



ELSEVIER

Journal of Econometrics 101 (2001) 109–122

JOURNAL OF
Econometrics

www.elsevier.nl/locate/econbase

Nonlinear estimation using estimated cointegrating relations

Robert M. de Jong

Department of Economics, Michigan State University, 215 Marshall Hall, East Lansing, MI 48824, USA

Received 15 January 1998; received in revised form 15 May 2000; accepted 1 August 2000

Abstract

The Granger–Engle procedure consists of two steps. In the first step, a long-run cointegrating relationship is estimated, and in the second stage, this estimated long-run relationship is used to estimate a distributed lag model. This paper establishes the limit distribution of the second-stage estimator if the model estimated in the second stage is other than linear. One may expect that the estimation of the cointegrating relationship does not affect the limit distribution of the second-stage estimator; however, it is shown that unless a regularity condition holds, this intuition is false. Clearly this regularity condition holds in the standard linear case. A simple example where the limit distribution changes is the addition of the square of the cointegrating relationship to the second stage distributed lag model that is estimated by least squares. Surprisingly however, it turns out that if a constant is included in the long-run least-squares regression, the (possibly nonlinear) second-stage estimator will be asymptotically normally distributed. © 2001 Elsevier Science S.A. All rights reserved.

JEL classification: C22; C32

Keywords: Cointegration; Unit root; Time series; Nonlinearity

1. Introduction

In this paper, the estimation of parametric models that are nonlinear in the cointegrating relationship will be considered. The concept of cointegration was

E-mail address: dejongr@pilot.msu.edu (R.M. de Jong).

introduced by Granger (1981) and extended in Engle and Granger (1987), Engle (1987), Engle and Yoo (1987), Phillips and Ouliaris (1990), Phillips (1991), and Johansen (1988, 1991). The literature on this subject is now enormous. This paper considers the following problem. Assume that x_t and y_t are I(1) processes and assume that there is some θ_0 such that $\varepsilon_t = y_t - \theta'_0 x_t$ is I(0). The first step of the famous Granger–Engle procedure is then to obtain an estimator $\hat{\theta}$ of θ_0 , for example the OLS estimator. In the second step, $y_{t-1} - \hat{\theta}'x_{t-1}$ is used as a regressor to estimate a relationship such as, for example,

$$\Delta y_t = \beta_0(y_{t-1} - \theta'_0 x_{t-1}) + \eta_t. \quad (1)$$

Clearly if θ_0 is known, the last relationship only involves I(0) random variables, which makes the standard asymptotic theory valid. However, to estimate β_0 , one needs to replace θ_0 by the consistent estimate $\hat{\theta}$. This implies that a two-stage estimation procedure is used here. The results obtained in the papers mentioned above can be interpreted as stating that in the case of the linear model above and least-squares estimation, the estimation in the first step does not affect the limit distribution of the second-stage estimator.

It is tempting to add extra functions of $y_{t-1} - \hat{\theta}'x_{t-1}$ to Eq. (1), in the hope of improving our model, or as a test of model specification. For example, for model specification testing purposes, one may be tempted to add a square of $y_{t-1} - \hat{\theta}'x_{t-1}$ to our estimated equation, and test the null hypothesis that the coefficient for that variable is zero using the standard t -statistic. In Section 3, it is shown that adding the square of the estimated cointegrating relationship invalidates the standard asymptotic distribution results. In general, it is shown that verification of an orthogonality condition is necessary to justify the use of standard asymptotic distribution results. Also, recently some models have been proposed that are nonlinear in the cointegrating relationship; see for example Granger and Terasvirta (1993) and Davidson and Peel (1998). The STECM model proposed by Granger and Terasvirta (1993) is

$$\Delta y_t = \beta_{01} + \beta_{02}w_t + (\beta_{03} + \beta_{04}w_t)F(\varepsilon_{t-d}) + \eta_t, \quad (2)$$

where w_t is an I(0) regressor and $F(\cdot)$ is some distribution function, possibly depending on parameters that will have to be estimated also; for example, the logistic distribution function

$$F(a) = (1 - \exp(-r(a - c)))^{-1}, \quad (3)$$

where $r \geq 0$. This paper points out how this model can be estimated and how an asymptotically normally distributed estimator of this model can be obtained. This paper considers the general case of nonlinear parametric models; it for example provides an asymptotic theory for the case where the model is

$$\Delta y_t = h(y_{t-1} - \theta'_0 x_{t-1}, \beta_0) + \eta_t, \quad (4)$$

where η_t is assumed to be stationary while $y_t \in \mathbb{R}$ and $x_t \in \mathbb{R}^k$ are I(1), and y_t is cointegrated with x_t . The even more general case of minimizing a criterion function in the presence of an estimated cointegrating relationship is also covered by the results. Note here that if a model that is nonlinear in the cointegrating relationship is estimated, it is unclear what long-run dynamics are implied by the model, since in general we will not be able to use a result such as the Granger–Engle Representation Theorem to make this connection. The results below assume that the reader is willing to consider nonlinear two-step estimation in the presence of a cointegrating relationship.

One may suspect that the estimation of θ_0 is innocent because of the order- n consistency of its estimator. The first theorem of this paper illustrates that in general, this intuition is incorrect. It turns out that in general an orthogonality condition needs to be met in order for the limit distribution of the second-stage estimator to be not affected by the first-stage estimation. This condition is reminiscent of the condition that appears in the theory of two-stage estimators in the i.i.d. framework. This condition, obviously, is met in the linear case, but need not necessarily be met in the general case. The second theorem of this paper shows that it is possible to obtain an asymptotically normally distributed estimator of the short-run parameter by including an intercept in the long-run least-squares regression. This result implies an important recommendation for applied econometricians interested in estimating this type of model.

The mistake of acting as if the long-run parameter is known (without checking the appropriate condition for the specific model that is estimated) is easily made and examples of this mistake can be found in the literature. For example, Hendry and Ericsson (1991) consider, in their terminology, a dynamic equilibrium correction model. Basically, they add a third-order polynomial in the long-run regression error, with some restrictions on parameters of this third-order polynomial, to their short-run specification, and justify acting as if the long-run parameter is known by referring to the superconsistency of the long-run parameter estimator. See also Ericsson et al. (1998, p. 296) for a discussion. Davidson and Peel (1998) estimate a bilinear model and act as if θ_0 is known. As will be argued later, unless an orthogonality condition holds, this procedure is incorrect.

Section 2 of this paper states the main result of this paper. In Section 3, three simple examples are provided that illustrate the results. This paper concludes with a Mathematical Appendix.

2. Main result

Assume that the second-stage estimator is defined as

$$\hat{\beta} = \operatorname{argmin}_{\beta \in B} n^{-1} \sum_{t=1}^n f(w_t, \varepsilon_t + (\theta_0 - \hat{\theta})' z_t + a_n, \beta), \quad (5)$$

where w_t and $\varepsilon_t = y_t - \theta'_0 x_t$ are stationary random variables and B is the parameter space. We could choose $z_t = x_t$ and $a_n = 0$ in what follows, implying that $y_t - \hat{\theta}'x_t$ is included, or alternatively we could choose $z_t = x_t - \bar{x}$ and $a_n = -\bar{\varepsilon}$, implying that $y_t - \bar{y} - \hat{\theta}'(x_t - \bar{x})$ is included. If a constant would have been included in the (least squares) long-run cointegrating regression, the residual from that regression would equal

$$\varepsilon_t + (\theta_0 - \hat{\theta})(x_t - \bar{x}) - \bar{\varepsilon}, \tag{6}$$

where $\hat{\theta}$ denotes the least-squares estimator (excluding the estimator of the intercept) from the regression of y_t on x_t and a constant. Therefore, the last choice of z_t and a_n refers to the inclusion of a constant in the long-run regression. For the consistency proof, it will usually be required that B is compact. Also, assume that the estimator $\hat{\theta}$ of θ_0 satisfies

$$n(\hat{\theta} - \theta_0) \xrightarrow{d} D \tag{7}$$

for some random variable D . In this paper, the notation \xrightarrow{d} and \xrightarrow{p} denotes convergence in distribution and in probability, respectively. Assume that

$$n^{-1/2} \sum_{t=1}^n (\partial/\partial\beta')f(w_t, \varepsilon_t + a_n, \beta_0) \xrightarrow{d} W, \tag{8}$$

where W is a normally distributed random variable. Both are natural assumptions to make in the present framework. Define

$$M = E(\partial/\partial\beta)(\partial/\partial\beta')f(w_t, \varepsilon_t, \beta_0), \tag{9}$$

and

$$N = E(\partial/\partial\beta'')(\partial/\partial\varepsilon)f(w_t, \varepsilon_t, \beta_0). \tag{10}$$

Everywhere in this paper, $|\cdot|$ denotes the Euclidean norm. Let $G_Z(\cdot)$ denote a Gaussian process that satisfies Assumption (1.4) below. In order to establish the results below, the following assumption is assumed to hold:

Assumption 1.

1. $\hat{\beta} \xrightarrow{p} \beta_0$, where β_0 is in the interior of the parameter space B , B is compact and convex, and $B \subseteq \mathbb{R}^m$.
2. $f(\cdot, \cdot, \cdot)$ is a continuous function from $\mathbf{W} \times \mathbb{R} \times B$ to \mathbb{R} , where $\mathbf{W} \subseteq \mathbb{R}^q$, and for $j = 1, \dots, m$,

$$(\partial/\partial\beta'')(\partial/\partial\varepsilon)f(w, \varepsilon, \beta) \tag{11}$$

and

$$(\partial/\partial\beta')(\partial/\partial\beta_j)f(w, \varepsilon, \beta) \tag{12}$$

are continuous on $\mathbf{W} \times \mathbb{R}$ uniformly on B and continuous in all arguments.

3. $n^{-3/2} \sum_{t=1}^n ((\partial/\partial\beta')(\partial/\partial\varepsilon)f(w_t, \varepsilon_t, \beta_0) - N)z_t \xrightarrow{p} 0$.
4.
$$\left(n^{-3/2} \sum_{t=1}^n z_t, n(\hat{\theta} - \theta_0), n^{-1/2} \sum_{t=1}^n (\partial/\partial\beta')f(w_t, \varepsilon_t + a_n, \beta_0) \right) \xrightarrow{d} \left(\int_0^1 G_Z(\xi) d\xi, D, W \right), \tag{13}$$

and $\sup_{1 \leq t \leq n} n^{-1/2}|z_t| = O_p(1)$.

5. (ε_t, w_t) is a strictly stationary sequence of random variables that is L_0 -approximable by an α -mixing process.
6. For some open neighborhood Γ of 0 and for some $\phi > 0$ and for $j = 1, \dots, k$,

$$E \sup_{\gamma \in \Gamma} \sup_{\beta \in B} |(\partial/\partial\beta')(\partial/\partial\beta_j)f(w_t, \varepsilon_t + \gamma, \beta)|^{1+\phi} < \infty, \tag{14}$$

and

$$E \sup_{\gamma \in \Gamma} \sup_{\beta \in B} |(\partial/\partial\beta')(\partial/\partial\varepsilon)f(w_t, \varepsilon_t + \gamma, \beta)|^{1+\phi} < \infty. \tag{15}$$

7. M is invertible.

8. $a_n \xrightarrow{p} 0$.

The consistency assumption (1.1) is an innocent one, because consistency proofs are relatively insensitive to the presence of the first-stage estimation. For an example of a possible regularity condition that ensures this, note that, under Assumption 1, by the Taylor series expansion,

$$\begin{aligned} & \sup_{\beta \in B} \left| n^{-1} \sum_{t=1}^n f(w_t, \varepsilon_t + (\theta_0 - \hat{\theta})'z_t, \beta) - f(w_t, \varepsilon_t, \beta) \right| \\ & \leq |n(\hat{\theta} - \theta_0)| n^{-3/2} \sum_{t=1}^n \sup_{\beta \in B} |(\partial/\partial\varepsilon)f(w_t, \varepsilon_t + (\theta_0 - \tilde{\theta})'z_t, \beta)| \sup_{1 \leq t \leq n} n^{-1/2}|z_t| \\ & \leq O_p(1) n^{-3/2} \sum_{t=1}^n \sup_{\beta \in B} \sup_{\gamma \in \Gamma} |(\partial/\partial\varepsilon)f(w_t, \varepsilon_t + \gamma, \beta)| \end{aligned} \tag{16}$$

with arbitrary large probability for n large enough, where Γ denotes any neighborhood of 0, and $\tilde{\theta}$ is a mean value that is order- n -consistent. Therefore,

a regularity condition such as

$$E \sup_{\beta \in B} \sup_{\gamma \in \Gamma} |(\partial/\partial \varepsilon)f(w_t, \varepsilon_t + \gamma, \beta)| < \infty \tag{17}$$

ensures that the expression of Eq. (16) is $O_p(n^{-1/2})$. This in turn implies asymptotic equivalence of both criterion functions, and therefore suffices to obtain the consistency result for $\hat{\beta}$ if such a result is available for the case where θ_0 is known. Assumption (1.2) is a regularity condition. Assumption (1.3) is an assumption on the average of a product of a mean zero I(0) times a mean zero I(1) random variable; typically, it is possible to set forth conditions implying that this term is $O_p(n^{-1/2})$, but the present condition suffices here. One possibility for obtaining the $O_p(n^{-1/2})$ result is to set forth conditions implying that term satisfies convergence in distribution to a stochastic integral; see for example Chan and Wei (1988), Davidson (1994), or De Jong and Davidson (1998). Assumption (1.4) assumes that the x_t are I(1) and assumes the presence of “standard” limit random variables D and W . This condition can also be derived from more primitive assumptions. Note that convergence in distribution of $n^{-3/2} \sum_{t=1}^n z_t$ and $n^{-1/2} \sup_{1 \leq t \leq n} |z_t|$ follows (by the continuous mapping theorem) from weak convergence of $n^{-1/2} z_{[n\xi]}$ to some limit process $G_Z(\xi)$ that takes on values in $C[0, 1]$. Assumptions (1.5) and (1.6) serve to be able to apply (uniform) laws of large numbers where needed. The assumption of L_0 -approximability by an α -mixing process is one of asymptotically weak dependence; see for example Pötscher and Prucha (1997) for details regarding this assumption. ARMA processes, as well as smooth functions of ARMA processes, can be shown to satisfy the L_0 -approximability condition under regularity conditions.

Assumptions (1.4) and (1.5) are high-level assumptions. De Jong and Davidson (1998) show that Assumptions (1.4) and (1.5) follow from the following assumption:

Assumption 2. $n(\hat{\theta} - \theta_0) \xrightarrow{d} D$, and $v_t = (w_t', \varepsilon_t, \Delta z_t')$ is strictly stationary and strong mixing with strong mixing coefficients $\alpha(m)$ such that $\alpha(m) \leq Cm^{-r/(r-2)}$ for some C and some $r > 2$ such that

$$E|v_t|^r < \infty \tag{18}$$

and

$$\lim_{n \rightarrow \infty} E \left(n^{-1} \sum_{t=1}^n \sum_{s=1}^n v_t v_s' \right) = \Sigma. \tag{19}$$

The central result of this paper is the following:

Theorem 1. Under Assumption 1,

$$n^{1/2}(\hat{\beta} - \beta_0) \xrightarrow{d} M^{-1}N \int_0^1 G_Z(\xi)' d\xi D - M^{-1}W. \tag{20}$$

There are two particularly interesting special cases here. The first is the case where $N = 0$. For this case, the estimation of θ_0 has no consequences for the second-stage estimation and the usual asymptotic normality result holds. However, if N does not equal to zero, in general the first-stage estimation has consequences for the second-stage estimator. For the case of the Granger–Engle procedure with a linear least-squares second-stage estimation step, this intuition is correct because $N = 0$ for that case. However, the reason for the asymptotic negligibility is the fact that for that case, $N = 0$, rather than the superconsistency of $\hat{\theta}$. For the case where N does not equal zero, it is hard to calculate critical values because in general the limit distribution depends on D , W , and $\int_0^1 G_Z(\xi)' d\xi$, and all three asymptotically random quantities are not independent of each other. The second interesting case is the case where $\int_0^1 G_Z(\xi) d\xi = 0$ almost surely. For that case, asymptotic normality of our estimator is obtained. If a constant would have been included in the long-run regression, we could set $z_t = x_t - \bar{x}$, and now typically,

$$n^{-1/2}(x_{[\xi n]} - \bar{x}) \xrightarrow{d} G_X(\xi) - \int_0^1 G_X(\xi) d\xi = G_Z(\xi) \tag{21}$$

and $G_Z(\xi)$ integrates to 0 almost surely over $[0,1]$. Finally, note that $a_n = -\bar{\varepsilon}$ implies that

$$n^{-1/2} \sum_{t=1}^n (\partial/\partial\beta')f(w_t, \varepsilon_t + a_n, \beta_0) \tag{22}$$

is distributed asymptotically as

$$\begin{aligned} & n^{-1/2} \sum_{t=1}^n (\partial/\partial\beta')f(w_t, \varepsilon_t, \beta_0) - n^{-1/2} \sum_{t=1}^n \varepsilon_t E(\partial/\partial\beta')(\partial/\partial\varepsilon)f(w_t, \varepsilon_t, \beta_0) \\ &= (I: E(\partial/\partial\beta')(\partial/\partial\varepsilon)f(w_t, \varepsilon_t, \beta_0)) \begin{pmatrix} n^{-1/2} \sum_{t=1}^n (\partial/\partial\beta')f(w_t, \varepsilon_t, \beta_0) \\ - n^{-1/2} \sum_{t=1}^n \varepsilon_t \end{pmatrix} \end{aligned} \tag{23}$$

under the stated regularity conditions. Therefore, assuming that

$$\begin{pmatrix} n^{-1/2} \sum_{t=1}^n (\partial/\partial\beta')f(w_t, \varepsilon_t, \beta_0) \\ - n^{-1/2} \sum_{t=1}^n \varepsilon_t \end{pmatrix} \xrightarrow{d} N(0, J) \tag{24}$$

where

$$J = \begin{pmatrix} \Sigma_{11} & \Sigma'_{21} \\ \Sigma_{21} & \sigma_\varepsilon^2 \end{pmatrix}, \tag{25}$$

the asymptotic variance of W for the case $a_n = -\bar{\varepsilon}$ is $(I:N)J(I:N)'$. The estimation of N is straightforward. J can be consistently estimated using a Newey–West type estimator, as is demonstrated in for example Hansen (1992). The Newey–West estimator was introduced in econometrics by Newey and West (1987), and consistency results for this estimator are given in Gallant and White (1988), Andrews (1991), Pötscher and Prucha (1997), Andrews and Monahan (1992), and De Jong and Davidson (1997). Summarizing, we obtain the following theorem, that can provide a basis for inference:

Theorem 2. Assume that the preliminary estimator $\hat{\theta}$ is obtained by linear regression of y_t on x_t , and assume that a constant was included in that regression (where the estimator of the constant is no part of the $\hat{\theta}$ vector). Assume that Assumption 1 holds, except that Assumption (1.4) is replaced by

$$\left(n^{-3/2} \sum_{t=1}^n x_t, n(\hat{\theta} - \theta_0), n^{-1/2} \sum_{t=1}^n (\partial/\partial\beta')f(w_t, \varepsilon_t + a_n, \beta_0), n^{-1/2} \sum_{t=1}^n \varepsilon_t \right) \xrightarrow{d} \left(\int_0^1 G_X(\xi) d\xi, D, W, V \right), \tag{26}$$

where V is distributed $N(0, \sigma_\varepsilon^2)$ and $\sup_{1 \leq t \leq n} n^{-1/2}|x_t| = O_p(1)$. Then

$$n^{1/2}(\hat{\beta} - \beta_0) \xrightarrow{d} N(0, M^{-1}(I:N)J(I:N)'M^{-1}). \tag{27}$$

3. Examples

The standard Granger–Engle procedure deals with minimizing the function

$$\sum_{t=1}^n (\Delta y_t - \beta(y_{t-1} - \hat{\theta}'x_{t-1}))^2 \tag{28}$$

so here

$$f(w, \varepsilon, \beta) = (w - \beta\varepsilon)^2. \tag{29}$$

Differentiation with respect to β gives

$$(\partial/\partial\beta)f(w, \varepsilon, \beta) = -2(w - \beta\varepsilon)\varepsilon \tag{30}$$

and therefore

$$\begin{aligned} (\partial/\partial\varepsilon)(\partial/\partial\beta)f(\Delta y_t, \varepsilon_{t-1}, \beta) &= -2\Delta y_t + 4\beta\varepsilon_{t-1} \\ &= -2(\Delta y_t - \beta\varepsilon_{t-1}) + 2\beta\varepsilon_{t-1}. \end{aligned} \tag{31}$$

Because ε_t and $\Delta y_t - \beta \varepsilon_{t-1}$ both represent regression errors, it is natural that in this case, $N = 0$.

Next, suppose that we add a quadratic term to the specification, i.e. we consider

$$f(w, \varepsilon, \beta) = (w - \beta_1 \varepsilon - \beta_2 \varepsilon^2)^2. \tag{32}$$

Then we have

$$\begin{aligned} (\partial/\partial\varepsilon)(\partial/\partial\beta_2)f(w, \varepsilon, \beta) &= (\partial/\partial\varepsilon)(-2(w - \beta_1 \varepsilon - \beta_2 \varepsilon^2)\varepsilon^2) \\ &= -4\varepsilon w + 6\beta_1 \varepsilon^2 + 8\beta_2 \varepsilon^3, \end{aligned} \tag{33}$$

and noting that it is natural to assume that $E(\Delta y_t - \beta_{01} \varepsilon_{t-1} - \beta_{02} \varepsilon_{t-1}^2)\varepsilon_t = 0$, in general we have $N_2 = 2\beta_{01} E\varepsilon_t^2 + 4\beta_{02} E\varepsilon_t^3$. Therefore, adding the square of the estimated cointegrating relationship will in general result in a nonstandard limit distribution for the least-squares estimator. Also note that, interestingly, testing the null hypothesis $H_0: \beta_1 = \beta_2 = 0$ can be done using the standard asymptotic results, because under that null hypothesis, $N = 0$. In general, however, we conclude that adding the square of the estimated cointegrating relationship invalidates standard asymptotic results. Finally, consider the nonlinear least-squares estimation of the model

$$\Delta y_t = h(y_{t-1} - \theta'_0 x_{t-1}, \beta_0) + \eta_t. \tag{34}$$

For this model, it is easy to show that

$$\begin{aligned} N &= -2E\eta_t(\partial/\partial\beta')(\partial/\partial\varepsilon)h(\varepsilon_{t-1}, \beta_0) \\ &\quad + 2E(\partial/\partial\varepsilon)h(\varepsilon_{t-1}, \beta_0)(\partial/\partial\beta')h(\varepsilon_{t-1}, \beta_0). \end{aligned} \tag{35}$$

While we can naturally assume that the first part of the above expression equals zero, the second part is potentially nonzero. For nonlinear least-squares estimation, the verification that the second part is zero will allow the application of standard asymptotic techniques.

4. Conclusions

From the above discussion, we conclude the following. If no constant is included in the cointegrating regression, standard normal inference “as if θ_0 were known” will not be valid for the second-stage regression, unless we can verify that $N = 0$ in our application. If a constant is included in the regression however, we obtain the asymptotic normality result that we would have obtained if θ_0 were known. Note, however, that the limit distributions “as if θ_0 were known” are different depending whether or not a constant is included in

the long-run regression, unless $N = 0$. In other words, one should not include a constant in the long-run regression and use the short-run asymptotics that would result if $\varepsilon_t = y_t - \theta'_0 x_t$ were included in the short-run regression, unless $N = 0$.

Acknowledgements

I thank Jeff Wooldridge for suggesting this problem to me and for remarks on a first draft of this paper, and James Davidson for suggesting the analysis when a constant is included in the first-stage regression. Ingmar Prucha provided excellent comments on a later version of this paper. The comments of the referees are gratefully acknowledged.

Appendix A. Mathematical

In order to prove Theorem 1, we will need the following uniform law of large numbers:

Lemma A.1. Suppose that v_t is a sequence of strictly stationary random variables that is L_0 -approximable by an α -mixing process, and suppose that the v_t take their values in \mathbf{V} , where $\mathbf{V} \subseteq \mathbb{R}^q$. Assume that Θ is a compact subset of \mathbb{R}^k . Furthermore, let $q: \mathbf{V} \times \Theta \rightarrow \mathbb{R}$ be a function that is continuous in both arguments. Also, assume that for some $\phi > 0$,

$$E \sup_{\theta \in \Theta} |q(v_t, \theta)|^{1+\phi} < \infty. \quad (\text{A.1})$$

Then

$$\sup_{\theta \in \Theta} \left| n^{-1} \sum_{t=1}^n q(v_t, \theta) - E q(v_t, \theta) \right| \xrightarrow{P} 0. \quad (\text{A.2})$$

Proof. The proof follows from Theorem 5.2 of Pötscher and Prucha (1997). Their Assumption B is satisfied by our continuity requirement for $q(\cdot, \cdot)$; their Assumptions C and D follow from strict stationarity and the condition of Eq. (A.1). Assumption 5.2 is verified by appealing to Theorem 6.13 of Pötscher and Prucha (1997) \square .

Proof of Theorem 1. First note that by the mean value theorem for n large enough,

$$\begin{aligned}
 0 &= n^{-1/2} \sum_{t=1}^n (\partial/\partial\beta')f(w_t, \varepsilon_t + a_n + (\theta_0 - \hat{\theta})'z_t, \hat{\beta}) \\
 &= n^{-1/2} \sum_{t=1}^n (\partial/\partial\beta')f(w_t, \varepsilon_t + a_n, \beta_0) \\
 &\quad + n^{-1} \sum_{t=1}^n (\partial/\partial\beta)(\partial/\partial\beta)f(w_t, \varepsilon_t + a_n + (\theta_0 - \tilde{\theta})'z_t, \tilde{\beta})n^{1/2}(\hat{\beta} - \beta_0) \\
 &\quad - n^{-3/2} \sum_{t=1}^n (\partial/\partial\beta)(\partial/\partial\varepsilon)f(w_t, \varepsilon_t + a_n + (\theta_0 - \tilde{\theta})'z_t, \tilde{\beta})z_t'n(\hat{\theta} - \theta_0),
 \end{aligned}
 \tag{A.3}$$

where $\tilde{\theta}$ and $\tilde{\beta}$ denote mean values (following the usual abuse of notation, since the Taylor expansion is element-wise); note that $\tilde{\beta}$ is consistent and note that $\tilde{\theta}$ is order- n -consistent. Next, note that for some K_δ to be specified later, for $i, j = 1, \dots, m$,

$$\begin{aligned}
 &\left| n^{-1} \sum_{t=1}^n (\partial/\partial\beta_i)(\partial/\partial\beta_j)f(w_t, \varepsilon_t + a_n + (\theta_0 - \tilde{\theta})'z_t, \tilde{\beta}) \right. \\
 &\quad \left. - (\partial/\partial\beta_i)(\partial/\partial\beta_j)f(w_t, \varepsilon_t, \beta_0) \right| \\
 &\leq n^{-1} \sum_{t=1}^n \sup_{|\gamma| \leq \eta} |(\partial/\partial\beta_i)(\partial/\partial\beta_j)f(w_t, \varepsilon_t + \gamma, \tilde{\beta}) \\
 &\quad - (\partial/\partial\beta_i)(\partial/\partial\beta_j)f(w_t, \varepsilon_t, \tilde{\beta})| \\
 &\quad + n^{-1} \sum_{t=1}^n |(\partial/\partial\beta_i)(\partial/\partial\beta_j)f(w_t, \varepsilon_t, \tilde{\beta}) - (\partial/\partial\beta_i)(\partial/\partial\beta_j)f(w_t, \varepsilon_t, \beta_0)| \\
 &\quad + n^{-1} \sum_{t=1}^n |(\partial/\partial\beta_i)(\partial/\partial\beta_j)f(w_t, \varepsilon_t + a_n + (\theta_0 - \tilde{\theta})'z_t, \tilde{\beta}) \\
 &\quad - (\partial/\partial\beta_i)(\partial/\partial\beta_j)f(w_t, \varepsilon_t, \tilde{\beta})| I(|(\tilde{\theta} - \theta_0)'z_t + a_n| > \eta).
 \end{aligned}
 \tag{A.4}$$

Now by two applications of Lemma A.1, i.e. a law of large numbers uniformly in β and γ , note that the first term is asymptotically equivalent to or less than

$$E \sup_{|\gamma| \leq \eta} \sup_{\beta \in B} |(\partial/\partial\beta_i)(\partial/\partial\beta_j)f(w_t, \varepsilon_t + \gamma, \beta) - (\partial/\partial\beta_i)(\partial/\partial\beta_j)f(w_t, \varepsilon_t, \beta)| \tag{A.5}$$

for all $\eta > 0$, and can be made arbitrary small by choosing η small enough by uniform continuity, while the second term is asymptotically equivalent to

$$E[(\partial/\partial\beta_i)(\partial/\partial\beta_j)f(w_t, \varepsilon_t, \tilde{\beta}) - (\partial/\partial\beta_i)(\partial/\partial\beta_j)f(w_t, \varepsilon_t, \beta_0)]. \tag{A.6}$$

Now let n approach infinity and conclude that this second term converges in probability to zero. The third term is asymptotically small because the probability that term equals zero satisfies

$$\lim_{n \rightarrow \infty} P\left(\sup_{1 \leq t \leq n} |n^{1/2}(\tilde{\theta} - \theta_0)(n^{-1/2}z_t) + a_n| \leq \eta\right) = 1 \tag{A.7}$$

and because η was arbitrary, it therefore suffices to consider the limit distribution of

$$n^{-1} \sum_{t=1}^n (\partial/\partial\beta)(\partial/\partial\beta')f(w_t, \varepsilon_t, \beta_0). \tag{A.8}$$

Under the conditions stated, a law of large numbers holds for this expression, and therefore this expression converges to its expectation M . For the expression

$$n^{-3/2} \sum_{t=1}^n (\partial/\partial\beta')(\partial/\partial\varepsilon)f(w_t, \varepsilon_t + a_n + (\theta_0 - \tilde{\theta})'z_t, \tilde{\beta})z_t' \tag{A.9}$$

a similar argument holds to show that $\tilde{\beta}$ and $\tilde{\theta}$ can be replaced asymptotically by β_0 and θ_0 , respectively. This can be seen as follows. Note that

$$\begin{aligned} & \left| n^{-3/2} \sum_{t=1}^n (\partial/\partial\beta')(\partial/\partial\varepsilon)f(w_t, \varepsilon_t + a_n + (\theta_0 - \tilde{\theta})'z_t, \tilde{\beta})z_t' \right. \\ & \quad \left. - (\partial/\partial\beta')(\partial/\partial\varepsilon)f(w_t, \varepsilon_t, \beta_0)z_t' \right| \\ & \leq \sup_{1 \leq t \leq n} |n^{-1/2}z_t| n^{-1} \sum_{t=1}^n |(\partial/\partial\beta')(\partial/\partial\varepsilon)f(w_t, \varepsilon_t + a_n + (\theta_0 - \tilde{\theta})'z_t, \tilde{\beta}) \\ & \quad - (\partial/\partial\beta')(\partial/\partial\varepsilon)f(w_t, \varepsilon_t, \beta_0)|, \end{aligned} \tag{A.10}$$

and for the last term the same reasoning as before applies to show convergence to zero in probability. Therefore, it suffices to consider the limit distribution of

$$n^{-3/2} \sum_{t=1}^n (\partial/\partial\beta')(\partial/\partial\varepsilon)f(w_t, \varepsilon_t, \beta_0)z_t'. \tag{A.11}$$

Note that by rearranging

$$\begin{aligned} & n^{-3/2} \sum_{t=1}^n (\partial/\partial\beta')(\partial/\partial\varepsilon)f(w_t, \varepsilon_t, \beta_0)z_t' \\ &= Nn^{-3/2} \sum_{t=1}^n z_t' + n^{-3/2} \sum_{t=1}^n ((\partial/\partial\beta')(\partial/\partial\varepsilon)f(w_t, \varepsilon_t, \beta_0) - N)z_t'. \end{aligned} \quad (\text{A.12})$$

The last term is $o_p(1)$ by assumption. The first term by assumption converges in distribution to

$$N \int_0^1 G_Z(\xi)' d\xi. \quad (\text{A.13})$$

In conclusion,

$$n^{1/2}(\hat{\beta} - \beta_0) \xrightarrow{d} M^{-1}N \int_0^1 G_Z(\xi)' d\xi D - M^{-1}W. \quad \square \quad (\text{A.14})$$

References

- Andrews, D.W.K., 1991. Heteroscedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica* 59, 817–858.
- Andrews, D.W.K., Monahan, J.C., 1992. An improved heteroscedasticity and autocorrelation consistent covariance matrix estimator. *Econometrica* 60, 953–966.
- Chan, N.H., Wei, C.Z., 1988. Limiting distributions of least-squares estimates of unstable autoregressive processes. *Annals of Statistics* 16, 367–401.
- Davidson, J., 1994. *Stochastic Limit Theory*. Oxford University Press, Oxford.
- Davidson, J., Peel, D., 1998. A nonlinear error correction mechanism based on the bilinear model. *Economics Letters* 58, 165–170.
- De Jong, R.M., Davidson, J., 1997. Consistency of kernel estimators of heteroscedastic and autocorrelated covariance matrices. Working paper.
- De Jong, R.M., Davidson, J., 1998. The functional central limit theorem and weak convergence to stochastic integrals: results for weakly dependent and fractionally integrated processes. Working paper.
- Engle, R.F., 1987. On the theory of cointegrated time series. Invited paper presented at the Econometric Society European Meeting 1987, Copenhagen.
- Engle, R.F., Granger, C.W.J., 1987. Cointegration and error correction: representation, estimation, and testing. *Econometrica* 55, 251–276.
- Engle, R.F., Yoo, S.B., 1987. Forecasting and testing in cointegrated systems. *Journal of Econometrics* 35, 143–159.
- Ericsson, N.R., Hendry, D.F., Prestwich, K.M., 1998. The demand for broad money in the United Kingdom. 1878–1993. *Scandinavian Journal of Economics* 100, 289–324.
- Gallant, A.R., White, H., 1988. *A Unified Theory of Estimation and Inference for Nonlinear Dynamic Models*. Basil Blackwell, New York.
- Granger, C.W.J., 1981. Some properties of time series and their use in econometric model specification. *Journal of Econometrics* 16, 121–130.
- Granger, C.W.J., Terasvirta, T., 1993. *Modeling Nonlinear Economic Relations*. Oxford University Press, Oxford.

- Hansen, B.E., 1992. Consistent covariance matrix estimation for dependent heterogeneous processes. *Econometrica* 60, 967–972.
- Hendry, D.F., Ericsson, N.R., 1991. An econometric analysis of U.K. money demand in “Monetary Trends in the United States and the United Kingdom” by Milton Friedman and Anna J. Schwartz. *American Economic Review* 81, 8–38.
- Johansen, S., 1988. Statistical analysis of cointegrated vectors. *Journal of Economic Dynamics and Control* 12, 231–254.
- Johansen, S., 1991. Estimation and hypothesis testing of cointegrated vectors in Gaussian vector autoregressive models. *Econometrica* 59, 1551–1580.
- Newey, W.K., West, K.D., 1987. A simple, positive semi-definite, heteroscedasticity and autocorrelation consistent covariance matrix. *Econometrica* 55, 703–708.
- Phillips, P.C.B., Ouliaris, S., 1990. Asymptotic properties of residual based tests for cointegration. *Econometrica* 58, 165–193.
- Phillips, P.C.B., 1991. Optimal inference in cointegrated systems. *Econometrica* 59, 283–306.
- Pötscher, B.M., Prucha, I.R., 1997. *Dynamic nonlinear Econometric Models: Asymptotic Theory*. Springer, Berlin.