E\Bigg\|n^{-1/2}\sum_i \mathring{f}_i^T(\bar{\Omega}_i^{-1} - \Omega_i^{-1})U_i\Bigg\|^2
&= \operatorname{trace}(E\{\mathring{f}_1^T\bar{\Omega}_1^{-1}(\Omega_1-\bar{\Omega}_1)\Omega_1^{-1}U_1U_1^T\Omega_1^{-1}(\Omega_1-\bar{\Omega}_1)\bar{\Omega}_1^{-1}\mathring{f}_1\})\\
&= \operatorname{trace}(E\{\mathring{f}_1^T\Omega_1^{-1}(\Omega_1-\bar{\Omega}_1)\Omega_1^{-1}(\Omega_1-\bar{\Omega}_1)\bar{\Omega}_1^{-1}\mathring{f}_1\})\\
&= \operatorname{trace}(E\{\mathring{f}_1^T(\Omega_1^{-1}(\Omega_1-\bar{\Omega}_1)\bar{\Omega}_1^{-1} - \Omega_1^{-1}(\Omega_1-\bar{\Omega}_1)\Omega_1^{-1})\mathring{f}_1\})\\
&\le KE|m_1(X_1)^2\|\Omega_1-\bar{\Omega}_1\||\\
&\le K\{E|m_1(X_1)|^{4v/(2v-1)}\}^{(2v-1)/2v}\{E\|\Omega_1-\bar{\Omega}_1\|^{2v}\}^{1/2v} = o(1),
\end{aligned}$$

by Hölder's inequality and Lemma 1. ■

Proof of the Corollary. Immediate, using Newey [10] Lemma A.2. ■

LEMMA 1. *Let $\hbar(\cdot)$ be a Borel function on $\mathbb{R}^r$ such that $E|\hbar(X)|^\gamma < \infty$, $\gamma \ge 1$. Then $E\{\sum_i |\hbar(X_i) - \hbar(X)|^\gamma W_i(X)\} = o(1)$.*

Proof. By [12] Lemma 1. ■

LEMMA 2. $E|\|U_1\|^2 m_i(X_1)^2| < \infty$, $i = 0, 1$, *and* $E|\|U_1\|m_2(X_1)|^{2v/(2v-1)} < \infty$.

Proof. By Hölder's inequality D3 and D4,

$$E|\|U_1\|^2 m_i(X)^2| \le \{E|m_i(X_1)|^{4v/(2v-1)}\}^{(2v-1)/2v}\{E|\|U_1\|^{4v}\}^{1/2v} < \infty, \quad i = 1,2.$$

$$E|\|U_1\|m_2(X)|^{2v/(2v-1)} \le \{E|m_2(X_1)|^{4v/(4v-3)}\}^{(4v-3)/(4v-2)} \times \{E\|U_1\|^{4v}\}^{1/(4v-2)} < \infty.$$ ■

LEMMA 3. $\max_i \|\hat{\Omega}_i - \tilde{\Omega}_i\| = O_p(k^{-1/2})$.

Proof. Note that

$$\hat{\Omega}_i - \tilde{\Omega}_i = \sum_j h_j(\tilde{\theta}_n) h_j(\tilde{\theta}_n)^T W_{ij} + \sum_j h_j(\tilde{\theta}_n) U_j^T W_{ij} + \sum_j U_j h_j(\tilde{\theta}_n)^T W_{ij}.$$

Then applying a mean value theorem argument, it suffices to prove that

$$\|\bar{\theta}_n - \theta^0\|^2 \max_i \sum_j \|\mathring{f}_j(\bar{\theta}_n)\|^2 W_{ij} = O_p(k^{-1}), \tag{B.1}$$

$$\|\bar{\theta}_n - \theta^0\| \max_i \sum_j \|\mathring{f}_j(\bar{\theta}_n)\| \, \|U_j\| W_{ij} = O_p(k^{-1/2}), \tag{B.2}$$

for any $\bar{\theta}_n \in N(\theta^0): \|\bar{\theta}_n - \theta^0\| \leq \|\tilde{\theta}_n - \theta^0\|$. B.1 is, by D.3, bounded by

$$\|\bar{\theta}_n - \theta^0\|^2 \{\max_{i,j} W_{ij}\} \sum_j m_1(X_j)^2 = O_p(k^{-1});$$

since $\max_{i,j} W_{ij} \leq Kk^{-1}$. B.2 is, by Schwarz's inequality and D.3, bounded by

$$\left\{ \|\bar{\theta}_n - \theta^0\|^2 \sum_j m_1(X_j)^2 \|U_j\|^2 \right\}^{1/2} \left\{ \max_i \sum_j W_{ij}^2 \right\}^{1/2} = O_p(k^{-1/2}),$$

since $\max_i \sum_j W_{ij}^2 \leq Kk^{-1} \max_i \sum_j W_{ij} = Kk^{-1}$. ■

LEMMA 4. $\sum_i \|\hat{\Omega}_i - \tilde{\Omega}_i\|^2 = O_p(k^{-1})$.

Proof. The lemma follows from

$$\sum_i \left\{ \sum_j \|h_j(\tilde{\theta}_n)\|^2 W_{ij} \right\}^2 = O_p(k^{-1}),$$

$$\sum_i \left\| \sum_j h_j(\tilde{\theta}_n)^T U_j W_{ij} \right\|^2 = O_p(k^{-1}).$$

Then using a mean value theorem argument and D.3, it suffices to prove that, for any $\bar{\theta}_n \in N(\theta^0): \|\bar{\theta}_n - \theta^0\| \leq \|\tilde{\theta}_n - \theta^0\|$,

$$\|\bar{\theta}_n - \theta^0\|^4 \left\{ \max_i \sum_j m_s(X_j)^2 W_{ij} \right\} \sum_i \sum_j m_s(X_j)^2 W_{ij} = O_p(k^{-1}), \quad s = 1,2, \tag{B.3}$$

$$\|\bar{\theta}_n - \theta^0\|^2 \sum_i E \left\| \sum_j \mathring{f}_j U_j^T W_{ij} \right\|^2 = O_p(k^{-1}). \tag{B.4}$$

B.3 follows from B.1 and Lemma 1, and B.4 from

$$n^{-1} \sum_i E \left\| \sum_j \mathring{f}_j U_j^T W_{ij} \right\|^2 = E\left\{ \sum_j \|\mathring{f}_j U_j^T\|^2 W_{ij}^2 \right\} \leq Kk^{-1} E\left\{ \sum_j \|\mathring{f}_j^T U_j\|^2 W_{1j} \right\}$$

by Lemmas 1 and 2. ■

LEMMA 5. $E\{ \|\sum_{j\leq n} (U_j U_j^T - \Omega_j) W_{ij}\|^{2s} | X_i, i \geq 1\} \leq Kk^{-s} \sum_{j\leq n} \mu_j^{[4s]} W_{ij}$ for $s \leq v$.

Proof. Immediate from [12] Lemma 7. ■

LEMMA 6. $\max_i\|\bar{\Omega}_i - \tilde{\Omega}_i\| = O_p(n^{1/\nu}k^{-1})$.

Proof. Immediate from [12] Lemma 9. ■

LEMMA 7. $\Pr\{\min_i \underline{\lambda}(\Omega_i) \leq \epsilon\} = 0$ *for any sample size and some* $\epsilon > 0$.

Proof.

$$\Pr\{\min_i \underline{\lambda}(\Omega_i) \leq \epsilon\} \leq n \Pr\{\underline{\lambda}(\Omega_1) \leq \epsilon\} = 0.$$ ■

LEMMA 8. $\Pr\{\underline{\lambda}(\bar{\Omega}_1) \leq \epsilon\} = 0$ *for any sample size and some* $\epsilon > 0$.

Proof. Since $\{\Omega_i, i \geq 1\}$ are p.d. by D5, then

$$\underline{\lambda}(\bar{\Omega}_1) = \underline{\lambda}\left(\sum_j \Omega_j W_{1j}\right) \geq \sum_j \underline{\lambda}(\Omega_j) W_{1j}$$

(see, e.g., Bellman [3], pp. 114), therefore

$$\Pr\{\underline{\lambda}(\bar{\Omega}_1) \leq \epsilon\} \leq \Pr\left\{\sum_j \underline{\lambda}(\Omega_j)W_{1j} \leq \epsilon\right\} \leq \Pr\left\{\min_i \underline{\lambda}(\Omega_i) \sum_j W_{1j} \leq \epsilon\right\}$$

$$= \Pr\{\min_i \underline{\lambda}(\Omega_i) \leq \epsilon\} = 0$$

by Lemma 7. ■

LEMMA 9. $\Pr\{\min_i \underline{\lambda}(\bar{\Omega}_i) \leq \epsilon\} = 0$ *for any sample size and some* $\epsilon > 0$.

Proof.

$$\Pr\{\min_i \underline{\lambda}(\bar{\Omega}_i) \leq \epsilon\} \leq n \Pr\{\underline{\lambda}(\bar{\Omega}_1) \leq \epsilon\} = 0$$

by Lemma 8. ■

LEMMA 10. $\{\min_i \underline{\lambda}(\tilde{\Omega}_i)\}^{-1} = O_p(1)$.

Proof.

$$\min_i \underline{\lambda}(\bar{\Omega}_i) \leq \min_i \underline{\lambda}(\tilde{\Omega}_i) + \max_i\|\tilde{\Omega}_i - \bar{\Omega}_i\| = \min_i \underline{\lambda}(\tilde{\Omega}_i) + o_p(1)$$

by Lemma 6, then the lemma follows applying Lemma 9. ■

LEMMA 11. $\{\min_i \underline{\lambda}(\hat{\Omega}_i)\}^{-1} = O_p(1)$.

Proof.

$$\min_i \underline{\lambda}(\tilde{\Omega}_i) \leq \min_i \underline{\lambda}(\hat{\Omega}_i) + \max_i\|\hat{\Omega}_i - \tilde{\Omega}_i\| = \min_i \bar{\lambda}(\hat{\Omega}_i) + o_p(1)$$

by Lemma 3, then the lemma follows applying Lemma 10. ■

LEMMA 12. $n^{-1}\sum_i Z_i(\hat{\Omega}_i^{-1} - \Omega_i^{-1}) = o_p(1)$, *where* Z_i *can be equal to* $m_0(X_i)^2$, $m_0(X_i)\|U_i\|$, $\|U_i\|^2$, $m_1(X_i)^2$, $m_1(X_i)\|U_i\|$ or $m_2(X_i)\|U_i\|$.

Proof. It suffices to prove that

$$\{\min_i \underline{\lambda}(\tilde{\Omega}_i)\min_i \underline{\lambda}(\hat{\Omega}_i)\}^{-1}\max_i\|\hat{\Omega}_i - \tilde{\Omega}_i\| n^{-1}\sum_i \mathbb{Z}_i = o_p(1), \quad \textbf{(B.5)}$$

$$\{\min_i \underline{\lambda}(\bar{\Omega}_i)\min_i \underline{\lambda}(\tilde{\Omega}_i)\}^{-1}\max_i\|\tilde{\Omega}_i - \bar{\Omega}_i\| n^{-1}\sum_i \mathbb{Z}_i = o_p(1), \quad \textbf{(B.6)}$$

$$\{\min_i \underline{\lambda}(\bar{\Omega}_i)\min_i \underline{\lambda}(\Omega_i)\}^{-1} n^{-1}\sum_i \mathbb{Z}_i\|\bar{\Omega}_i - \Omega_i\| = o_p(1). \quad \textbf{(B.7)}$$

B.5 follows by Lemmas 3, 10, and 11, B.6 by Lemmas 6, 9, and 10, and B.7 by Lemmas 7, 9 and Markov's inequality, since, by Hölder's inequality,

$$E\|\mathbb{Z}_1\|\bar{\Omega}_1 - \Omega_1\|\,| \le \{E|\mathbb{Z}_1|^{2v/(2v-1)}\}^{(2v-1)/2v}\{E\|\bar{\Omega}_1 - \Omega_1\|^{2v}\}^{1/2v} = o(1),$$

by Lemmas 1 and 2. ∎

LEMMA 13. *For any* $1 \le q \le \lfloor v \rfloor$,

$$E\left(m_1(X_1)^2\mu_1^{[2]}\left\|\sum_{j\le n}(U_jU_j^T - \Omega_j)W_{1j}\right\|^{2q}\right) = O((nk^{-(v+1)})^{q/v}). \quad \textbf{(B.8)}$$

Proof. By Lemma 5, Hölder's inequality and $\sum_{j\le n} W_{ij} \le 1$, the left-hand side of (B.8) is bounded by

$$Kk^{-q}E\left(m_1(X_1)^2\mu_1^{[2]}\sum_{j\le n}\mu_j^{[4q]}W_{1j}\right)$$

$$\le Kk^{-q}E\left[m_1(X_1)^2\mu_1^{[2]}\left(\sum_{j\le n}\mu_j^{[4q]v/q}W_{1j}\right)^{q/v}\left(\sum_{j\le n}W_{1j}\right)^{(v-q)/v}\right]$$

$$\le Kk^{-q}E\left[m_1(X_1)^2\mu_1^{[2]}\left(\sum_{j\le n}\mu_j^{[4v]}W_{1j}\right)^{q/v}\right]$$

$$\le Kk^{-q}\left[E\left(m_1(X_1)^2\mu_1^{[2]}\sum_{j\le n}\mu_j^{[4v]}W_{ij}\right)\right]^{q/v}(E(m_1(X_1)^2\mu_1^{[2]}))^{(v-q)/v}$$

$$\le Kk^{-(q+q/v)}E(m_1(X_1)^2\mu_1^{[2]})\left[E\left(\sum_{1<j\le n}\mu_j^{[4v]}\right)\right]^{q/v}$$

$$\le K(nk^{-(1+v)})^{q/v}E(m_1(X_1)^2\mu_1^{[2]})(E(\mu_1^{[4v]}))^{q/v} = O((nk^{-(1+v)})^{q/v}).$$ ∎

LEMMA 14. *For any* $1 \le q \le \lfloor v \rfloor$,

$$E\left\{\mu_1^{[2]}m_1(X_1)^2\sum_i\mu_i^{[4q]}W_{1i}^{2q}\right\} = O(k^{1-q}(nk^{-(v+1)})^{q/v}). \quad \textbf{(B.9)}$$

Proof. The left side of B.9 is bounded by

$$Kk^{1-2q}E\left\{\mu_1^{[2]}m_1(X_1)^2\sum_i\mu_i^{[4q]}W_{1i}\right\}.$$

Thus, the lemma follows as Lemma 13. ∎