Nonlinear minimization estimators in the presence of cointegrating relations

Robert M. de Jong
Department of Economics
217 Marshall Hall
Michigan State University
East Lansing, MI 48824[*]

May 28, 2001

Abstract

In this paper, we consider estimation of a long-run and a short-run parameter jointly in the presence of nonlinearities. The theory developed establishes limit behavior of minimization estimators of the long-run and short-run parameters jointly. Typically, if the long-run parameter that is present in a cointegrating relationship is estimated, its estimator will be superconsistent. Therefore, we may conjecture that the joint minimization estimation of both parameters jointly will result in the same limit distribution for the short-run parameter as if the long-run parameter was known. However, we show that unless a regularity condition holds, this intuition is false in general. This regularity condition, that clearly holds in the standard linear case, is identical to the condition for validity of a two-step Granger-Engle type procedure. Also, it is shown that if the cointegrated variables are measured in deviation from their averages, the standard asymptotic normality result (that one would obtain if the long-run parameter was known) holds.

1 Introduction

In this paper, the properties of minimization estimators of parametric models that are non-linear in the cointegrating relationship will be considered. The concept of cointegration was 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 huge. 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), and $E\varepsilon_t = 0$. The 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,

$$\Delta y_t = \beta_0(y_{t-1} - \theta_0' x_{t-1}) + \eta_t. \tag{1}$$

Note that throughout this paper, (ε_t, η_t) is assumed to satisfy weak dependence properties. A different estimation procedure could be to estimate β_0 and θ_0 jointly, for example by nonlinear least squares. Also, we may want to estimate a nonlinear model instead of the linear model of Equation (1). Recently, some models have been proposed that are nonlinear in the cointegrating relationship; for example the Smooth Transition Error Correction Model (STECM) of Granger and Terasvirta (1993), and Davidson and Peel (1998) propose a model that is bilinear in the cointegrating relationship. Nobay and Peel (1997) also use a model that is bilinear in the cointegrating relationship, but they assume that the cointegrating vector is known. Granger and Lee (1989) suggest an asymmetric adjustment process for inventories, and a model that is nonlinear in the cointegrating vector is proposed in this article.

Clearly, for such models no equivalent of the Granger Representation Theorem exists, and there is no transparent overall model from which short-run and long-run dynamics follow. Therefore, all I can do in this paper is assume that the proposed nonlinear model implies that y_t is an I(1) process. For a practical situation, this leaves the researcher with the considerable burden of proving this before the results of this paper will apply. For example, for the nonlinear model

$$\Delta y_t = f(y_{t-1} - \theta_0' x_{t-1}, \beta_0) + \eta_t \tag{2}$$

where x_t is assumed to be I(1), the properties of y_t are determined directly by the choice of $f(.,.)$. Even the simple choice $f(\varepsilon, \beta) = \varepsilon$ will cause explosive, non-I(1) behavior of y_t. However, for nonlinear models such as the one above, it is still very well possible to obtain conditions that will ensure I(1) behavior of y_t. For example, it is a well-known result that for ε_t generated as $\varepsilon_t = g(\varepsilon_{t-1}) + \eta_t$, where η_t satisfies some weak dependence condition and $g(.)$ is a continuously differentiable function, ε_t has weak dependence properties whenever

$\sup_\varepsilon |(\partial/\partial\varepsilon)g(\varepsilon)| < 1$. See for example Pötscher and Prucha (1991) for results of this type. For the situation of Equation (2), which can be rewritten as

$$\varepsilon_t = f(\varepsilon_{t-1}, \beta) + \varepsilon_{t-1} - \theta_0' \Delta x_t + \eta_t, \tag{3}$$

this implies that weak dependence properties for ε_t can be shown as long as Δx_t and η_t have weak dependence properties and

$$\sup_\beta \sup_\varepsilon |(\partial/\partial\varepsilon)f(\varepsilon, \beta) - 1| < 1. \tag{4}$$

For the case $f(\varepsilon, \beta) = \beta\varepsilon$, i.e. the standard linear cointegration situation, an analysis such as the above will give the result that weak dependence properties of ε_t can be proven as long as $-2 < \beta < 0$, which is well-known from the literature on (linear) cointegration. But in spite of results such as above, in practical situations, this is a problem that needs to be analyzed separately, and is not necessarily straightforward.

One possible technique for estimation of such models is to use a Granger-Engle type technique where an estimator $\hat\theta$ of θ_0 is plugged in, and in the second stage, the short-run parameter β_0 is estimated. The asymptotic properties of that procedure were studied in de Jong (2001). The central question of this paper is the following: is it possible to give a general asymptotic theory for minimization estimators in the presence of a cointegrating relationship, when the short-run and long-run parameters are estimated jointly?

Saikkonen (1995) studied the problem of characterizing the asymptotic behavior of the joint maximum likelihood estimator, but a general framework for the study of this type of estimator has - to the best of the author's knowledge - not been attempted before.

It is straightforward to provide examples where the theory presented in this paper can be useful. For example, we may want to estimate Equation (1), but add a square of $y_{t-1} - \hat\theta' x_{t-1}$ as an extra regressor. We could now perform nonlinear least squares estimation, and use the t-statistic for testing the correctness of the original linear specification. The results of this paper show that in general the standard asymptotic theory for this model (treating $\hat\theta$ as if θ_0 was known) is invalid, thereby making the t-values of such a regression useless in general. A second example could be an ordered probit model for the demand for luxury cars in some time period. One may want to include the difference between consumption and long-run consumption in our regression. Assuming that consumption and income are cointegrated, this difference could be obtained as the error of the linear regression of consumption on income, which would result in a two-step procedure. We could also choose to minimize some criterion function with respect to both parameters jointly, and for such an analysis, the framework of this paper applies. Third, our analysis includes the STECM model proposed by Granger and Terasvirta (see Granger and Terasvirta (1993)). Their model is

$$\Delta y_t = \beta_{01} + \beta_{02} v_t + (\beta_{03} + \beta_{04} v_t) G(\varepsilon_{t-d}) + \eta_t \tag{5}$$

3

where $G(.)$ is some distribution function, possibly depending on parameters that will have to be estimated also; for example, the logistic distribution function

$$G(\varepsilon) = (1 + \exp(-\beta_5(\varepsilon - \beta_6)))^{-1} \tag{6}$$

where $\beta_5 \geq 0$. Note that testing for $\beta_{03} = \beta_{04} = 0$ adds an extra difficulty, because θ_0 will not be identified under the null hypothesis if both parameters are estimated jointly using, for example, nonlinear least squares.

In de Jong (2001) it was shown that for a model such as the STECM, standard asymptotically normal inference for the two-step procedure is invalid unless y_t and x_t are included in deviation from their average. Because the least squares estimator $b = (b_1, b_2')'$ of the linear regression of y_t on x_t and a constant satisfies

$$y_t - b_1 - b_2'x_t = y_t - \bar{y} - b_2'(x_t - \bar{x}), \tag{7}$$

this implies the somewhat counterintuitive conclusion that adding a constant to the long-run regression will give us the standard asymptotic normality, while excluding the constant from the long-run regression may invalidate that conclusion. If y_t and x_t are not included in deviation from their average, it was shown that an orthogonality condition needs to be met for the second stage estimator to be asymptotically normally distributed according to the standard theory (i.e., treating $\hat{\theta}$ as if it equaled θ_0).

This paper establishes a similar phenomenon for the full minimization estimator. In this paper, we show that if y_t and x_t are included in deviation from their average and we perform full minimization estimation with respect to both parameters, the short-run parameter is asymptotically normally distributed with the same distribution as if θ_0 was known. If y_t and x_t are not included in deviation from their average, the same orthogonality condition that was obtained for the two-step procedure in de Jong (2001) is also necessary to justify minimization estimation of $(\beta_0', \theta_0')'$ jointly using the standard asymptotic normality result for known θ_0.

Section 2 of this paper states the consistency result of this paper. In Section 3, we discuss the asymptotic distribution of the minimization estimator. Section 4 concerns covariance matrix estimation. Section 5 specializes our results to the case of the STECM model. This paper concludes with a Mathematical Appendix.

2 Consistency result

In this paper, we consider minimization estimators that equal

$$\text{argmin}_{(\beta,\theta)\in(B\times\Theta)} n^{-1} \sum_{t=1}^{n} f(w_t, \varepsilon_t + (\theta_0 - \theta)'z_t + a_n, \beta) \tag{8}$$

with probability one as $n \to \infty$, where w_t and $\varepsilon_t = y_t - \theta_0' x_t$ are stationary random variables and $B \times \Theta$ is the parameter space. We assume that z_t, x_t, and θ are elements of $\mathbb{R}^k$, and β is assumed to be an element of $\mathbb{R}^r$. Note that ε_t is unobserved here, and $a_n \xrightarrow{p} 0$ by assumption. Setting $z_t = x_t$ and $a_n = 0$ implies that $y_t - \theta' x_t$ is included, while $z_t = x_t - \bar{x}$ and $a_n = -\bar{\varepsilon}$ implies that $(y_t - \bar{y}) - \theta'(x_t - \bar{x})$ is included. Therefore, the analysis below is sufficiently general to include both cases. Typically, one would need to include Δy_t and/or Δx_t, or possibly lags of these, among the elements of w_t.

In this paper, the notation $\xrightarrow{d}$ and $\xrightarrow{p}$ denotes convergence in distribution and in probability, respectively. Let $\Rightarrow$ denote weak convergence with respect to the Skorokhod metric, as defined and discussed e.g. in Davidson (1994), Chapter 26-28.

The weak dependence concept that we will use is that of strong mixing. For the definition of strong (α-) and uniform (ϕ-) mixing random variables see e.g. Gallant and White (1988, p. 23) and Pötscher and Prucha (1991, p. 164).

The intuition behind the consistency proof of the next theorem is as follows. Suppose for the moment that $a_n = 0$ and $z_t = x_t$. Then we can rewrite our criterion function as

$$n^{-1} \sum_{t=1}^{n} f(w_t, \varepsilon_t + n^{1/2}(\theta_0 - \theta)'(n^{-1/2} x_t), \beta), \tag{9}$$

which, assuming that $n^{-1/2} x_{[\xi n]} \Rightarrow X(\xi)$ where $X(\xi)$ is some random element of $C^k[0, 1]$ (e.g. Brownian motion), suggests that by the law of large numbers, perhaps the criterion function behaves similar to

$$n^{-1} \sum_{t=1}^{n} E_{w_t, \varepsilon_t} f(w_t, \varepsilon_t + n^{1/2}(\theta_0 - \theta)'(n^{-1/2} x_t), \beta), \tag{10}$$

where the expectation E_{w_t, ε_t} denotes the expectation with respect to the measure of (w_t, ε_t) only. By continuity, we may conjecture that in some sense, the last expression resembles

$$\int_0^1 E_{w_t, \varepsilon_t} f(w_t, \varepsilon_t + n^{1/2}(\theta_0 - \theta)' X(\xi), \beta) d\xi. \tag{11}$$

Note that the last expression is random asymptotically, but minimized at $\theta = \theta_0$, $\beta = \beta_0$ by assumption. $|.|$ will denote the Euclidean norm in what follows.

The assumption that we need for the consistency proof of this paper is the following:

Assumption 1

1. *The parameter space B is compact.*

2. *$a_n \xrightarrow{p} 0$.*

3. $n^{1/2}(\hat{\theta} - \theta_0) = O_P(1)$.

4. For any normally distributed random vector X, $E_{w_t, \varepsilon_t} f(w_t, \varepsilon_t - \delta' X, \beta)$ is uniquely minimized at $(\beta', \delta')' = (\beta_0', 0')'$ with probability 1.

5. (ε_t, w_t) is a strictly stationary sequence of random variables and is an α-mixing process.

6. $f(w, a, \beta)$ is a function from $W \times A \times B$ to $\mathbb{R}$, and is continuous in all its arguments and for all compact sets A,

$$E \sup_{a \in A} \sup_{\beta \in B} |f(w_t, \varepsilon_t - a, \beta)| < \infty. \tag{12}$$

7. $n^{-1/2} z_{[\xi n]} \Rightarrow Z(\xi)$, where $Z(\xi)$ is a Gaussian random element of $C^k[0, 1]$ such that there does not exist a nonzero k-vector λ such that $\lambda' Z(\xi) = 0$ a.s., and

$$\limsup_{n \to \infty} n^{-1} \sum_{t=2}^{n} E|z_t - z_{t-1}|^{2+\delta} < \infty \tag{13}$$

for some $\delta > 0$.

Assumption 1.4 implies, for functions $f(., ., .)$ that are differentiable with respect to both β and ε, that under regularity conditions

$$E(\partial/\partial\beta) f(w_t, \varepsilon_t, \beta_0) = 0 \tag{14}$$

and

$$E(\partial/\partial\varepsilon) f(w_t, \varepsilon_t, \beta_0) = 0, \tag{15}$$

and for the derivation of the limit distribution of $(\hat{\beta}, \hat{\theta})$, those two properties will be crucial. From Assumption 1 only, however, the following consistency result follows:

Theorem 1 *Under Assumption 1, $(\hat{\beta}', n^{1/2}(\hat{\theta} - \theta_0)')' \xrightarrow{p} (\beta_0', 0')'$.*

Note that in Theorem 1, $n^{1/2}(\hat{\theta} - \theta_0) = O_P(1)$ is assumed rather than derived. Typically n-consistent estimators of θ_0 exist in this framework, and the assumed $n^{1/2}$ rate is lower than this. The mathematical problem of relaxing this assumption is similar to the problem of establishing consistency for minimization estimators that are defined as minimizing over a parameter space such as $\mathbb{R}^k$ instead of some compact set, and seems hardly avoidable because of the scaling of z_t that has to take place. Therefore, this property has to be established on

6

an *ad hoc* basis for each estimator that is considered. One generic solution is to assume that we have a preliminary estimator $\tilde{\theta}$ that satisfies $n(\tilde{\theta} - \theta_0) = O_P(1)$. Then we could define $\hat{\theta}$ as minimizing the criterion function over $\{\theta \in \Theta : |\theta - \tilde{\theta}| \leq Cn^{-1/2}\}$. For this estimator, the reasoning leading up to Theorem 1 is easily copied, but Assumption 2 has become trivial. Other than this solution, we may be able to draw on the repertoire of techniques that exists in the literature to restrict attention to a compact parameter space. See for example Pötscher and Prucha (1991) for a discussion of such techniques.

3   Asymptotic distribution

In this section, we will derive the limit distribution of the estimator analyzed earlier. Define

$$B = (\sum_{t=1}^{n}(\partial/\partial\beta)f(w_t, \varepsilon_t + a_n, \beta_0), -\sum_{t=1}^{n}z_t'(\partial/\partial\varepsilon)f(w_t, \varepsilon_t + a_n, \beta_0))' \equiv (B_1', B_2')', \quad (16)$$

where the differentiation with respect to ε is with respect to the second argument of $f(., ., .)$, and

$$A(\beta, \theta) = \begin{pmatrix} A_{11}(\beta, \theta) & A_{12}(\beta, \theta) \\ A_{21}(\beta, \theta) & A_{22}(\beta, \theta) \end{pmatrix} \tag{17}$$

where

$$A_{11}(\beta, \theta) = \sum_{t=1}^{n}(\partial/\partial\beta)(\partial/\partial\beta')f(w_t, \varepsilon_t + (\theta_0 - \theta)'z_t + a_n, \beta) \tag{18}$$

$$A_{21}(\beta, \theta) = A_{12}(\beta, \theta)' = -\sum_{t=1}^{n}z_t(\partial/\partial\beta)(\partial/\partial\varepsilon)f(w_t, \varepsilon_t + (\theta_0 - \theta)'z_t + a_n, \beta) \tag{19}$$

and

$$A_{22}(\beta, \theta) = \sum_{t=1}^{n}z_t z_t'(\partial^2/\partial\varepsilon^2)f(w_t, \varepsilon_t + (\theta_0 - \theta)'z_t + a_n, \beta). \tag{20}$$

Also, define $A_{ij} = A_{ij}(\beta_0, \theta_0)$ for $i, j = 1, 2$, and

$$L = E(\partial^2/\partial\varepsilon^2)f(w_t, \varepsilon_t, \beta_0) \tag{21}$$

and

$$M = E(\partial/\partial\beta')(\partial/\partial\varepsilon)f(w_t, \varepsilon_t, \beta_0). \tag{22}$$

In order to prove the central result of this section, we need the following assumption:

Assumption 2

1. $$((\hat{\beta} - \beta_0)', n^{1/2}(\hat{\theta} - \theta_0)')' \xrightarrow{p} 0, \tag{23}$$

    where β_0 and θ_0 are in the interiors of the parameter spaces B and Θ.

2. Assumptions (1.4), (1.5), and (1.7) hold.

3. $(\partial/\partial\beta)f(w, \varepsilon, \beta)$, $(\partial/\partial\beta)(\partial/\partial\beta')f(w, \varepsilon, \beta)$ and $(\partial/\partial\beta)(\partial/\partial\varepsilon)f(w, \varepsilon, \beta)$ are continuous on $W \times A \times B$, and for some open neighborhood Γ of 0 and for some $\phi > 0$ and for $j = 1, \dots, k$,

    $$E \sup_{|\gamma| \in \Gamma} \sup_{\beta \in B} |(\partial/\partial\beta')(\partial/\partial\beta_j)f(w_t, \varepsilon_t + \gamma, \beta)|^{1+\phi} < \infty, \tag{24}$$

    $$E \sup_{|\gamma| \in \Gamma} \sup_{\beta \in B} |(\partial/\partial\beta')(\partial/\partial\varepsilon)f(w_t, \varepsilon_t + \gamma, \beta)|^{1+\phi} < \infty. \tag{25}$$

    and

    $$E \sup_{|\gamma| \in \Gamma} \sup_{\beta \in B} |(\partial^2/\partial\varepsilon^2)f(w_t, \varepsilon_t + \gamma, \beta)|^{1+\phi} < \infty. \tag{26}$$

4. $v_t = (w_t', \varepsilon_t, \Delta z_t')'$ is strong mixing with strong mixing coefficients $\alpha(m)$ such that $\alpha(m) \le Cm^{-r/(r-2)}$ for some C and some $r > 2$ such that

    $$E|\Delta z_t|^r < \infty, \tag{27}$$

    $$E|(\partial/\partial\beta)f(w_t, \varepsilon_t, \beta_0)|^r < \infty, \tag{28}$$

    $$E|(\partial/\partial\varepsilon)f(w_t, \varepsilon_t, \beta_0)|^r < \infty, \tag{29}$$

    $$E|(\partial/\partial\varepsilon)(\partial/\partial\beta')f(w_t, \varepsilon_t, \beta_0)|^r < \infty, \tag{30}$$

5. $$n^{-1/2}z_{[\xi n]} \Rightarrow Z(\xi), \tag{31}$$

    where $Z(\xi)$ is a Gaussian random element of $C^k[0, 1]$.

6. $$\left(\begin{pmatrix} n^{-1}A_{11}(\beta_0, \theta_0) & n^{-3/2}A_{12}(\beta_0, \theta_0) \\ n^{-3/2}A_{21}(\beta_0, \theta_0) & n^{-2}A_{22}(\beta_0, \theta_0) \end{pmatrix}, \begin{pmatrix} n^{-1/2}B_1 \\ n^{-1}B_2 \end{pmatrix} \right) \xrightarrow{d} (\tilde{A}, \tilde{B}), \tag{32}$$

    where $(\tilde{A}, \tilde{B})$ is defined as follows:

8

$$(a) \quad n^{-1/2} B_1 \xrightarrow{d} \tilde{B}_1, \tag{33}$$

$$(b) \quad n^{-1} B_2 \xrightarrow{d} \tilde{B}_2, \tag{34}$$

$$(c) \quad n^{-1} A_{11} \xrightarrow{p} E(\partial/\partial\beta)(\partial/\partial\beta') f(w_t, \varepsilon_t, \beta_0) = \tilde{A}_{11}, \tag{35}$$

$$(d) \quad n^{-2} A_{22} \xrightarrow{d} L \int_0^1 Z(\xi) Z(\xi)' d\xi = \tilde{A}_{22}, \tag{36}$$

$$(e) \quad n^{-3/2} A_{21} \xrightarrow{d} - \int_0^1 Z(\xi) d\xi M' = \tilde{A}_{21}. \tag{37}$$

(f) $\tilde{A}$ is invertible with probability 1.

It is straightforward to impose weak dependence conditions ensuring that the above assumptions hold. Such results are by now standard tools in time series literature. In this paper, I chose to impose the above high-level assumption and note that using e.g. the results from Davidson (1994), it is relatively straightforward to list weak dependence conditions ensuring that the above properties hold.

The assumption of Equation (23) can be verified by applying Theorem 1. Equation (31) assumes that a functional central limit theorem holds for $n^{-1/2} z_{[n\xi]}$. B_1 is assumed to satisfy a central limit theorem-type result in Equation (33). Note that for the choice $a_n = -\bar{\varepsilon}$,

$$n^{-1/2} \sum_{t=1}^n (\partial/\partial\beta') f(w_t, \varepsilon_t - \bar{\varepsilon}, \beta_0) \tag{38}$$

will be asymptotically equivalent to

$$n^{-1/2} \sum_{t=1}^n (\partial/\partial\beta') f(w_t, \varepsilon_t, \beta_0) - n^{-1/2} \sum_{t=1}^n \varepsilon_t E(\partial/\partial\beta')(\partial/\partial\varepsilon) f(w_t, \varepsilon_t, \beta_0) \tag{39}$$

under the stated regularity conditions. We can rewrite this as

$$(I : E(\partial/\partial\beta')(\partial/\partial\varepsilon) f(w_t, \varepsilon_t, \beta_0))(n^{-1/2} \sum_{t=1}^n (\partial/\partial\beta) f(w_t, \varepsilon_t, \beta_0), -n^{-1/2} \sum_{t=1}^n \varepsilon_t)', \tag{40}$$

and therefore in general the asymptotic variance of B_1 will be

$$V = (I : M) \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} (I : M)', \tag{41}$$

where

$$\Sigma_{11} = \lim_{n \to \infty} E(n^{-1/2} \sum_{t=1}^{n} (\partial/\partial\beta') f(w_t, \varepsilon_t, \beta_0))(n^{-1/2} \sum_{t=1}^{n} (\partial/\partial\beta) f(w_t, \varepsilon_t, \beta_0)), \tag{42}$$

$$\Sigma_{21} = \Sigma_{12}' = -\lim_{n \to \infty} E(n^{-1/2} \sum_{t=1}^{n} (\partial/\partial\beta) f(w_t, \varepsilon_t, \beta_0))(n^{-1/2} \sum_{t=1}^{n} \varepsilon_t) \tag{43}$$

and

$$\Sigma_{22} = \lim_{n \to \infty} E(n^{-1/2} \sum_{t=1}^{n} \varepsilon_t)^2 \tag{44}$$

where we assume existence of Σ_{11}, Σ_{21}, and Σ_{22}. Note that from the results in de Jong (2001), it follows that under Assumption 1 and 2, $\hat{M} \xrightarrow{p} M$ where

$$\hat{M} = n^{-1} \sum_{t=1}^{n} (\partial/\partial\beta')(\partial/\partial\varepsilon) f(w_t, y_t - \hat{\theta}' x_t, \hat{\beta}). \tag{45}$$

For B_2 a "convergence to stochastic integrals" result (see e.g. Davidson (1994), Chapter 30) typically holds (using the fact that $E(\partial/\partial\varepsilon) f(w_t, \varepsilon_t, \beta_0) = 0$ by assumption). For $z_t = x_t - \bar{x}$, a "convergence to stochastic integrals" type result for B_2 will typically hold as well. $n^{-1} A_{11}$ will satisfy a weak law of large numbers under the stated conditions, and $n^{-3/2} A_{21}$ can be rewritten as

$$-n^{-3/2} \sum_{t=1}^{n} z_t((\partial/\partial\beta)(\partial/\partial\varepsilon) f(w_t, \varepsilon_t + a_n, \beta_0) - M') - n^{-3/2} \sum_{t=1}^{n} z_t M' \tag{46}$$

and typically the first term will be $O_P(n^{-1/2})$; such a result can be obtained, for example, by assuming that a "convergence to stochastic integrals" result holds for this first term. A similar argument can be made for A_{22}.

Using Theorem 1 and Assumptions 1 and 2, the following result can be obtained:

Theorem 2 *Under Assumption 1 and 2,*

$$(n^{1/2}(\hat{\beta} - \beta_0)', n(\hat{\theta} - \theta_0)')' \xrightarrow{d} \tilde{A}^{-1}\tilde{B}. \tag{47}$$

The above theorem implies that in general the limit distribution of the short-run parameter is not normal in general unless $M = 0$ or $\int_0^1 Z(\xi)d\xi = 0$ almost surely. This is because under either of these two conditions, $\tilde{A}_{21} = 0$, implying that

$$n^{1/2}(\hat{\beta} - \beta_0) \xrightarrow{d} \tilde{A}_{11}^{-1}\tilde{B}_1, \tag{48}$$

and therefore the same limit distribution is obtained as if θ_0 was known. The property $\int_0^1 Z(\xi)d\xi = 0$ occurs if $y_t - \bar{y} - \theta'(x_t - \bar{x})$ was included in the original regression. Setting $z_t = x_t - \bar{x}$, and assuming that $n^{-1/2}x_{[\xi n]} \Rightarrow X(\xi)$ where $X(\xi)$ denotes some random element of $C^k[0,1]$, we have

$$n^{-1/2}z_{[\xi n]} \Rightarrow X(\xi) - \int_0^1 X(\xi)d\xi \tag{49}$$

which clearly integrates to zero almost surely over $\xi \in [0,1]$. Therefore, if we include y_t and x_t in deviation from their average, the "usual" asymptotic normality result (that we would have if θ_0 was known) for the short-run parameter estimator $\hat{\beta}$ holds, and we may want to interpret the diagonal structure of the $\tilde{A}$ matrix as indicating that the short- and long-run parameter estimators have become somewhat disconnected now. The possibility of obtaining this type of result was suggested to me by James Davidson after reading a first draft of this paper. In general, however, if y_t and x_t are not included in deviation from their average, the estimation of the long-run parameter seems to affect the limit distribution of the short-run parameter unless $M = 0$.

One of the referees pointed out that the result that including y_t and x_t in deviation from their averages renders "standard" asymptotic theory may be less surprising when a parallel is drawn with results such as those in Sims, Stock, and Watson (1990) in a slightly different setting. In the linear model

$$y_t = \beta_1 x_t + \beta_2 w_t + \eta_t$$

where x_t is I(1) and w_t and η_t are I(0), where η_t is a regression error satisfying $E\eta_t = 0$, the least squares estimator for β_2 fails to have the usual limit distribution if $Ew_t \neq 0$. When a constant is added to the estimated regression however, in spite of the fact that the "true" value for that constant is 0, the usual limit distribution is obtained, as is pointed out in Sims, Stock and Watson (1990). Also, previous authors have pointed out the benefits of including a "redundant" regressor in regressions such as the ones considered here; see for example Johansen (1994).

One leading case where $M = 0$ is obtained is the standard Granger-Engle procedure. If we consider

$$f(w, \varepsilon, \beta) = (w - \beta\varepsilon)^2 \tag{50}$$

11

where our model is

$$\Delta y_t = \beta_0(y_{t-1} - \theta_0' x_{t-1}) + \eta_t, \tag{51}$$

then

$$M = E(\partial/\partial\beta)(\partial/\partial\varepsilon)(w_t - \beta\varepsilon_t)^2 \mid_{\beta=\beta_0} = -2E\eta_t + 2\beta_0 E\varepsilon_t = 0 \tag{52}$$

in general. Therefore, the fact that the long-run parameter is estimated does not affect the distribution of the short-run parameter here. However, suppose we add a square to our specification, i.e. our model becomes

$$\Delta y_t = \beta_{01}(y_{t-1} - \theta_0 x_{t-1}) + \beta_{02}(y_{t-1} - \theta_0 x_{t-1})^2 + \eta_t \tag{53}$$

and

$$f(w, \varepsilon, \beta_1, \beta_2) = (w - \beta_1\varepsilon - \beta_2\varepsilon^2)^2. \tag{54}$$

We could do this for purposes of model specification. Note that the only choice of β_{02} that makes y_t I(1) will be $\beta_{02} = 0$, because for other choices of β_{02} the implied dynamics will generally be explosive. If our desire is to test the hypothesis $\beta_{02} = 0$, then in general, $M_2 = 2\beta_{01}E\varepsilon_t^2 + 4\beta_{02}E\varepsilon_t^3 + 4E\varepsilon_t\eta_t \neq 0$, so using the regular t-statistic for testing for $\beta_{02} = 0$ will in general be an incorrect procedure unless $\beta_{01} = 0$. Note that, interestingly, if $\beta_{01} = \beta_{02} = 0$ we will have $M = 0$ if $E\varepsilon_t\eta_t = 0$, but this does not imply that the usual chi-square test of the hypothesis $\beta_{01} = \beta_{02} = 0$ can be used here, because - as noted earlier - θ_0 will be unidentified under that hypothesis.

4 Covariance matrix estimation

In order to have a complete asymptotic theory for the estimation of models that are nonlinear in the cointegrating relationship, the problem of the estimation of V of Equation (41) remains. From the results in de Jong (2001), it follows that under Assumption 1 and 2, $\hat{M} \xrightarrow{p} M$. However, for the estimation of Σ_{11}, Σ_{21}, Σ_{12} and Σ_{22}, typically we will need heteroskedasticity and autocorrelation consistent covariance matrix estimators. Consistency proofs for these estimators are well-known; see for example White (1984), Newey and West (1987), Gallant and White (1988), Andrews (1991), Hansen (1992), and de Jong and Davidson (2000) for weak consistency proofs, and de Jong (2000) for a strong consistency proof. However, the proofs in these references are not directly applicable to the present situation because both a short-run and a long-run parameter are present. In the case where

only a root-n consistent estimator is present, consistency will follow from one of the above-mentioned references. Similarly to Hansen (1992), define $\hat{\Omega} = \sum_{l=-n+1}^{n-1} k(l/\gamma_n)\hat{\Gamma}(l)$ where $\hat{\Gamma}(l) = n^{-1}\sum_{t=1}^{n-l} g(w_t, y_t - \hat{\theta}'x_t, \hat{\beta})g(w_{t+l}, y_{t+l} - \hat{\theta}'x_{t+l}, \hat{\beta})'$ for $l \geq 0$ and $\hat{\Gamma}(l) = \hat{\Gamma}(-l)'$ for $l < 0$, where $g(w_t, y_t - \hat{\theta}'x_t, \hat{\beta}) = (\partial/\partial\beta')f(w_t, y_t - \hat{\theta}'x_t, \hat{\beta})$ if $a_n = 0$ and $g(w_t, y_t - \hat{\theta}'x_t, \hat{\beta}) = ((\partial/\partial\beta)f(w_t, y_t - \hat{\theta}'x_t, \hat{\beta}), \varepsilon_t)'$ if $a_n = -\bar{\varepsilon}$. Define $\tilde{\Omega}$ similarly to $\hat{\Omega}$ but using $g(w_t, \varepsilon_t, \hat{\beta})$ instead of $g(w_t, y_t - \hat{\theta}'x_t, \hat{\beta})$. Then note that for $\tilde{\Omega}$, standard results will hold because it only depends on a root-n consistent estimator but no longer on the long-run estimator $\hat{\theta}$. Therefore, we seek to establish asymptotic equivalence between $\hat{\Omega}$ and $\tilde{\Omega}$ and refer to the above references for exact listings of regularity conditions for the consistency of $\tilde{\Omega}$. The following theorem establishes such a result:

Theorem 3 *Assume that Assumption 1 and 2 hold. Also, assume that $\gamma_n n^{-1/2} = o(1)$ and assume that $k(.)$ is continuous at all but a finite number of points and satisfies $\int_{-\infty}^{\infty} |k(x)|dx < \infty$. In addition, assume that for some open neighborhood Γ of 0,*

$$E \sup_{|\gamma|\in\Gamma} \sup_{\beta\in B} |(\partial/\partial\varepsilon)g(w_t, \varepsilon_t + \gamma, \beta)|^2 < \infty \tag{55}$$

and

$$E \sup_{|\gamma|\in\Gamma} \sup_{\beta\in B} |g(w_t, \varepsilon_t + \gamma, \beta)|^2 < \infty, \tag{56}$$

Then $\tilde{\Omega} - \hat{\Omega} \xrightarrow{p} 0$.

5 The STECM model

The theory as set out in the earlier sections of this paper can be applied to derive the limit distribution of the nonlinear least squares estimator of the short- and long-run parameters in the STECM model of Equation (5). We assume that $G(\varepsilon) = R(\beta_5(\varepsilon - \beta_6))$, for some distribution function $R(.)$. We noted before that this asymptotic theory cannot be used for constructing an chi-square test for the null hypothesis H_0: $\beta_{03} = \beta_{04} = 0$ because of the identification problem. Also, we need to bound the parameter space B for the β_5 parameter away from zero (e.g. assume that the parameter space for β_5 is of the form $[\beta_{5L}, \beta_{5U}]$ where $0 < \beta_{5L} < \beta_{5U}$), because for $\beta_5 = 0$, an additional identification problem arises. For this model, we will directly show that

$$n^{1/2}(\hat{\theta} - \theta_0) = o_P(1) \tag{57}$$

using an argument similar to the consistency argument of Theorem 1. We will show this property using a compactification argument. The criterion function now is

$$n^{-1} \sum_{t=1}^{n} f(w_t, \varepsilon_t + (\theta_0 - \theta)' z_t + a_n, \beta)$$

$$= n^{-1} \sum_{t=1}^{n} (\Delta y_t - (\beta_1 + \beta_2 w_t) - (\beta_3 + \beta_4 w_t) R(\varepsilon_t + (\theta_0 - \theta)' z_t + a_n))^2, \tag{58}$$

and note that this function is bounded in θ. For this model, it is possible to show that under the stated regularity conditions,

$$\sup_{\beta \in B} \sup_{\delta \in \mathbb{R}^k} \sup_h \left| n^{-1} \sum_{t=1}^{n} f(w_t, \varepsilon_t + \delta' h(t/n), \beta) - E f(w_t, \varepsilon_t + \delta' h(t/n), \beta) \right| \xrightarrow{p} 0, \tag{59}$$

where the "sup" over h is as in the proof of Theorem 1. Therefore, the consistency argument given in the proof of Theorem 1 directly applies here, establishing that $n^{1/2}(\hat{\theta} - \theta_0) \xrightarrow{p} 0$ and $\hat{\beta} \xrightarrow{p} \beta_0$. To show the above law of large numbers, note that the assertion is equivalent to

$$\sup_{\beta \in B} \sup_{\alpha \in [0,1]^k} \sup_h \left| n^{-1} \sum_{t=1}^{n} f(w_t, \varepsilon_t + P(\alpha)' h(t/n), \beta) - E f(w_t, \varepsilon_t + P(\alpha)' h(t/n), \beta) \right| \xrightarrow{p} 0, (60)$$

where $P(\alpha) = (\Phi^{-1}(\alpha_1), \Phi^{-1}(\alpha_2), \ldots, \Phi^{-1}(\alpha_k))'$ and $\Phi(.)$ denotes the normal distribution function. Rewriting the expression in this way has the advantage that now the parameter space is compact, and therefore the uniform law of large numbers can be proven analogously to the proof of Theorem 1.

Summarizing, we obtain the following consistency result for the STECM nonlinear least squares estimator:

Assumption 3 $(\hat{\beta}', \hat{\theta}')'$ *is obtained by minimizing the expression of Equation (58) over* $B \times \mathbb{R}^k$, *where either* $z_t = x_t$ *and* $a_n = 0$, *or* $z_t = x_t - \bar{x}$ *and* $a_n = -\bar{\varepsilon}$. *In addition,*

1. *The parameter space* B *is compact, while the parameter space* Θ *is* $\mathbb{R}^k$.

2. $R(.)$ *is a continuous and bounded function.*

3. $E(\Delta y_t - (\beta_1 + \beta_2 w_t) - (\beta_3 + \beta_4 w_t) R(\beta_5 (\varepsilon_t - \phi)))^2$ *is minimized at* $(\beta', \phi')' = (\beta_0', 0)'$, *and this minimum exceeds* $\lim_{\phi \to \infty} E(\Delta y_t - (\beta_{01} + \beta_{02} w_t) - (\beta_{03} + \beta_{04} w_t) R(\beta_{05} (\varepsilon_t - \phi)))^2$ *and* $\lim_{\phi \to -\infty} E(\Delta y_t - (\beta_{01} + \beta_{02} w_t) - (\beta_{03} + \beta_{04} w_t) R(\beta_{05} (\varepsilon_t - \phi)))^2$.

14

4. (ε_t, w_t) is a strictly stationary sequence of random variables that is L_2-near epoch dependent on v_t, where v_t is an α-mixing sequence.

5. $E(\Delta y_t)^2 < \infty$ and $Ew_t^2 < \infty$, and $E|\varepsilon_t|^{2+\phi} < \infty$ for some $\phi > 0$.

6. $n^{-1/2}x_{[\xi n]} \Rightarrow X(\xi)$, where $X(\xi)$ is a random element of $C^k[0,1]$, and

$$\limsup_{n\to\infty} n^{-1} \sum_{t=1}^{n} E|x_t - x_{t-1}|^{2+\delta} < \infty \tag{61}$$

for some $\delta > 0$.

The above assumption leads to the following theorem:

Theorem 4 *Under Assumption 3,* $(\hat{\beta}', n^{1/2}(\hat{\theta} - \theta_0)')' \xrightarrow{p} (\beta_0', 0')'$.

We can also specialize Theorem 2 to case of the STECM model:

Assumption 4

1. Assumption 3 holds, and β_0 and θ_0 are in the interiors of the parameter spaces B and Θ.

2. $r(a) = (\partial/\partial\varepsilon)R(\varepsilon)$ and $r'(a) = (\partial^2/\partial\varepsilon^2)R(\varepsilon)$ are bounded.

3. For some $\phi > 0$, $E|w_t|^{2+\phi} < \infty$.

4. Assumptions 2.5 and 2.6 hold.

Using the above assumption, the following result is now established:

Theorem 5 *Under Assumption 3 and 4,*

$$(n^{1/2}(\hat{\beta} - \beta_0)', n(\hat{\theta} - \theta_0)')' \xrightarrow{d} \tilde{A}^{-1}\tilde{B}. \tag{62}$$

For the above STECM model, from elementary calculations it can be shown that

$$M = E \begin{pmatrix} 2(\beta_{03} + \beta_{04}w_t)r_t\beta_{05} \\ 2(\beta_{03} + \beta_{04}w_t)w_tr_t\beta_{05} \\ 2\beta_{05}(\beta_{03} + \beta_{04}w_t)r_tL_t - 2\eta_t\beta_{05}r_t \\ 2\beta_{05}(\beta_{03} + \beta_{04}w_t)w_tr_tL_t - 2\eta_tw_t\beta_{05}r_t \\ 2(\beta_{03} + \beta_{04}w_t)^2r_t^2(\varepsilon_t - \beta_{06}) - 2\eta_t(\beta_{03} + \beta_{04}w_t)r_t'(\varepsilon_t - \beta_{06}) \\ -2(\beta_{03} + \beta_{04}w_t)^2r_t^2\beta_{05} + 2\eta_t(\beta_{03} + \beta_{04}w_t)r_t'\beta_{05} \end{pmatrix}, \tag{63}$$

15

where $R_t = R(\beta_{05}(\varepsilon_t - \beta_{06}))$, $r_t = r(\beta_{05}(\varepsilon_t - \beta_{06}))$ and $r'_t = r'(\beta_{05}(\varepsilon_t - \beta_{06}))$. Clearly, the above result illustrates that $M \neq 0$ in general for the STECM model, implying that in general, if x_t and y_t are not included in deviation from their average, t-values and F-tests are invalid. Also note again that we can not test the null H_0: $\beta_{03} = \beta_{04} = 0$ using asymptotic results "as if θ_0 was known", even if $M = 0$ for that null hypothesis, because θ_0 will be unidentified under that null hypothesis.

Proofs

The lemma below is reminiscent of Theorem 2.7 of Kim and Pollard (1990); because we defined weak convergence in the "traditional" way (using the Skorokhod topology), we need to prove a result similar to Kim and Pollard's, but assuming that weak convergence holds in the "traditional" way. Let $(\Omega, \mathcal{F}, P)$ denote the probability space.

Lemma 1 *Assume that $Q_n(\omega, \alpha) \Rightarrow Q(\omega, \alpha)$ for $\alpha \in A$, and assume that $Q_n(\omega, \alpha)$ and $Q(\omega, \alpha)$ are almost surely continuous on A, where A is a compact subset of $\mathbb{R}^k$. Assume that $Q(\omega, \alpha)$ is uniquely minimized with probability 1 at $\alpha = \alpha_0$. Then*

$$\mathrm{argmin}_{\alpha \in A} Q_n(\omega, \alpha) \xrightarrow{d} \alpha_0. \tag{64}$$

Proof of Lemma 1:

By the Skorokhod Representation theorem (see e.g. Davidson (1994), theorem 26.25), there exists a sequence $Q^n(\omega, \alpha)$ of elements of $C[A]$ that are distributed identically to $Q_n(\omega, \alpha)$ such that

$$\sup_{\alpha \in A} |Q^n(\omega, \alpha) - Q(\omega, \alpha)| \xrightarrow{as} 0. \tag{65}$$

Define $\hat{\alpha} = \mathrm{argmin}_{\alpha \in A} Q_n(\omega, \alpha)$ and $\tilde{\alpha} = \mathrm{argmin}_{\alpha \in A} Q^n(\omega, \alpha)$. Note that $\tilde{\alpha}$ and $\hat{\alpha}$ have the same distribution, and therefore for all $\eta > 0$,

$$P(|\hat{\alpha} - \alpha_0| > \eta) = P(|\tilde{\alpha} - \alpha_0| > \eta), \tag{66}$$

and therefore showing consistency of $\tilde{\alpha}$ is sufficient to show consistency of $\hat{\alpha}$. Next, note that

$$0 \leq Q(\omega, \tilde{\alpha}) - Q(\omega, \alpha_0)$$
$$\leq 2 \sup_{\alpha \in A} |Q^n(\omega, \alpha) - Q(\omega, \alpha)| \xrightarrow{as} 0, \tag{67}$$

and therefore by uniqueness of α_0, the result follows.

Proof of Theorem 1:

Define $\hat{\delta} = n^{1/2}(\hat{\theta} - \theta_0)$ and $\delta = n^{1/2}(\theta - \theta_0)$. We will show that $(\hat{\beta}', \hat{\delta}')' \xrightarrow{p} (\beta_0', 0')'$ under the conditions of the theorem. Define $\alpha = (\beta', \delta')'$, and define α_0 and $\hat{\alpha}$ analogously. Let

$$H(\beta, \gamma, \delta) = Ef(w_t, \varepsilon_t - \delta'\gamma, \beta), \tag{68}$$

and note that $H(.,.,.)$ is continuous in all its arguments by the dominated convergence theorem and Assumption 1. Our proof is based on the observation that

$$Q_n(\omega, \beta, \delta) = n^{-1} \sum_{t=1}^{n} f(w_t, \varepsilon_t + a_n - \delta'n^{-1/2}z_t, \beta) \tag{69}$$

converges weakly (as a function of (β, δ)) to

$$Q(\omega, \beta, \delta) = \int_0^1 H(\beta, Z(\xi), \delta)d\xi, \tag{70}$$

and the last criterion function is minimal at $(\beta', \delta')' = (\beta_0', 0')'$ with probability 1 by assumption. The "usual" consistency proof (see e.g. Pötscher and Prucha (1991) for a discussion) is based on the fact that the limit objective function is nonrandom. One of the difficulties of this consistency proof is the fact that the limit objective function $Q(.,.,.)$ is asymptotically random. First note that by assumption, $\hat{\delta} = O_P(1)$, so we can find a compact set C_η such that $P(\hat{\delta} \in C_\eta) \geq 1 - \eta$. Therefore, we can assume that $\hat{\delta} \in C_\eta$ for the remainder of this proof. By Lemma 1, it is possible to deduce

$$\mathrm{argmin}_{\alpha \in A} Q_n(\omega, \alpha) \xrightarrow{p} \mathrm{argmin}_{\alpha \in A} Q(\omega, \alpha) \tag{71}$$

where $\alpha = (\beta', \delta')'$, by verifying the conditions of the Lemma 1. We will set the parameter space A equal to $B \times C_\eta$. The conditions of Lemma 1 are verified as follows. Q is uniquely minimized at $(\beta_0', 0)'$. Q is almost surely continuous by the dominated convergence theorem, compactness of $C_\eta \times B$, and the continuity of $H(.,.,.)$ in all its arguments, and is uniquely minimized at $(\beta_0', 0')'$ with probability 1 by assumption. It therefore only remains to prove weak convergence of Q_n to Q on $C[B \times C_\eta]$. This will follow if we show

$$\sup_{\beta \in B, \delta \in C_\eta} |n^{-1} \sum_{t=1}^{n} (f(w_t, \varepsilon_t + a_n - \delta'n^{-1/2}z_t, \beta) - H(\beta, n^{-1/2}z_t, \delta))| \xrightarrow{p} 0 \tag{72}$$

and

$$n^{-1} \sum_{t=1}^{n} H(\beta, n^{-1/2}z_t, \delta) \Rightarrow \int_0^1 H(\beta, Z(\xi), \delta)d\xi. \tag{73}$$

17

The second result is relatively easy to show. By the continuous mapping theorem, for each $(\beta', \delta')'$ we have convergence in distribution. By Billingsley (1968), Theorem 8.1, it suffices to verify tightness to obtain the second result. Tightness follows from Theorem 8.2 of Billingsley (1968). To show it, note that $\sup_{1 \leq t \leq n} n^{-1/2}|z_t| = O_P(1)$ and therefore with asymptotically arbitrary large probability $\sup_{1 \leq t \leq n} n^{-1/2}|z_t| \leq K$ for some $K > 0$. If the latter condition holds,

$$\sup_{\alpha \in A} \sup_{\alpha':|\alpha-\alpha'|<\rho} |n^{-1} \sum_{t=1}^{n} H(\beta, n^{-1/2}z_t, \delta) - H(\beta', n^{-1/2}z_t, \delta')|$$

$$\leq n^{-1} \sum_{t=1}^{n} \sup_{\alpha \in A} \sup_{\alpha':|\alpha-\tilde{\alpha}|<\rho} \sup_{|z| \leq K} |H(\beta, z, \delta) - H(\tilde{\beta}, z, \tilde{\delta})| \to 0 \tag{74}$$

as $\rho \to 0$ by uniform continuity. This completes the tightness proof. To show the result of Equation (72), using a construction as in Billingsley (1968), Theorem 9.1, we note that $n^{-1/2}z_{[\xi n]} = Z_n(\xi) - Y_n(\xi)$ where

$$Y_n(\xi) = (n\xi - [n\xi])n^{-1/2}(z_{[n\xi+1]} - z_{[n\xi]}), \tag{75}$$

and z_{n+1} and z_0 are both zero by definition here. The construction is such that $Z_n(\xi)$ is continuous. Now

$$b_n \equiv \sup_{\xi \in [0,1]} |Y_n(\xi)| \leq n^{-1/2} \sup_{2 \leq t \leq n} |z_t - z_{t-1}|, \tag{76}$$

and in addition we note that for all $\varepsilon > 0$, by the Markov inequality,

$$\limsup_{n \to \infty} P(n^{-1/2} \sup_{2 \leq t \leq n} |z_t - z_{t-1}| > \varepsilon) \leq \limsup_{n \to \infty} \sum_{t=2}^{n} P(n^{-1/2}|z_t - z_{t-1}| > \varepsilon)$$

$$\leq \limsup_{n \to \infty} \varepsilon^{-2-\delta} n^{-1-\delta/2} \sum_{t=2}^{n} E|z_t - z_{t-1}|^{2+\delta} = 0 \tag{77}$$

by assumption. Next, note that by tightness of Z_n, for each $\eta > 0$ there exists a compact subset K_η of $C^k[0,1]$ such that

$$\limsup_{n \to \infty} P(Z_n \in K_\eta) \geq 1 - \eta. \tag{78}$$

Therefore, with probability exceeding $1 - \eta$, by the triangle inequality,

$$\sup_{\alpha \in A} |n^{-1} \sum_{t=1}^{n} (f(w_t, \varepsilon_t + a_n - \delta'n^{-1/2}z_t, \beta) - H(\beta, n^{-1/2}z_t, \delta))|$$

18

$$= \sup_{\alpha \in A} |n^{-1} \sum_{t=1}^{n} (f(w_t, \varepsilon_t + a_n - \delta'(Z_n(t/n) - Y_n(t/n)), \beta) - H(\beta, n^{-1/2} z_t, \delta))|$$

$$\leq \sup_{\alpha \in A} \sup_{h \in K_\eta} |n^{-1} \sum_{t=1}^{n} f(w_t, \varepsilon_t - \delta'h(t/n), \beta) - Ef(w_t, \varepsilon_t - \delta'h(t/n), \beta)|$$

$$+ n^{-1} \sum_{t=1}^{n} \sup_{\alpha \in A} \sup_{h \in K_\eta} |f(w_t, \varepsilon_t + a_n + \delta'Y_n(t/n) - \delta'h(t/n), \beta) - f(w_t, \varepsilon_t - \delta'h(t/n), \beta)|. \quad (79)$$

Note that measurability of the last supremum is guaranteed by compactness of K_η and continuity of $f(., ., .)$ in all arguments. We will show that the last two expressions converge to zero in probability. The expectation of the second term is smaller than

$$E \sup_{\beta \in B} \sup_{\phi, \tilde{\phi}: |\phi - \tilde{\phi}| \leq a_n + Cb_n} |f(w_t, \varepsilon_t + \phi, \beta) - f(w_t, \varepsilon_t + \tilde{\phi}, \beta)| \quad (80)$$

for some fixed constant C, and the above expression converges to zero as $n \to \infty$ by continuity and because $a_n \xrightarrow{p} 0$ and $b_n \xrightarrow{p} 0$. Next, we apply Theorem 5.2 of Pötscher and Prucha (1991) (a uniform law of large numbers) to show that the first term of Equation (79) converges to zero in probability. We will need the fact that that by compactness of K_η, $\sup_{h \in K_\eta} \sup_{x \in [0,1]} |h(x)| < \infty$ for all $\varepsilon > 0$. Pötscher and Prucha's Assumption 5.1 (compactness of the parameter space) holds by assumption. Assumptions B, C, and D are easily verified from our assumptions. To verify their Assumption 5.2, we first note that

$$\sup\{f(w_t, \varepsilon_t - \tilde{\delta}'\tilde{h}(t/n), \tilde{\beta}) : \tilde{h} \in B(h, \rho) \cap K_\eta, \tilde{\beta} \in B(\beta, \rho'), \tilde{\delta} \in B(\delta, \rho'')\}$$

$$= \sup\{f(w_t, \varepsilon_t - \delta'h(t/n) + \xi, \tilde{\beta}) : |\xi| \leq C(\rho + \rho''), \tilde{\beta} \in B(\beta, \rho'), \delta \in C_\eta\}, \quad (81)$$

and the proof is concluded if we can show that a LLN for those last quantities holds. For this, we use Theorem 6.9 of Pötscher and Prucha (1991), and note that it is easily verified that the conditions of this theorem hold here.

Proof of Theorem 2:

Using a Taylor series expansion around $(\beta_0', \theta_0')'$, for n large enough we have

$$0 = n^{-1/2} \sum_{t=1}^{n} (\partial/\partial\alpha) f(w_t, \varepsilon_t + (\theta_0 - \hat{\theta})'z_t + a_n, \hat{\beta})$$

$$= n^{-1/2} \sum_{t=1}^{n} (\partial/\partial\alpha) f(w_t, \varepsilon_t + a_n, \beta_0)$$

$$+ n^{-1/2} \sum_{t=1}^{n} (\partial/\partial\alpha)(\partial/\partial\alpha') f(w_t, \varepsilon_t + (\theta_0 - \tilde{\theta})' z_t + a_n, \tilde{\beta})(\hat{\alpha} - \alpha_0) \tag{82}$$

for some mean values $(\tilde{\theta}, \tilde{\beta})$. Therefore,

$$(n^{1/2}(\hat{\beta} - \beta_0)', n(\hat{\theta} - \theta_0)')' = \begin{pmatrix} n^{-1} A_{11}(\tilde{\beta}, \tilde{\theta}) & n^{-3/2} A_{12}(\tilde{\beta}, \tilde{\theta}) \\ n^{-3/2} A_{21}(\tilde{\beta}, \tilde{\theta}) & n^{-2} A_{22}(\tilde{\beta}, \tilde{\theta}) \end{pmatrix}^{-1} \tilde{B}. \tag{83}$$

In De Jong (2001) it was shown that $A_{11}(\tilde{\theta}, \tilde{\beta})$ and $A_{12}(\tilde{\theta}, \tilde{\beta})$ can be asymptotically replaced by $A_{11}(\theta_0, \beta_0)$ and $A_{12}(\theta_0, \beta_0)$ under the conditions of the theorem (using the assumptions of Equations (24), (25) and (26)). Note that for $A_{22}(\tilde{\theta}, \tilde{\beta})$, this result is easily obtained using an analogous argument.

Proof of Theorem 3:

Let A^{ij} denote element (i,j) of a matrix A and let B^i denote element i of a vector B. Then for some mean value $\tilde{\theta}$,

$$|\tilde{\Omega}^{ij} - \hat{\Omega}^{ij}|$$

$$\leq \max_{i,j} | \sum_{l=-n+1}^{n-1} k(l/\gamma_n) n^{-1} \sum_{t=1}^{n-l} (g(w_t, y_t - \hat{\theta}' x_t, \hat{\beta})^i g(w_{t+l}, y_{t+l} - \hat{\theta}' x_{t+l}, \hat{\beta})^j$$

$$- g(w_t, \varepsilon_t, \hat{\beta})^i g(w_{t+l}, \varepsilon_{t+l}, \hat{\beta})^j)|$$

$$\leq \max_{i,j} | \sum_{l=-n+1}^{n-1} k(l/\gamma_n)(\hat{\theta} - \theta_0)' n^{-1} \sum_{t=1}^{n-l} x_t ((\partial/\partial\varepsilon) g(w_t, y_t - \tilde{\theta}' x_t, \hat{\beta})^i) g(w_{t+l}, y_{t+l} - \tilde{\theta}' x_{t+l})^j$$

$$+ x_t ((\partial/\partial\varepsilon) g(w_{t+l}, y_{t+l} - \tilde{\theta}' x_{t+l}, \hat{\beta})^i) g(w_t, y_t - \tilde{\theta}' x_t)^j$$

$$\leq (\sup_{1 \leq t \leq n} n^{-1/2} |x_t|)(n(\hat{\theta} - \theta_0)) \sum_{l=-n+1}^{n-1} |k(l/\gamma_n)| \times$$

$$2 \max_{i,j} n^{-3/2} \sum_{t=1}^{n-l} |((\partial/\partial\varepsilon)g(w_t, \varepsilon_t + (\theta_0 - \tilde{\theta})'x_t, \hat{\beta})^i)g(w_{t+l}, y_{t+l} - \tilde{\theta}'x_{t+l})^j|. \tag{84}$$

Next, note that by Theorem 2 under the stated conditions $\sup_{1 \le t \le n} n^{-1/2}|x_t|$ and $n(\hat{\theta} - \theta_0)$ are $O_P(1)$. Also,

$$\sup_{1 \le t \le n} |(\theta_0 - \hat{\theta})'x_t| = O_P(n^{-1/2}), \tag{85}$$

and therefore for n large enough with probability arbitrarily close to one, by the Cauchy-Schartz inequality,

$$(n^{-1} \sum_{t=1}^{n-l} |((\partial/\partial\varepsilon)g(w_t, \varepsilon_t + (\theta_0 - \hat{\theta})'x_t, \hat{\beta})^i)g(w_{t+l}, y_{t+l} - \hat{\theta}'x_{t+l})^j|)^2$$

$$\le n^{-1} \sum_{t=1}^{n} \sup_{\gamma \in \Gamma} \sup_{\beta \in B} |(\partial/\partial\varepsilon)g(w_t, \varepsilon_t + \gamma, \beta)|^2 n^{-1} \sum_{t=1}^{n} \sup_{\gamma \in \Gamma} \sup_{\beta \in B} |g(w_t, \varepsilon_t + \gamma, \beta)|^2 \tag{86}$$

which is $O_P(1)$ by assumption. Therefore, the remaining probability order is that of

$$n^{-1/2} \sum_{l=-n+1}^{n-1} |k(l/\gamma_n)| = O_P(\gamma_n n^{-1/2} \int_{-\infty}^{\infty} |k(x)|dx) = o_P(1) \tag{87}$$

by assumption.

References

Andrews, D.W.K. (1991): "Heteroscedasticity and autocorrelation consistent covariance matrix estimation", Econometrica, 59, 817-858.

Billingsley, P. (1968): *Convergence of probability measures.* New York: Wiley.

Davidson, J. (1994): *Stochastic limit theory.* Oxford: Oxford University Press.

Davidson, J. and D. Peel (1998): "A nonlinear error correction mechanism based on the bilinear model", Economics Letters, 58, 165-170.

De Jong, R.M. and J. Davidson (2000): "Consistency of kernel estimators of heteroscedastic

and autocorrelated covariance matrices", Econometrica 68, 407-424.

De Jong, R.M. (2001): "Nonlinear estimation using estimated cointegrating relations", Journal of Econometrics, 101, p. 109-122.

De Jong, R.M. (2000): "A strong consistency proof for heteroscedasticity and autocorrelation consistent covariance matrix estimators", Econometric Theory, 16, p. 262-267.

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. and C.W.J. Granger (1987): "Cointegration and error correction: Representation, estimation, and testing", Econometrica, 55, 251-276.

Engle, R.F. and S.B. Yoo (1987): "Forecasting and testing in cointegrated systems", Journal of Econometrics, 35, 143-159.

Gallant, A.R. and H. White (1988): *A unified theory of estimation and inference for nonlinear dynamic models.* New York: Basil Blackwell.

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. and T.H. Lee (1989): "Investigation of production, sales and inventory relationships using multicointegration and on-symmetric error correction models", Journal of Applied Econometrics, 14, 145-159.

Granger, C.W.J. and T. Terasvirta (1993): *Modelling nonlinear economic relations.* Oxford: Oxford University Press.

Hansen, B.E. (1992): "Consistent covariance matrix estimation for dependent heterogeneous processes", Econometrica, 60, 967-972.

Johansen, S. (1994), "Determination of Cointegration Rank in the Presence of a Linear Trend", Oxford Bulletin of Economics and Statistics, 54, 383-397.

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.

Kim, J. and D. Pollard (1990): "Cube root asymptotics", Annals of Statistics, 18, 191-219 .

Newey, W.K. and K.D. West (1987): "A simple, positive semi-definite, heteroscedasticity and autocorrelation consistent covariance matrix", *Econometrica*, 55, 703-708.

Nobay, A.R. and D.A. Peel (1997): "Risk premia and the term structure of interest rates: evidence from non-linear error correction mechanisms", working paper, Financial Markets

Group, London School of Economics.

Phillips, P.C.B. and S. Ouliaris (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. and I.R. Prucha (1991): "Basic structure of the asymptotic theory in dynamic nonlinear econometric models, part 1: consistency and approximation concepts", Econometric Reviews, 10, 125-216.

Saikkonen, P. (1995): "Problems with the asymptotic theory of maximum likelihood estimation in integrated and cointegrated systems", Econometric Theory 11, 888-911.

Sims, C., Stock, J. and M. Watson (1990): "Inference in linear time series models with some unit roots", Econometrica, 58, 113-144.

White, H. (1984): *Asymptotic theory for econometricians.* New York: Academic Press.