Adaptive Estimation of Error Correction Models

Hodgson, Douglas J.

Working Paper No. 410
September 1995

Adaptive Estimation of Error Correction Models

Douglas J. Hodgson

Rochester Center for Economic Research
Working Paper No. 410

September 1995

ADAPTIVE ESTIMATION OF ERROR CORRECTION MODELS

Douglas J. Hodgson[1]

Department of Economics

University of Rochester

Harkness Hall

Rochester, NY

14627-0156

June 9, 1995

The paper considers the adaptive maximum likelihood estimation of reduced rank vector error correction models. It is shown that such models can be asymptotically efficiently estimated even in the absence of knowledge of the shape of the density function of the innovation sequence, provided that this density is symmetric. The construction of the estimator, involving the nonparametric kernel estimation of the unknown density using the residuals of a consistent preliminary estimator, is described, and its asymptotic distribution is derived. Asymptotic efficiency gains over the Gaussian pseudo-MLE are evaluated for elliptically symmetric innovations.

[1] This paper is based on Chapter 2 of my Ph.D. thesis at Yale University. I am grateful to Peter Phillips and Oliver Linton for advice and assistance, and to the Alfred P. Sloan Foundation and the Social Sciences and Humanities Research Council of Canada for financial support.

I. INTRODUCTION

Contemporary empirical researchers in macroeconomics and finance make considerable use of error correction representations in the modeling of cointegrated systems. Such representations are always possible (Engle and Granger (1987)), and derive their name from the fact that the deviations of a system from its cointegrating relationships are explicitly modeled as impacting upon subsequent short-run dynamics. An error correction representation can be derived from a vector autoregression (VAR) by taking first differences. The fact that the system is cointegrated implies that among the regressors in the differenced VAR will be a term in the lagged levels of the variables, with an accompanying coefficient matrix that has reduced rank equal to the number of cointegrating relationships.

A natural approach to estimating such a model is reduced rank regression. For the case of stationary VAR's, reduced rank regression estimators have been analyzed by Ahn and Reinsel (1988) and Velu, Reinsel, and Wichern (1986). In the nonstationary case, the reduced rank structure implies an error correction representation, as the reduced rank matrix can be decomposed into a matrix of cointegrating vectors and a matrix of error correction coefficients, or factor loadings, characterizing the effects of the deviations from the cointegrating relationships on the transitory dynamics of the system. The estimation of cointegrated systems by reduced rank regression has been analyzed by Johansen (1988, 1991) and Ahn and Reinsel (1990). These authors derive maximum likelihood estimators (MLE's) of the model assuming Gaussian innovations to the underlying VAR. The Gaussian reduced rank MLE has been widely employed in applied cointegration analysis. For example, Johansen (1992), Johansen and Juselius (1990), and Friedman and Kuttner (1992) estimate monetary models, Johansen and Juselius (1992) estimate exchange rate models, Kunst and Neusser (1990) estimate real business cycle models, and Kasa (1992) estimates models of stock prices and dividends.

If the assumption of Gaussianity is correct, then the estimators of Johansen (1988) and Ahn and Reinsel (1990) are asymptotically efficient and should have performance superior to that of alternatives such as the least squares estimator of Engle and Granger (1987). Indeed, Ahn and Reinsel (1990) report a simulation study comparing their estimator of the cointegrating parameter in a Gaussian bivariate model with OLS. They find a large improvement in mean squared error, for all sample sizes considered (50 through 400), when the efficient MLE is used. An extensive and general analysis of efficient estimation of cointegrated models in the Gaussian case is provided by Phillips (1991).

Although the MLE's discussed above are asymptotically efficient when the innovations are Gaussian, they are inefficient when the innovations are non-Gaussian. In the latter case, the efficient MLE will take a different form. As Ahn and Reinsel's (1990) own simulations show, it matters in the estimation of cointegrating vectors whether or not an efficient estimator is used. Some of the applied studies cited above (Johansen and Juselius (1990, 1992) and Kasa (1992)) test and reject the Gaussianity hypothesis for the estimated residuals. The rejections are due primarily to excess kurtosis. This result is not surprising, given that many economic time series, especially speculative prices, are well documented to be driven by leptokurtic processes (see, for example, Mandelbrot (1963), Fama (1963, 1965), Mittnik and Rachev (1993), and McGuirk, Robertson, and Spanos (1993)). That Gaussian reduced rank estimators can give poor estimates when using thick-tailed data has been demonstrated by Phillips (1993) in the context of empirical exchange rate models.

In problems for which a Gaussian MLE is inappropriate, adaptive estimation, which can be employed when the underlying density function of the data generating process is of unknown shape, provides a highly attractive alternative. An adaptive estimator shares the asymptotic optimality properties of the MLE, differing from the latter in that a nonparametric estimate of the score function of the log-likelihood replaces the analytic expression that would be used if the density were known. An adaptive estimator can be viewed as an MLE when the shape of the likelihood is unknown. A simulation study by McDonald and White (1993) finds that adaptive estimators compare quite favourably with OLS, LAD, GMM, and M-estimators in the estimation of a (non-cointegrating) non-Gaussian linear regression model.

The present paper analyzes adaptive estimators for reduced rank regression in cointegrated error correction models. It builds on and follows the work of Jeganathan (1994), who analyzes the adaptive estimation of linear cointegrating regressions. In section II, the model and notation are introduced. In section III, we show that this model falls within the locally asymptotically normal (LAN) and locally asymptotically mixed normal (LAMN) family of models, with the component of the model associated with short-run dynamics being LAN and the component associated with long-run dynamics being LAMN. In section IV, we introduce a class of estimators termed asymptotically centering (AC), and describe the optimality properties of these estimators in LAN/LAMN models. Although AC estimators are asymptotically equivalent to the MLE, they require knowledge of the shape of the density function of the innovations, knowledge we are assumed to lack. In section V, we show how to construct estimators that are asymptotically equivalent to AC estimators, and so share their optimality properties, but do not require knowledge of the shape of the density of the innovations. These estimators, termed *adaptive*, utilize nonparametric density estimates to consistently estimate the score and information of the log-likelihood function. We derive the asymptotic distribution of an adaptive estimator and, for the special case of elliptically symmetric innovation densities, evaluate its efficiency gains over the Gaussian pseudo-MLE. Section VI extends the basic stochastic model specification to allow for non-zero intercept terms in the cointegrating vectors. Section VII discusses possible extensions of the research. The Appendix contains the proofs of all lemmas and theorems.

The following notation is used throughout the paper. I_s denotes the identity matrix of dimension s, $|x|$ the Euclidean norm of the vector x, $I(\cdot)$ the indicator function, $N(x, V)$ the distribution of a random variable that is normal with mean vector x and covariance matrix V, and $MN(x, V)$ a mixed normal distribution, i.e. one in which the covariance matrix V is random. The vectorization operator *vec(X)* stacks the transposed rows of the matrix X, while the inequalities $X > Y$ and $X \geq Y$, when applied to matrices, signify that the difference X-Y is positive definite and positive semi-definite, respectively. We simplify the notation by writing $\int_0^1 B$ in place of $\int_0^1 B(r)dr$ when $B(r)$ is a Brownian motion process defined on the interval $[0,1]$. $L(X|P)$ denotes the distribution (or law) of X with respect to the probability measure P. When P is the distribution of X itself, $L(X|P)$ is abbreviated to $L(X)$. The weak convergence of probability measures is denoted by the symbol $\Rightarrow$.

II. THE MODEL AND NOTATION

We assume that the q-dimensional stochastic process X_t is observed for all $t=1,\ldots,n$. Considered individually, each of the q series is integrated of order one, but we shall assume that there exist r cointegrating relationships among the variables, with $1 \leq r < q$, and that r is known. We also assume that the data generating process for X_t can be characterized by the following VAR, of known order k:

(1) $$X_t = \Pi_1 X_{t-1} + \ldots + \Pi_k X_{t-k} + \varepsilon_t .$$

In addition, we assume that initial observations $X_{1-k},\ldots,X_0$ are available. The implications for the lag polynomial $\Pi(z) = I - \sum_{j=1}^{k} \Pi_j z^j$ of our assumption that r cointegrating vectors exist are that $\det\{\Pi(z)\} = 0$ has q-r roots on the unit circle and r roots outside the unit circle.

So far, our model is identical to those of Johansen (1988) and Ahn and Reinsel (1990). Where we differ from these authors is in our assumptions regarding the distribution of the iid innovation process ε_t. Whereas they assume that this distribution is Gaussian, we allow it to belong to a much broader class and to be unknown to the investigator. This class is restricted by assuming that the true distribution has a Lebesgue density function, $p(\varepsilon)$, that is symmetric about the origin in the sense that $p(\varepsilon) = p(-\varepsilon)$, and that satisfies the moment condition $E\left(|\varepsilon_{it}|^{2+\delta}\right) < \infty$ for some $\delta > 0$, for all $i = 1,\ldots,q$. The symmetry assumption is important for our purposes because it implies that the q-vector of partial derivatives is anti-symmetric, i.e. that $\partial p(\varepsilon) / \partial\varepsilon = -\partial p(-\varepsilon) / \partial\varepsilon$. Consequently, the q-dimensional (negative of the) score function of $p(\varepsilon)$, which we denote by $\psi(\varepsilon) = \left(\partial p(\varepsilon) / \partial\varepsilon\right) / p(\varepsilon)$, is also anti-symmetric in ε. This latter property facilitates our derivation of an adaptive estimator because it allows us to apply a result of Jeganathan (1988) in showing that the sample score function of the model can be consistently estimated through the use of a nonparametric kernel estimator of $\psi(\varepsilon)$. Finally, we assume that the information matrix of $p(\varepsilon)$, $\Omega \equiv \int \psi(\varepsilon)\psi(\varepsilon)' p(\varepsilon) d\varepsilon$, is finite and positive definite.[1] Note that in the special case where $p(\varepsilon)$ is Gaussian, $\Omega = \Sigma_\varepsilon^{-1}$, where $\Sigma_\varepsilon = \mathrm{cov}(\varepsilon)$. If Gaussianity does not hold, then $\Omega > \Sigma_\varepsilon^{-1}$.

[1] It follows from this assumption that $0 < \lambda^2 \equiv \int |\psi(\varepsilon)|^2 p(\varepsilon) d\varepsilon < \infty$, since $\lambda^2 = trace(\Omega)$.

As noted in the Introduction, first differencing the VAR (1) yields the following error-correction representation:

$$(2) \qquad \Delta X_t = ABX_{t-1} + \sum_{j=1}^{k-1} \Phi_j \Delta X_{t-j} + \varepsilon_t,$$

where the $q \times r$ matrix A is a matrix of error correction coefficients and the rows of the $r \times q$ matrix B are cointegrating vectors. This model is analyzed by Johansen (1988) and Ahn and Reinsel (1990) under the Gaussianity assumption. The primary goal of our paper is to adaptively estimate A and B.

As the model stands, A and B are unidentified. In what follows, we only consider the estimation of an identified model. Following Ahn and Reinsel (1990), this is achieved by partitioning the variables in X_t as $X_t = [X_{1t}', X_{2t}']'$, with X_{1t} having r elements and X_{2t} having q-r elements, such that the subsystem X_{2t} contains q-r unit roots, and by writing $B = [I_r, -B_0]$, where B_0 has dimensions $r \times (q-r)$. The $r(q$-$r)$ elements of B_0 are the long-run coefficients in this model. We can then rewrite (2) as

$$(3) \qquad \Delta X_t = A[X_{1,t-1} - B_0 X_{2,t-1}] + \Phi Z_{t-1} + \varepsilon_t,$$

where $\Phi = [\Phi_1, \ldots, \Phi_{k-1}]$ and $Z_{t-1} = [\Delta X_{t-1}', \ldots, \Delta X_{t-k+1}']'$.

It is clear from equation (3) why the representation is termed an error correction model. The bracketed expression is an r-vector of transitory fluctuations of the system about its cointegrating relationships, while A determines the reactions of the system to these fluctuations. The system's remaining transitory dynamics are characterized by Φ, the matrix of coefficients on lagged first differences, which is treated as a nuisance parameter in most applications. Our primary objective is to efficiently estimate A and B_0, adapting for the unknown density $p(\varepsilon)$. However, we would also like to adapt for the unknown nuisance parameter Φ. It will be shown below that this can be done for B_0, but not for A. In fact, we will show that B_0 can be efficiently estimated adapting for unknown $p(\varepsilon)$, Φ, *and* A. In other words, if our interest is confined to estimating the system's long-run dynamics, we can do as well, asymptotically, not knowing its short-run dynamics as we can knowing them. Conversely, we can estimate the short-run dynamics as well not knowing the long-run dynamics as we can knowing them. However, we can always improve our estimates of certain short-run components, even asymptotically, if other short-run components are known, vis-a-vis the case where the latter are unknown.

The development of the asymptotic theory in subsequent sections will be facilitated by the arrangement of all of the model's unknown parameters into a single vector. To this end, we define $\alpha = vec(A)$, $\varphi = vec(\Phi)$, and $\beta = vec(B_0)$, of dimensions qr, $q^2(k-1)$, and $r(q-r)$, respectively. These vectors are gathered into the m-dimensional full parameter vector, $\theta = [\alpha', \varphi', \beta'] = [\eta', \beta']$, where $\eta \equiv (\alpha', \varphi')'$, $m = 2qr - r^2 + q^2(k-1)$, and θ belongs to the parameter space Θ, which is taken to be all of R^m (excepting points at which either A or B_0 is of deficient rank). Defining

$s = qr + q^2(k-1)$ as the total number of parameters in the stationary component of the model (i.e. the dimension of η) allows us to define the $m \times m$ scaling matrix $\delta_n = diag\left[n^{-1/2}I_s, n^{-1}I_{m-s}\right]$. We can then write the local representation of the full parameter vector θ as $\theta_n = \theta + \delta_n h_n$, where $\{h_n\}$ is a sequence of bounded m-vectors. Note that θ_n converges to θ, but does so at different rates in different directions of the parameter space. In directions associated with transitory dynamics, the rate of convergence is $\sqrt{n}$, whereas in those associated with non-stationary dynamics, the rate is n.

III. LAN AND LAMN LIMIT THEORY

In this section, we derive the asymptotic distribution of the log-likelihood ratio,

$$\Lambda_n(\theta_n, \theta) = \log\left(dP_{\theta_n,n} \big/ dP_{\theta,n} \right),$$

where $P_{\theta,n}$ is the distribution of the sample of size n with parameter θ. We find that the limit theory is such that the model falls within the LAN/LAMN family. This is important because a theory of optimal estimation applies to such models. In Section IV, optimal estimators for this family are characterized.

We prove that the component of the model associated with the parameters describing the long-run relationships in the model (i.e. the cointegrating coefficients B_0) has an LAMN limit theory, while the component associated with parameters describing short-run dynamics (i.e. A and Φ) has an LAN limit theory.

We show that these two components are asymptotically independent, allowing separate adaptive estimation of the coefficients associated with long-run and short-run dynamics, respectively, but not allowing adaptive estimation of the error-correction coefficients (A) separately from other parameters characterizing short-run dynamics (Φ).

To derive the limit theory of $\Lambda_n(\theta_n,\theta)$, we couch the model in terms of the framework of a general non-linear model, as described by Jeganathan (1994). We assume that the initial observations $\underline{X}_0 \equiv (X_{1-k},\ldots,X_0)$ have density $f_0(\underline{X}_0,\theta)$, with the property that $f_0(\underline{X}_0,\theta_n)-f_0(\underline{X}_0,\theta)=o_p(1)$ in $P_{\theta,n}$ as $\theta_n\to\theta$. We define $\underline{X}_t \equiv (\underline{X}_0,X_1,\ldots,X_t)$ and denote by F_t the $\sigma-$field generated by $\underline{X}_t$. Applying this notation to equation (27) of Jeganathan (1994), where we set $\sigma_t(\underline{X}_{t-1},\theta)=1$, yields:

$$X_t = g_t(\underline{X}_{t-1},\theta)+\varepsilon_t,$$

where ε_t is as in equation (2) and, also using (2), we have

$$\begin{aligned} g_t(\underline{X}_{t-1},\theta) &= X_{t-1}+ABX_{t-1}+\sum_{j=1}^{k-1}\Phi_j Y_{t-j} \\ &= X_{t-1}+ABX_{t-1}+\Phi Z_{t-1}. \end{aligned} \tag{4}$$

The following result is derived within the preceding non-linear framework and is useful in our subsequent development of the LAN/LAMN limit theory:

Lemma 3.1 : *Defining* $d_t(\theta_n,\theta) = g_t(\underline{X}_{t-1},\theta_n) - g_t(\underline{X}_{t-1},\theta)$, *we have*

$$(5) \qquad d_t(\theta_n,\theta)' = h_n'\,\delta_n H_{t-1}(\theta) - n^{-3/2} X_{2,t-1}'\, b_n'\, a_n' \,,$$

where $H_{t-1}(\theta) \equiv \left[(I_q \otimes W_{t-1})', (I_q \otimes Z_{t-1})', (-A' \otimes X_{2,t-1})'\right]'$, $W_{t-1} = X_{1,t-1} - B_0 X_{2,t-1}$, *and* $\{a_n\}$ *and* $\{b_n\}$ *are bounded sequences of matrices of dimensions* $q \times r$ *and* $r \times (q-r)$, *respectively.*

Remarks:

(a) The q-vector $d_t(\theta_n,\theta)$ plays an important role in the theory developed below. Since the essential goal of this paper is to construct an adaptive version of the maximum likelihood estimator, the derivation of expressions for the sample score vector and information matrix is necessary. Writing $d_t(\theta_n,\theta)'$ as in (5) is instructive in this regard because postmultiplying $d_t(\theta_n,\theta)'$ by the score of the innovation density evaluated at observation t, $-\psi(\varepsilon_t)$, and summing over t, results in an expression for a linear combination of the elements of the sample score vector (scaled by δ_n) plus an asymptotically negligible component. The scaled sample score is $-\sum_{t=1}^{n} \delta_n H_{t-1}(\theta)\psi(\varepsilon_t)$. We show below that this quantity is $O_p(1)$ in $P_{\theta,n}$ and that the second term on the right-hand side of (4), postmultiplied by $\psi(\varepsilon_t)$ and summed over t, converges in probability to zero. We shall also show that an expression for the asymptotic information matrix of the model can be obtained by postmultiplying $d_t(\theta_n,\theta)'$ by $\psi(\varepsilon_t)$, summing the squares over t, and considering the limiting behaviour of the sum. The result is a quadratic form in the asymptotic information matrix, the weak limit of the standardized sum $\sum_{t=1}^{n}\delta_n H_{t-1}(\theta)\Omega H_{t-1}(\theta)'\delta_n$. Once again, only the first term in (5) is asymptotically relevant.

Note that in these expressions for the sample score vector and information matrix appear the quantities $\psi(\varepsilon)$ and Ω, which are, respectively, the (negative of the) score and the information of the innovation density $p(\varepsilon)$. The basic problem addressed in this paper is the fact that we assume that $p(\varepsilon)$, and consequently also $\psi(\varepsilon)$ and Ω, are unknown to the investigator. In Section V, we show how this problem can be addressed through the estimation of $\psi(\varepsilon)$ and Ω using nonparametric density estimates.

(b) Note that $d_t(\theta_n,\theta)\in F_{t-1}$. Recognition of this fact simplifies the proof of Theorem 3.2 below because we deal throughout the proof with statistics containing $d_t(\theta_n,\theta)$ that are analyzed conditional on F_{t-1}. Furthermore, since ε_t is independently distributed, $\psi(\varepsilon_t)$ is independent of F_{t-1} and therefore of $d_t(\theta_n,\theta)$, so that the two Brownian motion processes appearing in our derivation of the limit theory for the nonstationary component of the scaled sample score $W_n(\theta)$, given below, are independent. Our LAMN limit theory for the nonstationary component is a consequence of this independence.

The key result in showing that our model falls within the LAN/LAMN family is Theorem 3.2, given below. A more general family of models than the LAN/LAMN is the *locally asymptotically quadratic* (LAQ) family. A model falls into the LAQ family when its sample likelihood ratio $\Lambda_n(\theta_n, \theta)$ can be asymptotically approximated by a quadratic function of the vector h_n (recall that $h_n = \delta_n^{-1}(\theta_n - \theta)$). Jeganathan (1994) and LeCam and Yang (1990) formally define LAQ families. Theorem 3.2 shows that the likelihood ratios in the present model can be asymptotically approximated by the quadratic given in equation (6).

<u>Theorem 3.2</u>: *The likelihood ratios* $\Lambda_n(\theta_n, \theta)$ *have the following asymptotic quadratic approximation*:

(6)
$$\Lambda_n(\theta_n, \theta) = -\sum_{t=1}^{n} h_n' \delta_n H_{t-1}(\theta)\psi(\varepsilon_t) - \frac{1}{2}\sum_{t=1}^{n} h_n' \delta_n H_{t-1}(\theta)\Omega H_{t-1}(\theta)' \delta_n h_n + o_p(1) \ in\, P_{\theta,n}.$$

<u>Remark</u>: The linear term in the quadratic approximation given by (6) is the scaled sample score vector, while the quadratic term is the scaled sample information matrix. We shall denote these quantities as follows:

(7)
$$W_n(\theta) = -\sum_{t=1}^{n} \delta_n H_{t-1}(\theta)\psi(\varepsilon_t)$$

(8) $$S_n(\theta)=\sum_{t=1}^{n}\delta_n H_{t-1}(\theta)\Omega H_{t-1}(\theta)'\delta_n.$$

In deriving our LAN/LAMN limit theory, the asymptotic behaviour of both of these statistics is of central importance. The limiting behaviour of $W_n(\theta)$ is analyzed following the statement of Definition 3.3 below, but we have already, in the course of proving Theorem 3.2, derived the asymptotics for $S_n(\theta)$. In equation (65) in the Appendix, we find that $S_n(\theta)\Rightarrow S(\theta)$, where

(9) $$S(\theta)=\begin{bmatrix}\Omega\otimes E[M_tM_t'] & 0\\ 0 & A'\Omega A\otimes\int_0^1 B_2B_2'\end{bmatrix}$$

is the asymptotic information matrix, and where we use the notation $M_{t-1}=[W_{t-1}',Z_{t-1}']'$ for the $r+q(k\text{-}1)$-dimensional vector of stationary variables in the error correction model. Note that since M_{t-1} appears in each of the q equations in the system, the number of coefficients in the stationary component of the model, s, is equal to q multiplied by the number of variables in M_{t-1}.

The structure of $S(\theta)$ is very important to our theory. Recall that the first s rows of $H_{t-1}(\theta)$ consist of stationary variables, while the remaining m-s rows are nonstationary. Since $S(\theta)$ is the limit of the scaled sum of generalized outer products of $H_{t-1}(\theta)$, its leading s-dimensional submatrix is associated with the stationary component of the model and the second submatrix on the diagonal, of dimension m-s, is associated with the nonstationary component. Not surprisingly, the former submatrix is non-random and the latter is random, corresponding to the fact that stationary models typically have non-random Fisher information matrices while cointegrated models typically have random ones. In our model, the information matrix has both non-random and random components, since we are modeling both stationary and nonstationary dynamics. This structure is important because, as will be formally shown below, it implies that the stationary component of the model has LAN limit theory and the nonstationary component has LAMN limit theory. Finally, the block diagonality of $S(\theta)$ has important implications for the question of which components of θ can be adaptively estimated treating the remaining components as nuisance parameters. We discussed this question in the second-last paragraph of Section II.

We now show formally that the likelihood ratios for this model have a limit theory that is LAN with respect to the component of θ associated with transitory dynamics, and LAMN for the component associated with long-run dynamics (i.e. the cointegrating parameters). The following definition is used.

Definition 3.3 *(Jeganathan (1994)): The family* $\{P_{\theta,n};\theta\in\Theta\}$ *is said to have LAMN likelihood ratios at* $\theta\in\Theta$ *if the quadratic approximation (6) holds, and, furthermore,* $L\left(W_n(\theta),S_n(\theta)\middle|P_{\theta,n}\right)\Rightarrow L\left(S^{1/2}N(0,I),S(\theta)\right)$, *where* $S(\theta)$ *is positive definite almost surely and* $N(0,I)$ *is a standard Gaussian independent of* $S(\theta)$. *In the special case where* $S(\theta)$ *is non-random, LAMN likelihoods are called LAN.*

According to this definition, a model falls into the LAMN family if its likelihood ratios can be asymptotically approximated by a quadratic in which the linear term has an asymptotic mixed normal distribution with a covariance matrix equal to the random quadratic term. Furthermore, the LAMN family is more general than the LAN family, as the latter requires the quadratic term to be non-random.

To prove our assertions regarding LAN and LAMN likelihood ratios for this model, we need only show that $L\left(W_n(\theta)\middle|P_{\theta,n}\right) \Rightarrow L\left(S^{1/2}(\theta)N(0,I)\right)$. Since the upper-left block of $S(\theta)$ is non-random, and the lower-right block is random, our earlier claims regarding the LAN and LAMN limit theory for the respective components will then hold. To analyze the limit distribution of the score $W_n(\theta)$, we write it as follows:

(10) $$W_n(\theta) = \sum_{t=1}^{n} \begin{bmatrix} -\frac{1}{\sqrt{n}}\psi(\varepsilon_t) \otimes M_{t-1} \\ \frac{1}{n} A'\,\psi(\varepsilon_t) \otimes X_{2,t-1} \end{bmatrix}.$$

The first s elements of the m-vector $W_n(\theta)$ are sums of stationary random variables scaled by $n^{-1/2}$, while the remaining m-s elements are sums of products of stationary and nonstationary random variables, scaled by n^{-1}. Hence, we would expect the first component to have a limiting Gaussian distribution with non-random covariance matrix and the second component to converge weakly to a stochastic integral of the form $\int dB \otimes B$. We now show that these conjectures are correct, beginning with an analysis of the first component of (10),

$$-\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\psi(\varepsilon_t)\otimes M_{t-1}.$$

Since $\psi(\varepsilon_t)$ is independent of M_{t-1} and $E[\psi(\varepsilon_t)]=0$, a central limit theorem for martingale difference sequences (e.g., White (1984, p. 130)) can be used to show that

$$\text{(11)} \qquad L\left(-\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\psi(\varepsilon_t)\otimes M_{t-1}\middle|P_{\theta,n}\right)\Rightarrow L\left(N\left(0,\Omega\otimes E[M_tM_t']\right)\right).$$

As for the second component of (10), we know that

$$\frac{1}{n}\sum_{t=1}^{n}A'\,\psi(\varepsilon_t)\otimes X_{2,t-1}\Rightarrow\int_0^1 A'\,dB_1\otimes B_2.$$

Since B_1 and B_2 are independent, it follows that

$$\text{(12)} \qquad L\left(\int_0^1 A'\,dB_1\otimes B_2\middle|P_{\theta,n}\right)=L\left(MN\left(0,A'\,\Omega A\otimes\int_0^1 B_2B_2'\right)\right),$$

as shown by Phillips and Park (1988). To show that $L\left(W_n(\theta)\middle|P_{\theta,n}\right)\Rightarrow L\left(MN\left(0,S(\theta)\right)\right)$, we need only verify the following equation:

$$\text{(13)} \qquad \frac{1}{n^{3/2}}\sum_{t=1}^{n}A'\,\psi(\varepsilon_t)\psi(\varepsilon_t)'\otimes X_{2,t-1}M_{t-1}'=o_p(1) \quad in \quad P_{\theta,n}.$$

To this end, we rewrite the left-hand side of (13) as

$$\frac{1}{n^{3/2}}\sum_{t=1}^{n} A'\left(\psi(\varepsilon_t)\psi(\varepsilon_t)'-\Omega\right)\otimes X_{2,t-1}M_{t-1}'+\frac{1}{n^{3/2}}\sum_{t=1}^{n} A'\,\Omega\otimes X_{2,t-1}M_{t-1}'$$
$$= o_p(1)+o_p(1)=o_p(1) \;\; in\; P_{\theta,n}.$$

The second term is $o_p(1)$ by equation (63) in the proof of Theorem 3.2, and the first term is $o_p(1)$ because

$$\frac{1}{n}\sum_{t=1}^{n}\left(\psi(\varepsilon_t)\psi(\varepsilon_t)'-\Omega\right)=o_p(1) \;\; in\; P_{\theta,n}$$

and

$$\max_{t\in\{1,\ldots,n\}}\left|\frac{X_{i;2,t-1}M_{\ell,t-1}}{\sqrt{n}}\right| = O_p(1) \;\; in\; P_{\theta,n} \quad \forall i=1,\ldots,q-r;\; \ell=1,\ldots,r+q(k-1).$$

This completes our derivation of the LAN/LAMN limit theory for our model. One important consequence of this theory is that the sequences of probability measures $\{P_{\theta,n}\}$ and $\{P_{\theta_n,n}\}$ are contiguous, and therefore have the property that the sequence of statistics $\{T_n\}$ is $o_p(1)$ in $P_{\theta_n,n}$ if and only if it is $o_p(1)$ in $P_{\theta,n}$ (see LeCam and Yang (1990, p. 20)). This fact is used below because statistics will be computed using residuals $\varepsilon(\theta_n)$ from a consistently estimated model in lieu of the true innovations $\varepsilon(\theta)$, and the fact that the latter statistics are $o_p(1)$ in $P_{\theta,n}$ will be used to show that the former are $o_p(1)$ in $P_{\theta_n,n}$.

IV. EFFICIENT ESTIMATION IN LAN/LAMN MODELS

In this section, we are concerned with the efficient estimation of our model under the assumption that the innovation density $p(\varepsilon)$ is known to the investigator. The results of the preceding section are important in this regard because they permit us to draw upon the theory of efficient estimation that has been developed for LAN/LAMN models. According to this theory, the class of estimators termed *asymptotically centering* (AC) by Jeganathan (1994) is optimal according to the *locally asymptotically minimax* (LAM) criterion. We begin this section by defining AC estimators and discussing their optimality properties. We then describe the construction of such estimators for our model. The following definition of AC estimators is given by Jeganathan (1994).

Definition 4.1: *If the model is LAMN or LAN at θ, a sequence $\{\hat{\theta}_n\}$ of estimators is called AC if*

(14) $$\delta_n^{-1}(\hat{\theta}_n - \theta) - S_n^{-1}(\theta)W_n(\theta) = o_p(1) \ \text{ in } P_{\theta,n},$$

where $W_n(\theta)$ and $S_n(\theta)$ are defined in (7) and (8).

Remark: We can see from this definition that an AC estimator for our model will have an asymptotic mixed normal distribution when appropriately scaled and centered. Furthermore, the asymptotic covariance matrix is the inverse of the Fisher information, so that an AC estimator has the same asymptotic distribution as the MLE and so shares the latter's efficiency properties. AC estimators are optimal according to the LAM criterion, which means that for any symmetric, bowl-shaped loss function and for any $\theta \in \Theta$, they achieve a lower bound for the limit inferior of the supremum of the risk over a ball around θ whose radius converges to zero as $n \to \infty$ (for a discussion, see, for example, Ghosh (1985, pp. 318-320)).

Before showing how to construct AC estimators for our model, we introduce some notation and assumptions. We assume that consistent estimates $\hat{S}_n$ of $S_n(\theta)$ exist. Under the assumptions of our model, conditions (C2) and (C5) of Jeganathan (1994) hold ((C2) because Θ is open and $\delta_n \to 0$, and (C5) because δ_n does not depend on θ). Let $\{\theta_n^*\}$ be a sequence of preliminary estimates such that

$$\delta_n^{-1}(\theta_n^* - \theta) = O_p(1) \;\; in \; P_{\theta,n} \quad \forall \theta \in \Theta,$$

and let θ_n^{**} be a discretized version of θ_n^*. The following definition of discretization is quoted from Jeganathan (1994):

> Partition the space R^m into cubes $C_i, i \geq 1$, of sides of length unity, and let $C_{ni} = \delta_n C_i = \{\delta_n u : u \in C_i\}$. If $\theta_n^* \in \Theta \cap C_{ni}$, take $\theta_n^{**} = t_{ni}$, where t_{ni} is some fixed point in $\Theta \cap C_{ni}$, which will necessarily be non-empty since $\theta_n^* \in \Theta$. The θ_n^{**} constructed in this way preserves the properties of θ_n^* in the sense that $\theta_n^{**} \in \Theta$ and $\delta_n^{-1}(\theta_n^{**} - \theta) = O_p(1)$ in $P_{\theta,n}$ for all $\theta \in \Theta$.

In practice, any preliminary estimator θ_n^* will effectively already be discretized, since it will be computed to a prespecified finite number of decimal places.

We define the quantity $W_n^*(\theta)$ as in Proposition 3 of Jeganathan (1994). (The definition is not repeated here because it would involve the introduction of a considerable amount of new notation. For our purposes, the important characteristic of $W_n^*(\theta)$ is the fact that $W_n^*(\theta) = W_n(\theta) - \hat{S}_n h_n + o_p(1)$ *in* $P_{\theta,n}$ for every bounded $\{h_n\}$. This is proved in Proposition 3 of Jeganathan (1994)).

Given the above definitions and assumptions, Theorem 2 of Jeganathan (1994) establishes that $\hat{\theta}_n$ as given in equation (15) below is an AC estimator:

(15) $$\hat{\theta}_n = \theta_n^{**} + \delta_n \hat{S}_n^{-1} W_n^*(\theta_n^{**}).$$

The following discussion, based on that of Jeganathan (1994), shows heuristically why (15) is an AC estimator and why the notion of a discretized estimator is employed. As noted above, Proposition 3 of Jeganathan (1994) proves that

(16) $$W_n^*(\theta_n) = W_n(\theta) - \hat{S}_n h_n + o_p(1) \ \ in\ P_{\theta,n}$$

for every bounded $\{h_n\}$ and for every $\theta \in \Theta$. Defining the estimator

(17) $$\overline{\theta}_n = \theta_n^* + \delta_n \hat{S}_n^{-1} W_n^*(\theta_n^*),$$

and replacing h_n in (16) with $\delta_n^{-1}(\theta_n^* - \theta)$ (which, recall, is $O_p(1)$ in $P_{\theta,n}$), it would seem that (16) and (17) could be combined to conclude that

$$\delta_n^{-1}(\bar{\theta}_n - \theta) = \hat{S}_n^{-1} W_n(\theta) + o_p(1) \ \ in\ P_{\theta,n},$$

so that $\bar{\theta}_n$ is an AC estimator. However, it is not strictly correct to replace h_n with $\delta_n^{-1}(\theta_n^* - \theta)$ in (16), since $\delta_n^{-1}(\theta_n^* - \theta)$, although confined with probability arbitrarily close to unity to a bounded interval, may assume any of an uncountably infinite number of values within such an interval. This would require the replacement of (16) with the stronger condition that

$$\sup_{|h|\leq\gamma}\left|W_n^*(\theta_n) - (W_n(\theta) - \hat{S}_n h)\right| = o_p(1) \ \ in\ P_{\theta,n}\ \forall\alpha > 0.$$

This problem can be avoided by replacing θ_n^* with θ_n^{**}. For then the quantity $\delta_n^{-1}(\theta_n^{**} - \theta)$ can only assume one of a finite number of points in any bounded interval.

V. ADAPTIVE ESTIMATION

The asymptotically efficient estimator derived in the preceding section is of no immediate use to us because of its assumption that the density $p(\varepsilon)$ is known to the investigator, an assumption we wish to avoid. In this section, we argue that our model can be efficiently estimated even if $p(\varepsilon)$ is unknown. We show how to construct an estimator that is asymptotically equivalent to the AC estimator given by (15). To accomplish this equivalence, we employ nonparametric kernel techniques to estimate the density, $p(\varepsilon)$, its score, $\psi(\varepsilon)$, and its information matrix, Ω; these estimates are then substituted into equations (7) and (8) to give us consistent estimates of the sample score and information, with which we can construct a one-step Newton-Raphson estimator similar in form to that given by (15).

Our analysis belongs to the body of research stemming from Stein's (1956) investigation of the problem of efficiently estimating a parameter of interest in the presence of an unknown infinite-dimensional nuisance parameter. The problem of adaptively estimating a location parameter using a sample of iid observations from a symmetric density of unknown shape was solved by Beran (1974) and Stone (1975), the former using Fourier series methods and the latter using a Gaussian kernel. To prevent misbehaviour of his score estimator, Stone (1975) required that extreme outliers be trimmed in its computation. Similar procedures were adopted by Bickel (1982), Kreiss (1987b), Manski (1984), Linton (1993), and Jeganathan (1994), among others, and are also employed here.

Our first step in this section is to formulate a nonparametric kernel estimator of the score $\psi(\varepsilon_t)$. We therefore introduce the following notation:

$$\pi(x,\sigma) = \left(\frac{1}{\left(\sigma\sqrt{2\pi}\right)^q}\right)\exp\left(\frac{-|x|^2}{2\sigma^2}\right),$$

$$\hat{p}_{\sigma,t}(x,\theta) = \frac{1}{2(n-1)}\sum_{\substack{i=1 \\ i\neq t}}^{n}\left\{\pi(x+\varepsilon_i(\theta),\sigma) + \pi(x-\varepsilon_i(\theta),\sigma)\right\}$$

and let $\hat{p}^j_{\sigma,t}(x,\theta)$ be the partial derivative of $\hat{p}_{\sigma,t}(x,\theta)$ with respect to the j^{th} element of x, for all $j=1,\ldots,q$. In these expressions, $\pi(x,\sigma)$ is a q-dimensional normal kernel with smoothing parameter σ. The larger is σ, the smoother is the estimate of p. We further define

$$\hat{\psi}^j_{n,t}(x,\theta) = \begin{cases} \dfrac{\hat{p}^j_{\sigma(n),t}(x,\theta)}{\hat{p}_{\sigma(n),t}(x,\theta)} & if \begin{cases} \hat{p}_{\sigma(n),t}(x,\theta) \geq m^j_n \\ |x| \leq \alpha^j_n \\ \left|\hat{p}^j_{\sigma(n),t}(x,\theta)\right| \leq c^j_n \hat{p}_{\sigma(n),t}(x,\theta) \end{cases} \\ 0 & otherwise \end{cases}$$

where $c^j_n \to \infty$, $\alpha^j_n \to \infty$, $\sigma(n) \to 0$, $m^j_n \to 0$. The trimming parameters c^j_n, α^j_n, and m^j_n serve to omit extreme outlying observations that would distort the behaviour of the score estimator $\hat{\Delta}_n$ given by equation (19) below. Our derivation of an adaptive estimator hinges on the consistency of $\hat{\Delta}_n$ as an estimator of Δ_n (given in (18)). The above conditions on the trimming parameters and on the smoothing parameter σ are used to prove this consistency.

At this point, a problem arises that is common in nonparametric estimation. The theory only describes the limiting behaviour of the smoothing and trimming parameters. In practical problems, with a fixed sample size, knowing this theoretical limiting behaviour provides little assistance in selecting the values to be used. An extensive literature exists regarding the selection of smoothing parameters in density estimation problems (see, for example, Marron (1987) for a survey), but the applicability of this literature to the case at hand has not been much investigated, nor has the problem of trimming parameter selection.

The question of smoothing and trimming parameter selection in the adaptive estimation of (non-cointegrating, single equation) linear regression models was addressed in a Monte Carlo simulation study by Hsieh and Manski (1987), which extends a similar study reported by Manski (1984). Hsieh and Manski used sample sizes of 25 and 50, and considered six possible distributions for the (univariate) errors (normal, variance-contaminated mixture of normals, t, bimodal mixture of normals, beta, and log-normal). In all cases, the standard deviation of the error density was set equal to unity. They found that the adaptive estimator's performance was fairly insensitive to the selection of trimming parameters (although being more sensitive to mild overtrimming than to mild undertrimming). They found that good values of c_n, α_n, and m_n for n=50 were 8, 8, and exp(-32), respectively. Regarding the smoothing parameter $\sigma(n)$, they found that the estimator was quite sensitive to its selection. Depending on the true distribution of the errors, it was found that from a set of preselected possible values of $\sigma(n)$, the best value was anywhere from 0.1 to 0.5 for n=50. It should be emphasized that the range of smoothing and trimming parameter values used by Hsieh and Manski (1987) is roughly appropriate for their simulated data because the innovations had unit standard deviation. In empirical applications, the range of values that an investigator would want to consider will depend upon the scale of the data. The smaller is the scale, the smaller will be the most appropriate smoothing and trimming parameter settings. For example, Silverman's (1986) rule of thumb formula for bandwidth selection in density estimation problems is linear in the standard deviation of the sample, while Hsieh and Manski's (1987) trimming procedure is effective at approximately eight standard deviations from the origin.

Hsieh and Manski (1987) found that estimator performance improved if a data-based bootstrap method was used to select $\sigma(n)$. Their procedure took the preliminary (least squares) parameter estimates as the true values, and the sample distributions of the regressors and the least squares residuals as the population distributions. From this artificial population, they drew a number of independent samples, choosing the value of $\sigma(n)$ that minimized the mean square error of the parameter estimate over these samples. They concluded by recommending the use of such a method in empirical applications and strongly recommending against using preselected values for $\sigma(n)$.

It has been assumed above that the same bandwidth setting is used to estimate both the density and its partial derivatives. However, Hardle, Hart, Marron, and Tsybakov (1992, p.219) suggest that better results may be obtained in practice by using different values of $\sigma(n)$ for the density and derivative estimation problems. For both problems, data-based methods such as cross-validation could conceivably be employed. For density estimation, the literature in this area is extensive. It is much less so for derivative estimation, although Hardle, Marron, and Wand (1990) explore the use of data-dependent methods and prove an optimality result for a least-squares cross-validation procedure. In addition to allowing the bandwidth to vary *between* the density and derivative estimation procedures, we may wish to let it vary *within* each procedure. Since observations are more dense in the body of the distribution than in the tails, it would seem reasonable to smooth more heavily in the latter, and various methods for using variable bandwidths have appeared in the literature. For our problem, increased smoothing in the tails would reduce the need to trim, which arises because undersmoothing causes the density estimate to approach zero more rapidly than the derivative estimate, causing an explosion in the estimated score at outlying observations.

Most work that has been carried out on the selection of smoothing parameters in density estimation problems analyzes univariate densities. In the model considered here, however, a multivariate density is being estimated, which may complicate matters since it may be desirable to employ a $q \times q$ matrix of smoothing parameters rather than a scalar. Alternatively, we may choose to rescale the data prior to employing the kernel. Furthermore, the multivariate character of our analysis increases the number of trimming parameters to be selected from three to $3q$. In practice, it would be simplest to use the same settings of the trimming parameters for all q equations, but there could be a potential accompanying loss of effectiveness in the estimator.

The theory developed in this paper assumes that the values of the smoothing and trimming parameters are selected in advance of the analysis. However, there is strong reason to believe that the performance of the adaptive estimator can be improved by the use of data-based methods to select these parameters. This point implies that it would be desirable to extend the theory to allow for data-dependent smoothing and trimming parameters.

Returning to the analysis, we define the q-dimensional score estimate as $\hat{\psi}_{n,t}(x,\theta) = \left(\hat{\psi}^1_{n,t}(x,\theta),\ldots,\hat{\psi}^q_{n,t}(x,\theta)\right)'$. Assuming that $p(\varepsilon)$ is symmetric about the origin implies that $\psi(\varepsilon)$ is anti-symmetric about the origin. Furthermore, $\hat{\psi}_{n,t}(x,\theta)$ is anti-symmetric about the origin in x by construction. We also define

(18) $$\Delta_n(\theta) = -\sum_{t=1}^{n} U_{nt}(\theta)\psi(\varepsilon_t(\theta))$$

(19) $$\hat{\Delta}_n(\theta) = -\sum_{t=1}^{n} U_{nt}(\theta)\hat{\psi}_{n,t}(\varepsilon_t(\theta),\theta)$$

where $U_{nt}(\theta)$ is some $m \times q$ matrix. This notation is introduced in order to facilitate our statement of Proposition 5.4, which will be used in our derivation of an adaptive estimator. As mentioned earlier, a central problem in constructing an adaptive estimator is the consistent estimation of the score vector for the sample. The discussion leading up to Proposition 5.4 develops the properties that the matrix $U_{nt}(\theta)$ must possess in order for $\hat{\Delta}_n(\theta)$ to be a consistent estimator of $\Delta_n(\theta)$. After these properties have been outlined, it will be shown in Lemma 5.5 that they are indeed satisfied by the matrix $\delta_n H_{t-1}(\theta)$. Therefore, after stating Proposition 5.4, the setting of $U_{nt}(\theta) = \delta_n H_{t-1}(\theta)$ will be used to show that $\hat{\Delta}_n(\theta)$ is a consistent estimator of the score vector $\Delta_n(\theta)$. This will allow the derivation of an expression for an adaptive estimator in Theorem 5.6. First, the desired properties of the matrix $U_{nt}(\theta)$ are stated in Conditions 5.2 and 5.3.

Define

$$\Delta_n^j(\theta) = -\sum_{t=1}^{n} U_{nt}^j(\theta)\psi^j(\varepsilon_t(\theta))$$
$$\hat{\Delta}_n^j(\theta) = -\sum_{t=1}^{n} U_{nt}^j(\theta)\hat{\psi}_{n,t}^j(\varepsilon_t(\theta),\theta)$$

where $U_{nt}^j(\theta)$ is the j^{th} column of $\psi^j(\varepsilon_t(\theta))$ and $\psi^j(\varepsilon_t(\theta))$ is the j^{th} element of $\psi(\varepsilon_t(\theta))$. We seek to prove that

$$\hat{\Delta}_n(\theta_n) - \Delta_n(\theta_n) = o_p(1) \;\; in \; P_{\theta,n},$$

for which it is sufficient to show that

$$(20) \qquad \hat{\Delta}_n^j(\theta_n) - \Delta_n^j(\theta_n) = o_p(1) \;\; in \, P_{\theta,n} \;\; \forall j = 1,\ldots,q,$$

because

$$\hat{\Delta}_n(\theta_n) - \Delta_n(\theta_n) = \sum_{j=1}^{q}\left(\hat{\Delta}^j_n(\theta_n) - \Delta^j_n(\theta_n)\right).$$

The consistency result (20) is a consequence of Proposition 5.4, the statement of which requires our introduction of the following three conditions.

<u>*Condition 5.1*</u> *(condition (28) in Jeganathan (1988, p. 35)):*

$$0 < \int_{-\infty}^{\infty}\left[\frac{p^j(x)}{p(x)}\right]^2 p(x)dx < \infty.$$

<u>*Condition 5.2*</u> *(condition (B.2) in Jeganathan(1988, pp. 38-39)):*

Verify that there is a suitable sequence $\{\delta_n\}$ *of normalizing matrices such that for every bounded* $\{h_n\}$ *(where* $\theta_n = \theta + \delta_n h_n$*) it holds that, for all* $j=1,\dots,q$:

(21) $$\sum_{t=1}^{n}\left[\left(g^j_{t-1}(\theta_n) - g^j_{t-1}(\theta)\right) - h_n' U^j_{nt}(\theta)\right]^2 = o_p(1) \text{ in } P_{\theta,n}$$

(where the superscript denotes the j^{th} *element of the vector g) for the m-vectors* $U^j_{n1},\dots,U^j_{nn}$ *such that*

(22) $$\sum_{t=1}^{n}\left|h_n' U^j_{nt}(\theta)\right|^2 = O_p(1) \text{ in } P_{\theta,n}, \text{ and}$$

(23) $$\max_{t\in\{1,\dots,n\}}\left|h_n U^j_{nt}(\theta)\right|^2 = o_p(1) \text{ in } P_{\theta,n}.$$

<u>*Condition 5.3*</u> *(condition (B.3) in Jeganathan (1988, pp. 44-45)):*

Verify that, for every $j = 1,\ldots,q$, *there are w-vectors* $V_{nt}^{j}(\theta), t = 1,\ldots,n,$ *and non-random (m-w)-vectors* $R_{nt}^{j}(\theta), t = 1,\ldots,n,$ *such that for every bounded* $\{h_n\}$ *and for every u, it holds that*

$$(24) \qquad \sum_{t=1}^{n} \left| u' \, U_{nt}^{j}(\theta_n) - u' \begin{bmatrix} V_{nt}^{j}(\theta) \\ R_{nt}^{j}(\theta) \end{bmatrix} \right|^2 = o_p(1) \ \ in \ P_{\theta,n}$$

and, for some $\delta \in [0,1)$,

$$(25) \qquad \max_{t \in \{1,\ldots,n\}} n \left| V_{nt}^{j}(\theta) \right|^2 = O_p(n^{\delta}) \ \ in \ P_{\theta,n}.$$

Our proof of adaptive estimation uses the following proposition (based on Proposition 15 of Jeganathan (1988, pp. 46-50)):

<u>Proposition 5.4</u>: *For every* $j = 1,\ldots,q$, *assume that Conditions 5.1, 5.2, and 5.3 hold, and that* $c_n^j \to \infty$, $\alpha_n^j \to \infty$, $m_n^j \to 0$, $\sigma(n) \to 0$, $\sigma(n) c_n^j \to 0$, *and* $n^{-(1-\delta)} \alpha_n^j \sigma(n)^{-(4+q)} \to 0$, *with* δ *as in (25). Furthermore, assume that* $p(\varepsilon)$ *is symmetric about the origin and that the sequences* $\{P_{\theta,n}\}$ *and* $\{P_{\theta_n,n}\}$ *are contiguous for every bounded* $\{h_n\}$. *Then, for every bounded* $\{h_n\}$, *it follows that, for every* $j = 1,\ldots,q$,

$$(26) \qquad \hat{\Delta}_n^j(\theta_n) - \Delta_n^j(\theta_n) = o_p(1) \ \ in \ P_{\theta,n}.$$

It follows from (26) that

(27) $\hat{\Delta}_n(\theta_n) - \Delta_n(\theta_n) = o_p(1)$ *in* $P_{\theta,n}$.

Proposition 5.4 shows that $\hat{\Delta}_n(\theta)$ is a consistent estimator of $\Delta_n(\theta)$ when the matrix $U_{nt}(\theta)$ satisfies the conditions laid out above. We can use this result to derive an adaptive estimator for our model by showing that $\delta_n H_{t-1}(\theta)$ satisfies the conditions specified for $U_{nt}(\theta)$. We do so in Lemma 5.5.

<u>Lemma 5.5</u>: *Setting* $U_{nt}(\theta) = \delta_n H_{t-1}(\theta)$, *Conditions 5.1-5.3 are satisfied by our model.*

With Proposition 5.4 and Lemma 5.5 together showing that the score for our model is consistently estimable, we are nearly ready to propose an adaptive estimator for θ. We need only find a consistent estimate of the scaled sample information matrix $S_n(\theta)$. Recall that

$$\begin{aligned} S_n(\theta) &= \sum_{t=1}^{n} \delta_n H_{t-1}(\theta) \Omega H_{t-1}(\theta)' \delta_n \\ &= \sum_{j=1}^{q} \sum_{\ell=1}^{q} \sum_{t=1}^{n} \delta_n H_{t-1}^{j}(\theta) H_{t-1}^{\ell}(\theta)' \delta_n \omega_{j\ell} \\ &= \sum_{j=1}^{q} \sum_{\ell=1}^{q} \omega_{j\ell} S_n^{j\ell}(\theta)' \end{aligned}$$

where $S_n^{j\ell}(\theta) = \sum_{t=1}^{n} \delta_n H_{t-1}^{j}(\theta) H_{t-1}^{\ell}(\theta)' \delta_n$.

For the moment, we shall assume the existence of a consistent estimate $\hat{\omega}_{j\ell}$ of $\omega_{j\ell}$ (the construction of such an estimate is discussed below). Now, given a δ_n^{-1}-

consistent sequence $\{\theta_n\}$, we have $S_n^{j\ell}(\theta_n) - S_n^{j\ell}(\theta) = o_p(1)$ in $P_{\theta,n}$ for every $j, \ell = 1, \ldots, q$, the result holding because $S_n^{j\ell}(\theta_n) \Rightarrow S^{j\ell}(\theta)$ and $S_n^{j\ell}(\theta) \Rightarrow S^{j\ell}(\theta)$.

We now define the estimate

$$\hat{S}_n(\theta_n) = \sum_{j=1}^{q} \sum_{\ell=1}^{q} \hat{\omega}_{j\ell} S_n^{j\ell}(\theta_n).$$

From the preceding discussion, it follows that

(28) $\quad \hat{S}_n(\theta_n) - S_n(\theta) = o_p(1)$ *in* $P_{\theta,n}$.

Using (27), (28), and arguments in Jeganathan (1994), we can show that

(29) $\quad \hat{\Delta}_n(\theta_n) = \hat{\Delta}_n(\theta) - S_n(\theta) h_n + o_p(1)$ *in* $P_{\theta,n}$.

The above results can be used to derive an adaptive estimator for θ, as shown in the following theorem:

<u>Theorem 5.6</u>: The estimator $\tilde{\theta}_n$ *given in (30) is adaptive for our model:*

(30) $\quad \tilde{\theta}_n = \theta_n^{**} + \delta_n \hat{S}_n^{-1}(\theta_n^{**}) \hat{\Delta}_n(\theta_n^{**})$.

In other words,

(31) $\quad \delta_n^{-1}(\hat{\theta}_n - \tilde{\theta}_n) = o_p(1)$ *in* $P_{\theta,n}$,

where $\hat{\theta}_n$ *is given by (15).*

Remarks:

(a) In deriving our estimate $\hat{S}_n(\theta_n)$ of $S_n(\theta)$, we assumed the existence of a consistent estimator of $\omega_{j\ell}$. The argument of Kreiss (1987b, p. 123) can be used to show that $\hat{\omega}_{j\ell}$ as defined in (32) below is such an estimator (the proof uses the fact that $\{P_{\theta,n}\}$ and $\{P_{\theta_n,n}\}$ are contiguous):

$$(32) \qquad \hat{\omega}_{j\ell} = \frac{1}{n}\sum_{t=1}^{n} \hat{\psi}^{j}_{n,t}\left(\varepsilon_t(\theta_n^{**}), \theta_n^{**}\right)\hat{\psi}^{\ell}_{n,t}\left(\varepsilon_t(\theta_n^{**}), \theta_n^{**}\right).$$

(b) We now derive the asymptotic distribution of the adaptive estimator $\tilde{\theta}_n$. Since $\tilde{\theta}_n$ is an AC estimator, we can use the definition of the latter to obtain

$$\delta_n^{-1}(\tilde{\theta}_n - \theta) = S_n^{-1}(\theta)W_n(\theta) + o_p(1) \ \ in\ P_{\theta,n}.$$

We showed above that

$$L\left(S_n^{-1}(\theta)W_n(\theta)\middle|P_{\theta,n}\right) \Rightarrow L\left(S(\theta)^{-1/2}N(0,I)\right),$$

so it follows that

$$(33) \qquad L\left(\delta_n^{-1}(\tilde{\theta}_n - \theta)\middle|P_{\theta,n}\right) \Rightarrow L\left(MN\left(0, S^{-1}(\theta)\right)\right),$$

where $S(\theta)$ is the block diagonal Fisher information matrix. It therefore follows that

(34) $$L\left(\sqrt{n}(\tilde{\eta}_n - \eta)\middle|P_{\theta,n}\right) \Rightarrow L\left(N\left(0, \Omega^{-1} \otimes \left(E[M_t M_t']\right)^{-1}\right)\right)$$

and

(35) $$L\left(n(\tilde{\beta}_n - \beta)\middle|P_{\theta,n}\right) \Rightarrow L\left(MN\left(0, (A'\,\Omega A)^{-1} \otimes \left(\int_0^1 B_2 B_2'\right)^{-1}\right)\right).$$

(c) The asymptotic efficiency gains to be obtained from employing the adaptive estimator developed here rather than the Gaussian pseudo-MLE when $p(\varepsilon)$ is not Gaussian can be investigated using (35). The covariance matrix of the scaled and centered Gaussian pseudo-MLE is

(36) $$\left(A'\,\Sigma_\varepsilon^{-1} A\right)^{-1} \otimes \left(\int_0^1 B_2 B_2'\right)^{-1}.$$

If $p(\varepsilon)$ actually is Gaussian, then $\Sigma_\varepsilon^{-1} = \Omega$, so that (36) is identical to the covariance matrix in (35). However, if Gaussianity fails, then $\Omega > \Sigma_\varepsilon^{-1}$, so that $\left(A'\,\Sigma_\varepsilon^{-1} A\right)^{-1} \geq (A'\,\Omega A)^{-1}$ and the Gaussian estimator is inefficient. The degree of inefficiency can be measured using the following ratio of generalized variances (cf. Mitchell (1989)), where ER is mnemonic for "efficiency ratio":

$$ER = \frac{\left[\det\left[(A'\,\Omega A)^{-1} \otimes \left(\int_0^1 B_2 B_2'\right)^{-1}\right]\right]^{1/r(q-r)}}{\left[\det\left[\left(A'\,\Sigma_\varepsilon^{-1} A\right)^{-1} \otimes \left(\int_0^1 B_2 B_2'\right)^{-1}\right]\right]^{1/r(q-r)}}$$

$$(37) \qquad = \frac{\left[\det\left(A'\,\Sigma_\varepsilon^{-1} A\right)\right]^{(q-r)/r(q-r)} \left[\det\left(\int_0^1 B_2 B_2'\right)\right]^{r/r(q-r)}}{\left[\det(A'\,\Omega A)\right]^{(q-r)/r(q-r)} \left[\det\left(\int_0^1 B_2 B_2'\right)\right]^{r/r(q-r)}}$$

$$= \left[\frac{\det\left(A'\,\Sigma_\varepsilon^{-1} A\right)}{\det(A'\,\Omega A)}\right]^{1/r}.$$

Now suppose, for example, that $p(\varepsilon)$ is elliptically symmetric, with characteristic function $cf(s) = \phi(s'\,\Sigma s)$, where $\Sigma_\varepsilon = k_\varepsilon \Sigma$, with $k_\varepsilon = -2\phi'(0)$, and where

$$(38) \qquad p(\varepsilon) = \left|\det \Sigma\right|^{-1/2} f^*\left(\varepsilon'\,\Sigma^{-1}\varepsilon\right).$$

Then, as shown by Mitchell (1989),

$$(39) \qquad \Omega = 4a_p k_\varepsilon \Sigma_\varepsilon^{-1},$$

where a_p is defined in Mitchell (1989, p.296). (In the Gaussian case, $k_\varepsilon = 1$ and $a_p = 1/4$, giving us the familiar result that $\Omega = \Sigma_\varepsilon^{-1}$.) Substituting (39) into the last line of (37) gives us

$$ER = \left[\frac{\det\left(A' \Sigma_\varepsilon^{-1} A\right)}{\det\left(4 a_p k_\varepsilon A' \Sigma_\varepsilon^{-1} A\right)}\right]^{1/r} = \frac{1}{4 a_p k_\varepsilon}. \tag{40}$$

This result is interesting because the ratio obtained is identical to that derived by Mitchell (1989) for the estimation of the location of a distribution from which a sequence of iid observations are drawn. Thus, the efficiency gains to be obtained through maximum likelihood estimation of the nonstationary components of a reduced rank VAR are identical to those to be obtained through maximum likelihood estimation in a very wide range of stationary and nonstationary models.

Mitchell (1989) illustrates this ratio for the case where $p(\varepsilon)$ has a t-distribution with ν degrees of freedom. She finds that $k_\varepsilon = \nu / (\nu - 2)$ and $a_p = (\nu + q) / 4(\nu + q + 2)$, so that

$$ER = (1 - 2/\nu) \cdot (1 + 2/(\nu + q)).$$

Thus, the asymptotic efficiency gains to be obtained from using the adaptive estimator are increasing in q, the number of variables in the model.

(d) The preceding remark suggests that our adaptive estimator will provide better relative performance the larger is the system. In finite samples, however, its performance will worsen with increasing q. This is because we have used kernel methods to estimate a density of dimension q. From a computational standpoint, kernel estimators can perform poorly when q is large. For sample sizes typical in econometrics, q need only equal three or four for a kernel estimator to give inaccurate results.

If we assume elliptical symmetry, however, this dimensionality problem can be alleviated. This is because, as equation (38) shows, one property of elliptically symmetric densities is that they can be expressed as a function of a scalar random variable, where the latter is a quadratic term in the underlying vector-valued random variable (see Fang, Kotz, and Ng (1990, p. 46)). We could therefore proceed by using a normal kernel estimator such as the one used in the multivariate case to get an estimate $\hat{f}^*$ of f^* in (38). We could then estimate $\hat{\psi}$ as above, using the derivative $2\Sigma^{-1}\varepsilon\hat{f}^{*\prime}$. Proposition 5.4 would still apply, with the condition $n^{-(1-\delta)}\alpha_n^j\sigma(n)^{-(4+q)} \to 0$ being replaced by $n^{-(1-\delta)}\alpha_n^j\sigma(n)^{-5} \to 0$. In doing the computations, Σ could be replaced with a consistent preliminary estimate.

VI . INCLUSION OF INTERCEPTS IN THE COINTEGRATING RELATIONS

The model considered above is entirely stochastic, which is clearly a limitation since most econometric models contain deterministic components. Especially common is the presence of non-zero intercepts in the cointegrating relations. In this section, we extend the model to allow for intercepts, but still maintain the restriction that none of the time series contain deterministic drift. Johansen (1994) investigates the various possible ways in which deterministic components may enter a reduced rank regression model. We consider only the model that he denotes by $H_1^*(r)$.

Maintaining our earlier notation, we may now write the model as

$$\text{(3')} \qquad \Delta X_t = A\left[X_{1,t-1} - B_1 - B_0 X_{2,t-1}\right] + \Phi Z_{t-1} + \varepsilon_t,$$

where B_1 is the r-vector of constants in the cointegrating vectors. By now allowing h_n to have $m+r$ elements, redefining δ_n as $diag\left[n^{-1/2} I_{qr+q^2(k-1)+r}, n^{-1} I_{r(q-r)}\right]$ and $H_{t-1}(\theta)$ as $\left[(I_q \otimes W_{t-1})', (I_q \otimes Z_{t-1})', -A, (-A' \otimes X_{2,t-1})'\right]'$, we obtain the following generalization of Theorem 3.2:

<u>*Proposition 6.1*</u>*: The likelihood ratios $\Lambda_n(\theta_n, \theta)$ have the following quadratic approximation:*

$$\text{(6')} \qquad \begin{aligned} \Lambda_n(\theta_n,\theta) &= -\sum_{t=1}^{n} h_n' \delta_n H_{t-1}(\theta) \psi(\varepsilon_t) \\ &\quad - \frac{1}{2}\sum_{t=1}^{n} h_n' \delta_n H_{t-1}(\theta) \Omega H_{t-1}(\theta)' \delta_n h_n + o_p(1) \; in \; P_{\theta,n}. \end{aligned}$$

Using analogous definitions for $W_n(\theta)$ and $S_n(\theta)$ to those given in (7) and (8), our LAN/LAMN limit theory will hold, with $S_n(\theta) \Rightarrow S(\theta)$, where

$$S(\theta)=\begin{bmatrix}\Omega\otimes E[M_tM_t'] & 0 & 0\\ 0 & A'\Omega A & A'\Omega A\otimes\overline{B}_2'\\ 0 & A'\Omega A\otimes\overline{B}_2 & A'\Omega A\otimes\int_0^1 B_2B_2'\end{bmatrix},$$

and $\overline{B}_2=\int_0^1 B_2$. We can then obtain an adaptive estimator $\widetilde{\theta}_n$ as in (30), with the asymptotic distribution:

(33') $$L\left(\delta_n^{-1}(\widetilde{\theta}_n-\theta)\middle|P_{\theta,n}\right)\Rightarrow L\left(MN(0,S(\theta)^{-1})\right),$$

so that

(34') $$L\left(\sqrt{n}(\widetilde{\eta}_n-\eta)\middle|P_{\theta,n}\right)\Rightarrow L\left(N\left(0,\Omega^{-1}\otimes E[M_tM_t']^{-1}\right)\right),$$

(35') $$L\left(n(\widetilde{B}_n-B)\middle|P_{\theta,n}\right)\Rightarrow L\left(MN\left(0,(A'\Omega A)^{-1}\otimes\left(\int_0^1 B_2B_2'-\overline{B}_2\overline{B}_2'\right)^{-1}\right)\right),$$

and

(41) $$L\left(\sqrt{n}(\widetilde{B}_{1n}-B_1)\middle|P_{\theta,n}\right)\Rightarrow L\left(MN\left(0,(A'\Omega A)^{-1}\left(1+\overline{B}_2'\left(\int_0^1 B_2B_2'-\overline{B}_2\overline{B}_2'\right)\overline{B}_2\right)\right)\right).$$

We can see from these results that not knowing the true value of the intercept vector adversely affects our ability to estimate the slope parameters B, even asymptotically. This point is illustrated by the fact that the asymptotic information matrix is not block diagonal and by the fact that the covariance matrix of $\widetilde{B}_n$ as given in (35') differs from that given in (35) - the case where the intercept is known - by a positive definite matrix with probability one.

VII. CONCLUSIONS

In this paper we have demonstrated that reduced rank error correction models can be adaptively estimated, assuming that the innovations in the underlying VAR are drawn from a symmetric density function. We have shown how to construct consistent nonparametric estimates of the score function of the unknown density of the innovations, and we have demonstrated that the asymptotic efficiency gains to be obtained from employing the adaptive estimator rather than a Gaussian pseudo-MLE are identical to those obtained in an extremely broad class of statistical and econometric models, including the most basic location parameter problem.

As they stand, the theory and methods developed here should be of substantial value to practitioners. Nevertheless, further developments would be desirable. The relaxation of the symmetry assumption is one direction in which the generality of the analysis could be increased significantly. Conversely, in cases where elliptical symmetry is a reasonable assumption, further investigation of techniques of kernel estimation to reduce a multidimensional density estimation problem to a one-dimensional problem would undoubtedly produce improved estimators for large systems. Finally, the range of empirical situations to which the methodology is applicable would also be increased by generalizing the analysis to allow for various possible specifications of deterministic components, including the case of drifting variables.

APPENDIX

Throughout the Appendix, we simplify notation by writing $g_{t-1}(\theta)$ in place of $g_t(\underline{X}_{t-1}, \theta)$.

Proof of Lemma 3.1: We decompose h_n into components of respective dimension qr, $q^2(k-1)$, and $r(q-r)$ by writing $h_n = (h_{\alpha n}', h_{\varphi n}', h_{\beta n}')'$. These components can be thought of as vectorizations of the matrices a_n, ϕ_n, and b_n, whose respective dimensions are $q \times r$, $q \times q(k-1)$, and $r \times (q-r)$. Using this notation, and (3), we have

$$g_{t-1}(\theta) = X_{t-1} + A\left[X_{1,t-1} - B_0 X_{2,t-1}\right] + \Phi Z_{t-1} \tag{42}$$

and

$$\begin{aligned} g_{t-1}(\theta_n) &= X_{t-1} + \left(A + \frac{1}{\sqrt{n}} a_n\right)\left[X_{1,t-1} - \left(B_0 + \frac{1}{n} b_n\right) X_{2,t-1}\right] + \left(\Phi + \frac{1}{\sqrt{n}} \phi_n\right) Z_{t-1} \\ &= g_{t-1}(\theta) + \frac{1}{\sqrt{n}} a_n\left[X_{1,t-1} - B_0 X_{2,t-1}\right] - \frac{1}{n} A b_n X_{2,t-1} \\ &\quad - \frac{1}{n^{3/2}} a_n b_n X_{2,t-1} + \frac{1}{\sqrt{n}} \phi_n Z_{t-1}. \end{aligned} \tag{43}$$

Subtracting (42) from (43) yields:

$$d_t(\theta_n, \theta) = \frac{1}{\sqrt{n}} a_n W_{t-1} + \frac{1}{\sqrt{n}} \phi_n Z_{t-1} - \frac{1}{n} A b_n X_{2,t-1} - \frac{1}{n^{3/2}} a_n b_n X_{2,t-1}. \tag{44}$$

Noting that $d_t(\theta_n, \theta)$ is a vector, and applying the formula for vectorizing products of matrices, we can rewrite (44) as follows:

$$
\begin{aligned}
d_t(\theta_n,\theta)' &= h_{\alpha n}'\frac{1}{\sqrt{n}}\left(I_q \otimes W_{t-1}\right) + h_{\varphi n}'\frac{1}{\sqrt{n}}\left(I_q \otimes Z_{t-1}\right) \\
&\quad - h_{\beta n}'\frac{1}{n}\left(A' \otimes X_{2,t-1}\right) - \frac{1}{n^{3/2}} X_{2,t-1}' b_n' a_n' \\
&= \left(h_{\alpha n}', h_{\varphi n}', h_{\beta n}'\right)\delta_n \begin{bmatrix} I_q \otimes W_{t-1} \\ I_q \otimes Z_{t-1} \\ -A' \otimes X_{2,t-1} \end{bmatrix} - \frac{1}{n^{3/2}} X_{2,t-1}' b_n' a_n',
\end{aligned}
$$

from which (2) immediately follows. •

<u>Proof of Theorem 3.2</u>: We begin by quoting Conditions A.1-A.5 and Proposition A.6 below, as stated by Jeganathan (1994). According to Proposition A.6, the likelihood ratios $\Lambda_n(\theta_n, \theta)$ have an asymptotic quadratic approximation if Conditions A.1-A.5 hold. The proof then proceeds in two steps, the first showing that the approximation given by Proposition A.6 is asymptotically equivalent to that given by (6), and the second showing that Conditions A.1-A.5 are satisfied for our model.

Conditions A.1-A.5 and Proposition A.6 are as follows:

<u>Condition A.1</u>: *The density $p(\varepsilon)$ is absolutely continuous with respect to Lebesgue measure in ε.*

<u>Condition A.2</u>: *The partial derivative vector $\frac{\partial p(\varepsilon)}{\partial \varepsilon}$ exists.*

<u>Condition A.3</u>: If $\{\theta_n\} \in \Theta$ is a sequence such that

(45) $$\sum_{t=1}^{n} E\left[\left|d_t'(\theta_n,\theta)\psi(\varepsilon_t)\right|^2 \middle| F_{t-1}\right] = O_p(1) \quad in\ P_{\theta,n},$$

then the quantities

(46) $$\sum_{t=1}^{n} \int \int_0^1 \left|d_t'(\theta_n,\theta)\left[\psi^*(\varepsilon - \kappa d_t(\theta_n,\theta)) - \psi^*(\varepsilon)\right]\right|^2 d\varepsilon d\kappa = o_p(1) \quad in\ P_{\theta,n},$$

where $\psi^*(\varepsilon) = \frac{\partial p(\varepsilon)}{\partial \varepsilon} \Big/ \sqrt{p(\varepsilon)}$, *and,* $\forall \omega > 0$,

(47) $$\sum_{t=1}^{n} E\left[\left|d_t'(\theta_n,\theta)\psi(\varepsilon_t)\right|^2 I\left(\left|d_t'(\theta_n,\theta)\psi(\varepsilon_t)\right| > \omega\right) \middle| F_{t-1}\right] = o_p(1) \quad in\ P_{\theta,n}.$$

<u>Condition A.4</u>: $E\left[\psi(\varepsilon_t) \middle| F_{t-1}\right] = 0,\ t \geq 1.$

<u>Condition A.5</u>: $f_0(\underline{X}_0,\theta_n) - f_0(\underline{X}_0,\theta) = o_p(1)$ *in* $P_{\theta,n}$ *as* $\theta_n \to \theta$.

<u>Proposition A.6</u> (Theorem 11 in Jeganathan (1994)):

Assume that Conditions A.1-A.5 hold. Then, for every $\{\theta_n\}$ such that (45) holds, we have

$$
(48) \qquad \begin{aligned} &\Lambda_n(\theta_n,\theta) - \Bigg\{ -\sum_{t=1}^{n} \big(g_{t-1}(\theta_n) - g_{t-1}(\theta)\big)' \psi(\varepsilon_t) \\ &\quad - \frac{1}{2}\sum_{t=1}^{n} E\Big[\big|\big(g_{t-1}(\theta_n) - g_{t-1}(\theta)\big)' \psi(\varepsilon_t)\big|^2 \big| F_{t-1}\Big]\Bigg\} \\ &= o_p(1) \ \ in\ P_{\theta,n}. \end{aligned}
$$

Lemma A.7 establishes the aforementioned asymptotic equivalence between (48) and (6). Combined with Lemma A.10, which shows that Conditions A.1-A.5 hold for our model, it proves the theorem. Lemmas A.8 and A.9 are used in the proof of Lemma A.10.

<u>Lemma A.7</u>: *For our model, the asymptotic approximation (6) follows from (48).*

<u>Proof</u>: Since the expression $d_t(\theta_n,\theta)$ appears in (48), we can substitute (5) into (48) to obtain

$$
(49) \qquad \begin{aligned} \Lambda_n(\theta_n,\theta) = &-\sum_{t=1}^{n} h_n' \delta_n H_{t-1}(\theta)\psi(\varepsilon_t) + \sum_{t=1}^{n} \frac{1}{n^{3/2}} X_{2,t-1}' b_n' a_n' \psi(\varepsilon_t) \\ &- \frac{1}{2}\sum_{t=1}^{n} E\left[\left|\left(h_n' \delta_n H_{t-1}(\theta) - \frac{1}{n^{3/2}} X_{2,t-1}' b_n' a_n'\right)\psi(\varepsilon_t)\right|^2 \left| F_{t-1}\right.\right] + o_p(1) \ \ in\ P_{\theta,n}. \end{aligned}
$$

To prove the Lemma, we will shall establish that all terms involving the quantity $n^{-3/2} X_{2,t-1}' b_n' a_n'$ are asymptotically negligible. This entails proving that the right-hand side of (49) is equal to

$$
(50) \qquad -\sum_{t=1}^{n} h_n' \delta_n H_{t-1}(\theta)\psi(\varepsilon_t) - \frac{1}{2}\sum_{t=1}^{n} E\Big[\big|h_n' \delta_n H_{t-1}(\theta)\psi(\varepsilon_t)\big|^2 \big| F_{t-1}\Big] + o_p(1) \ \ in\ P_{\theta,n}.
$$

The proof that (50) equals the right-hand side of (49) proceeds in two steps. Step (i) shows that the second term on the right-hand side of (49) is $o_p(1)$, and step (ii) shows that the second component inside the parentheses in the quadratic term can be ignored asymptotically.

Step (i): We seek to prove that

(51) $$\frac{1}{n^{3/2}}\sum_{t=1}^{n} X_{2,t-1}'\, b_n'\, a_n'\, \psi(\varepsilon_t) = o_p(1) \ \ in\ P_{\theta,n}.$$

Since each term in this sum is a scalar, the sum is equal to its own *vec*, allowing us to rewrite the left-hand side of (51) as

$$\frac{1}{n^{3/2}}\sum_{t=1}^{n}\left(X_{2,t-1}'\otimes\psi(\varepsilon_t)\right)vec\left(b_n'\, a_n'\right)$$
$$= vec\left(b_n'\, a_n'\right)'\left[\frac{1}{n^{3/2}}\sum_{t=1}^{n} X_{2,t-1}\otimes\psi(\varepsilon_t)'\right].$$

This expression is $o_p(1)$ in $P_{\theta,n}$ because

$$\frac{1}{n}\sum_{t=1}^{n} X_{2,t-1}\otimes\psi(\varepsilon_t)' \Rightarrow \int_0^1 B_2\otimes dB_1' ,$$

where B_1 and B_2 are independent Brownian motions with respective covariance matrices of Ω and $P_{21}\Psi_{11}\Omega_{a_1}\Psi_{11}'P_{21}'$. The latter matrix is derived on p.818 of Ahn and Reinsel (1990) and is stated in terms of the notation of that paper. This completes step (i).

Step (ii): We now consider the quadratic term in (49), and prove that it can be simplified as follows:

$$(52)\quad \sum_{t=1}^{n} E\left[\left|\left(h_n'\,\delta_n H_{t-1}(\theta) - \frac{1}{n^{3/2}} X_{2,t-1}'\,b_n'\,a_n'\right)\psi(\varepsilon_t)\right|^2 \middle| F_{t-1}\right]$$
$$= \sum_{t=1}^{n} E\left[\left|h_n'\,\delta_n H_{t-1}(\theta)\psi(\varepsilon_t)\right|^2 \middle| F_{t-1}\right] + o_p(1) \ \ in\ P_{\theta,n}.$$

This step involves proving the following two results:

$$(53)\quad \frac{1}{n^{3/2}}\sum_{t=1}^{n} E\left[h_n'\,\delta_n H_{t-1}(\theta)\psi(\varepsilon_t) X_{2,t-1}'\,b_n'\,a_n'\,\psi(\varepsilon_t) \middle| F_{t-1}\right] = o_p(1)\ \ in\ P_{\theta,n},\ \text{and}$$

$$(54)\quad \frac{1}{n^{3}}\sum_{t=1}^{n} E\left[\left(X_{2,t-1}'\,b_n'\,a_n'\,\psi(\varepsilon_t)\right)^2 \middle| F_{t-1}\right] = o_p(1)\ \ in\ P_{\theta,n}.$$

Since (54) will follow as a consequence of (53), the remainder of this step proves (53). We can use the facts that

$$h_n'\,\delta_n H_{t-1}(\theta) = \frac{1}{\sqrt{n}} W_{t-1}'\,a_n' + \frac{1}{\sqrt{n}} Z_{t-1}'\,\phi_n' - \frac{1}{n} X_{2,t-1}'\,b_n'\,A',$$

and that $X_{2,t-1}'\,b_n'\,a_n'\,\psi(\varepsilon_t)$ is a scalar, and so equals its own transpose, to rewrite (53) as follows:

$$
(55)\quad \begin{aligned}
&\frac{1}{n^2}\sum_{t=1}^{n} E\left[W_{t-1}'\, a_n'\, \psi(\varepsilon_t)\psi(\varepsilon_t)'\, a_n b_n X_{2,t-1} \middle| F_{t-1}\right] \\
&+\frac{1}{n^2}\sum_{t=1}^{n} E\left[Z_{t-1}'\, \phi_n'\, \psi(\varepsilon_t)\psi(\varepsilon_t)'\, a_n b_n X_{2,t-1} \middle| F_{t-1}\right] \\
&-\frac{1}{n^{5/2}}\sum_{t=1}^{n} E\left[X_{2,t-1}'\, b_n'\, A'\, \psi(\varepsilon_t)\psi(\varepsilon_t)'\, a_n b_n X_{2,t-1} \middle| F_{t-1}\right].
\end{aligned}
$$

We now show that each term in (55) is $o_p(1)$ in $P_{\theta,n}$. The first term can be rewritten as

$$
\begin{aligned}
&\frac{1}{n^2}\sum_{t=1}^{n} W_{t-1}'\, a_n'\, E\left[\psi(\varepsilon_t)\psi(\varepsilon_t)' \middle| F_{t-1}\right] a_n b_n X_{2,t-1} \\
&= \frac{1}{n^2}\sum_{t=1}^{n} W_{t-1}'\, a_n'\, \Omega a_n b_n X_{2,t-1} \\
&= \frac{1}{n^2}\sum_{t=1}^{n} \left(W_{t-1}' \otimes X_{2,t-1}'\right) vec\left(a_n'\, \Omega a_n b_n\right) \\
&= vec\left(a_n'\, \Omega a_n b_n\right)\left[\frac{1}{n^2}\sum_{t=1}^{n} \left(W_{t-1} \otimes X_{2,t-1}\right)\right] \\
&= o_p(1) \quad in\ P_{\theta,n},
\end{aligned}
$$

since

$$
\frac{1}{n}\sum_{t=1}^{n} \left(W_{t-1} \otimes X_{2,t-1}\right) \Rightarrow \int_0^1 dB_W \otimes B_2 + const,
$$

where B_W is a Brownian motion with covariance matrix $E\left[W_t W_t'\right]$ and the constant depends on the correlation between B_W and B_2.

The second term in (55) is

$$\frac{1}{n^2}\sum_{t=1}^{n} Z_{t-1}'\,\phi_n'\,E\left[\psi(\varepsilon_t)\psi(\varepsilon_t)'\,|F_{t-1}\right]a_n b_n X_{2,t-1}$$
$$=\frac{1}{n^2}\sum_{t=1}^{n} Z_{t-1}'\,\phi_n'\,\Omega a_n b_n X_{2,t-1}$$
$$=\frac{1}{n^2}\sum_{t=1}^{n}\left(Z_{t-1}'\otimes X_{2,t-1}'\right)vec\left(\phi_n'\,\Omega a_n b_n\right)$$
$$=vec\left(\phi_n'\,\Omega a_n b_n\right)'\left[\frac{1}{n^2}\sum_{t=1}^{n}\left(Z_{t-1}\otimes X_{2,t-1}\right)\right]$$
$$=o_p(1)\ \ in\, P_{\theta,n},$$

since

$$\frac{1}{n}\sum_{t=1}^{n}\left(Z_{t-1}\otimes X_{2,t-1}\right)\Rightarrow\int_0^1 dB_Z\otimes B_2+const,$$

where B_Z is a Brownian motion with covariance matrix $E\left[Z_t Z_t'\right]$ and the constant depends on the correlation between B_Z and B_2.

Finally, we rewrite the negative of the third term in (55) as:

$$\frac{1}{n^{5/2}}\sum_{t=1}^{n} X_{2,t-1}'\,b_n'\,A'\,E\left[\psi(\varepsilon_t)\psi(\varepsilon_t)'\,|F_{t-1}\right]a_n b_n X_{2,t-1}$$
$$=\frac{1}{n^{5/2}}\sum_{t=1}^{n} X_{2,t-1}'\,b_n'\,A'\,\Omega a_n b_n X_{2,t-1}$$
$$=\frac{1}{n^{5/2}}\sum_{t=1}^{n}\left(X_{2,t-1}'\otimes X_{2,t-1}'\right)vec\left(b_n'\,A'\,\Omega a_n b_n\right)$$
$$=vec\left(b_n'\,A'\,\Omega a_n b_n\right)'\left[\frac{1}{n^{5/2}}\sum_{t=1}^{n}\left(X_{2,t-1}\otimes X_{2,t-1}\right)\right]$$
$$=o_p(1)\ \ in\, P_{\theta,n},$$

since

$$\frac{1}{n^2}\sum_{t=1}^{n}\left(X_{2,t-1}\otimes X_{2,t-1}\right)\Rightarrow\int_0^1 B_2\otimes B_2 .$$

This completes step (ii).

We have proved that the right-hand side of (49) is equal to (50), so that Proposition A.6 yields the approximation:

$$\Lambda_n(\theta_n,\theta)=-\sum_{t=1}^{n}h_n'\,\delta_n H_{t-1}(\theta)\psi(\varepsilon_t)-\frac{1}{2}\sum_{t=1}^{n}E\left[\left|h_n'\,\delta_n H_{t-1}(\theta)\psi(\varepsilon_t)\right|^2\left|F_{t-1}\right.\right]+o_p(1),$$

from which it follows that

$$\Lambda_n(\theta_n,\theta)=-\sum_{t=1}^{n}h_n'\,\delta_n H_{t-1}(\theta)\psi(\varepsilon_t)-\frac{1}{2}\sum_{t=1}^{n}h_n'\,\delta_n H_{t-1}(\theta)\Omega H_{t-1}(\theta)'\,\delta_n h_n+o_p(1),$$

establishing Lemma A.7. •

The following two lemmas are used in the proof of Lemma A.10.

Lemma A.8 *(Lemma 19 in Jeganathan (1994))*:

Let W(y) be Lebesgue measurable such that $\int\left|W(y)\right|^2 dy<\infty$. *Then*

$$\int\left|W\left(\frac{y+\omega}{\delta}\right)-W(y)\right|^2 dy\rightarrow 0$$

as $\omega\rightarrow 0$ and $\delta\rightarrow 1$.

Lemma A.9 (equation (2.32) of Hall and Heyde (1980, p.46)):

For each $n \geq 1$, let $(\xi_{n1}, \ldots, \xi_{nn})$ be an array of random variables and let $(G_{n1}, \ldots, G_{nn})$ be an array of σ-fields such that $G_{n1} \subseteq \ldots \subseteq G_{nn}$ and ξ_{nt} is G_{nt}-measurable. Furthermore, let G_{n0} be the trivial σ-field. Then, for any constants $\tau > 0$ and $\omega > 0$,

$$P\left(\max_{t \in \{1,\ldots,n\}} |\xi_{nt}| > \tau\right) \leq \omega + P\left(\sum_{t=1}^{n} E\left[\xi_{nt}^2 \cdot I\left(|\xi_{nt}| > \tau\right) \middle| G_{n,t-1}\right] > \tau^2 \omega\right).$$

To complete the proof of Theorem 1, we must establish Lemma A.10.

Lemma A.10: Conditions A.1-A.5 are satisfied for our model.

Proof: Conditions A.1 and A.2 are primitive conditions on $p(\varepsilon)$ that are assumed to hold. Condition A.4 is satisfied by (108) of Jeganathan (1994), and Condition A.5 holds by an earlier assumption. Consequently, the remainder of the proof deals with the verification of Condition A.3. To this end, we must show that equations (45), (46), and (47) are satisfied for our model. Upon the substitution of (5), this entails checking the following three conditions, respectively:

$$(56) \quad \sum_{t=1}^{n} E\left[\left|h_n' \delta_n H_{t-1}(\theta)\psi(\varepsilon_t) - \frac{1}{n^{3/2}} X_{2,t-1}' b_n' a_n' \psi(\varepsilon_t)\right|^2 \middle| F_{t-1}\right] = O_p(1) \ in \ P_{\theta,n},$$

(57)
$$\sum_{t=1}^{n}\int\int_{0}^{1}\left|\left(h_n'\delta_n H_{t-1}(\theta)-\frac{1}{n^{3/2}}X_{2,t-1}'b_n'a_n'\right)\right.$$
$$\left.\cdot\left[\psi^*\left(\varepsilon-\kappa\left(h_n'\delta_n H_{t-1}(\theta)-\frac{1}{n^{3/2}}X_{2,t-1}'b_n'a_n'\right)\right)-\psi^*(\varepsilon)\right]\right|^2 d\varepsilon d\kappa=o_p(1)\ \ in\ P_{\theta,n},$$

and, for every $\omega>0$,

(58)
$$\sum_{t=1}^{n}E\left[\left|h_n'\delta_n H_{t-1}(\theta)\psi(\varepsilon_t)-\frac{1}{n^{3/2}}X_{2,t-1}'b_n'a_n'\psi(\varepsilon_t)\right|^2\right.$$
$$\left.\cdot I\left(\left|h_n'\delta_n H_{t-1}(\theta)\psi(\varepsilon_t)-\frac{1}{n^{3/2}}X_{2,t-1}'b_n'a_n'\psi(\varepsilon_t)\right|>\omega\right)\middle|F_{t-1}\right]=o_p(1)\ \ in\ P_{\theta,n}.$$

These equations are respectively verified in steps (i)-(iii) below.

Step (i): To prove (56), it is sufficient to verify equations (59) and (60) below:

(59) $$\sum_{t=1}^{n}E\left[\left|h_n'\delta_n H_{t-1}(\theta)\psi(\varepsilon_t)\right|^2\middle|F_{t-1}\right]=O_p(1)\ \ in\ P_{\theta,n},$$

(60) $$\sum_{t=1}^{n}E\left[\left|\frac{1}{n^{3/2}}X_{2,t-1}'b_n'a_n'\psi(\varepsilon_t)\right|^2\middle|F_{t-1}\right]=o_p(1)\ \ in\ P_{\theta,n}.$$

First note that (60) is the same as (54) above, and so is already established. We therefore proceed with (59), whose left-hand side equals

$$h_n'\left[\sum_{t=1}^{n}\delta_n H_{t-1}(\theta)\Omega H_{t-1}(\theta)'\delta_n\right]h_n.$$

For convenience, define $M_{t-1} = [W_{t-1}', Z_{t-1}']'$. Note that M_{t-1} is the $r + q(k-1)$-dimensional vector of stationary regressors in the error correction model. With this notation, we can carry out the following calculations:

$$
\begin{aligned}
H_{t-1}(\theta)\Omega H_{t-1}(\theta)' &= \begin{bmatrix} I_q \otimes M_{t-1} \\ -A' \otimes X_{2,t-1} \end{bmatrix} \Omega \begin{bmatrix} I_q \otimes M_{t-1}' & -A \otimes X_{2,t-1}' \end{bmatrix} \\
&= \begin{bmatrix} I_q \otimes M_{t-1} \\ -A' \otimes X_{2,t-1} \end{bmatrix} (\Omega \otimes 1) \begin{bmatrix} I_q \otimes M_{t-1}' & -A \otimes X_{2,t-1}' \end{bmatrix} \\
&= \begin{bmatrix} \Omega \otimes M_{t-1} \\ -A'\Omega \otimes X_{2,t-1} \end{bmatrix} \begin{bmatrix} I_q \otimes M_{t-1}' & -A \otimes X_{2,t-1}' \end{bmatrix} \\
&= \begin{bmatrix} \Omega \otimes M_{t-1}M_{t-1}' & -\Omega A \otimes M_{t-1}X_{2,t-1}' \\ -A'\Omega \otimes X_{2,t-1}M_{t-1}' & A'\Omega A \otimes X_{2,t-1}X_{2,t-1}' \end{bmatrix},
\end{aligned}
$$

from which it follows that

(61) $$\sum_{t=1}^{n} \delta_n H_{t-1}(\theta)\Omega H_{t-1}(\theta)' \delta_n = \sum_{t=1}^{n} \begin{bmatrix} \dfrac{\Omega \otimes M_{t-1}M_{t-1}'}{n} & \dfrac{-\Omega A \otimes M_{t-1}X_{2,t-1}'}{n^{3/2}} \\ \dfrac{-A'\Omega \otimes X_{2,t-1}M_{t-1}'}{n^{3/2}} & \dfrac{A'\Omega A \otimes X_{2,t-1}X_{2,t-1}'}{n^2} \end{bmatrix}.$$

To verify (59), we must show that the matrix on the right-hand side of (61) is $O_p(1)$ in $P_{\theta,n}$. To achieve this, we show that the following limit theory holds:

(62) $$\frac{1}{n}\sum_{t=1}^{n} \Omega \otimes M_{t-1}M_{t-1}' = \Omega \otimes E[M_t M_t'] + o_p(1) \quad in\ P_{\theta,n},$$

(63) $$\frac{1}{n^{3/2}}\sum_{t=1}^{n} \Omega A \otimes M_{t-1}X_{2,t-1}' = o_p(1) \quad in\ P_{\theta,n}, \text{ and}$$

(64) $$\frac{1}{n^2}\sum_{t=1}^{n} A'\,\Omega A \otimes X_{2,t-1}X_{2,t-1}' \Rightarrow A'\,\Omega A \otimes \int_0^1 B_2 B_2'.$$

To complete the proof of step (i), we prove these equations in turn. We obtain (62) as follows:

$$\begin{aligned}
&\frac{1}{n}\sum_{t=1}^{n} \Omega \otimes M_{t-1}M_{t-1}' \\
&= \Omega \otimes \frac{1}{n}\sum_{t=1}^{n} M_{t-1}M_{t-1}' \\
&= \Omega \otimes E[M_t M_t'] + o_p(1) \;\; in\; P_{\theta,n}.
\end{aligned}$$

The second equality in this expression holds since, by Lemma 1(iv) of Ahn and Reinsel (1990, p. 815),

$$\frac{1}{n}\sum_{t=1}^{n} M_{t-1}M_{t-1}' = E[M_t M_t'] + o_p(1) \;\; in\; P_{\theta,n}.$$

To show (63), we rewrite its left-hand side as follows:

$$\begin{aligned}
&\Omega A \otimes \frac{1}{n^{3/2}}\sum_{t=1}^{n} M_{t-1}X_{2,t-1}' \\
&= \Omega A \otimes o_p(1) \;\; in\; P_{\theta,n} \\
&= o_p(1) \;\; in\; P_{\theta,n}.
\end{aligned}$$

The $o_p(1)$ result in the first equality holds because

$$\frac{1}{n}\sum_{t=1}^{n} M_{t-1}X_{2,t-1}' \Rightarrow \int_0^1 dB_M B_2' + const,$$

where B_M is a Brownian motion with covariance matrix $E[M_tM_t']$ and the constant depends on the correlation between M_{t-1} and $X_{2,t-1}$.

We finish step (i) by proving (64), whose left-hand side we rewrite as

$$A'\Omega A \otimes \frac{1}{n^2}\sum_{t=1}^{n} X_{2,t-1}X_{2,t-1}'.$$

We obtain (64) because

$$\frac{1}{n^2}\sum_{t=1}^{n} X_{2,t-1}X_{2,t-1}' \Rightarrow \int_0^1 B_2B_2'.$$

In consequence of (62), (63), and (64), we have

$$(65) \qquad \sum_{t=1}^{n} \delta_n H_{t-1}(\theta)\Omega H_{t-1}(\theta)'\,\delta_n \Rightarrow \begin{bmatrix} \Omega\otimes E[M_tM_t'] & 0 \\ 0 & A'\Omega A\otimes \int_0^1 B_2B_2' \end{bmatrix}.$$

This completes step (i) by verifying (59) and therefore (56). As remarked in the text, (65) is important in its own right because it gives us an expression for the asymptotic information matrix of the model.

<u>Step (ii)</u>: To verify (57), it is sufficient to check the following two conditions:

$$(66) \qquad \sum_{t=1}^{n}\int_0^1\int \Big| h_n'\,\delta_n H_{t-1}(\theta)\Big[\psi^*\Big(\varepsilon-\kappa\big(h_n'\,\delta_n H_{t-1}(\theta)-n^{-3/2}X_{2,t-1}'\,b_n'\,a_n'\big)\Big) - \psi^*(\varepsilon)\Big]\Big|^2 d\varepsilon\, d\kappa = o_p(1) \ in\ P_{\theta,n}.$$

$$(67)\quad \sum_{t=1}^{n}\int_0^1\int \left| n^{-3/2} X_{2,t-1}' b_n' a_n' \left[\psi^*\left(\varepsilon - \kappa\left(h_n' \delta_n H_{t-1}(\theta) - n^{-3/2} X_{2,t-1}' b_n' a_n'\right)\right) - \psi^*(\varepsilon)\right]\right|^2 d\varepsilon d\kappa = o_p(1) \ in\ P_{\theta,n}.$$

Most of the remainder of this step is devoted to proving (66), since (67) will then follow directly. We begin our proof of (66) by noting that its left-hand side is less than or equal to

$$\left(\sum_{t=1}^{n} \left| h_n' \delta_n H_{t-1}(\theta)\right|^2\right) \int_0^1 \left(\max_{t\in\{1,\ldots,n\}} \int \left| \psi^*\left(\varepsilon - \kappa\left(h_n' \delta_n H_{t-1}(\theta) - n^{-3/2} X_{2,t-1}' b_n' a_n'\right)\right) - \psi^*(\varepsilon)\right|^2 d\varepsilon \right) d\kappa$$

$$= O_p(1) \cdot \int_0^1 o_p(1) d\kappa = o_p(1) \ in\ P_{\theta,n}.$$

The $O_p(1)$ result follows from (65). The $o_p(1)$ result will follow from Lemma A.8 and the fact that $\int \left|\psi^*(\varepsilon)\right|^2 d\varepsilon = \lambda^2 < \infty$, provided we can prove that

$$(68)\quad \max_{t\in\{1,\ldots,n\}} \left| h_n' \delta_n H_{t-1}(\theta) - n^{-3/2} X_{2,t-1}' b_n' a_n' \right| = o_p(1) \ in\ P_{\theta,n}.$$

To prove (68), note that its left-hand side is less than or equal to

$$\max_{t\in\{1,\ldots,n\}} \left| h_n' \delta_n H_{t-1}(\theta)\right| + \max_{t\in\{1,\ldots,n\}} \left| n^{-3/2} X_{2,t-1}' b_n' a_n' \right|.$$

To show that this sum is $o_p(1)$, it suffices to show

$$(69)\quad \max_{t\in\{1,\ldots,n\}} \left| h_n' \delta_n H_{t-1}(\theta)\right| = o_p(1) \ in\ P_{\theta,n}$$

and

(70) $$\max_{t\in\{1,\ldots,n\}} \left| n^{-3/2} X_{2,t-1}' b_n' a_n' \right| = o_p(1) \;\; in \; P_{\theta,n}.$$

We only prove (69), since (70) then follows easily. We begin by rewriting (69) as

(71) $$\max_{t\in\{1,\ldots,n\}} h_n' \begin{bmatrix} \dfrac{I \otimes M_{t-1} M_{t-1}'}{n} & \dfrac{-A \otimes M_{t-1} X_{2,t-1}'}{n^{3/2}} \\ \dfrac{-A' \otimes X_{2,t-1} M_{t-1}'}{n^{3/2}} & \dfrac{A' A \otimes X_{2,t-1} X_{2,t-1}'}{n^2} \end{bmatrix} h_n = o_p(1) \;\; in \; P_{\theta,n}.$$

We verify (71) by checking the following three equations:

$$\max_{t\in\{1,\ldots,n\}} \left| \frac{M_{t-1}^j M_{t-1}^\ell}{n} \right| = o_p(1) in P_{\theta,n} \; \forall j, \ell,$$

$$\max_{t\in\{1,\ldots,n\}} \left| \frac{M_{t-1}^j X_{2,t-1}^\ell}{n^{3/2}} \right| = o_p(1) in P_{\theta,n} \; \forall j, \ell,$$

$$\max_{t\in\{1,\ldots,n\}} \left| \frac{X_{2,t-1}^j X_{2,t-1}^\ell}{n^2} \right| = o_p(1) in P_{\theta,n} \; \forall j, \ell,$$

where the superscripts j and ℓ represent the j^{th} and ℓ^{th} elements of the respective vectors.

Using the inequalities

$$\max_{t\in\{1,\dots,n\}}\left|\frac{M_{t-1}^{j}M_{t-1}^{\ell}}{n}\right| \le \max_{t\in\{1,\dots,n\}}\left|\frac{M_{t-1}^{j}}{\sqrt{n}}\right| \max_{t\in\{1,\dots,n\}}\left|\frac{M_{t-1}^{\ell}}{\sqrt{n}}\right|,$$

$$\max_{t\in\{1,\dots,n\}}\left|\frac{M_{t-1}^{j}X_{2,t-1}^{\ell}}{n^{3/2}}\right| \le \max_{t\in\{1,\dots,n\}}\left|\frac{M_{t-1}^{j}}{\sqrt{n}}\right| \max_{t\in\{1,\dots,n\}}\left|\frac{X_{2,t-1}^{\ell}}{n}\right|,$$

$$\max_{t\in\{1,\dots,n\}}\left|\frac{X_{2,t-1}^{j}X_{2,t-1}^{\ell}}{n^{2}}\right| \le \max_{t\in\{1,\dots,n\}}\left|\frac{X_{2,t-1}^{j}}{n}\right| \max_{t\in\{1,\dots,n\}}\left|\frac{X_{2,t-1}^{\ell}}{n}\right|,$$

proving (69) reduces to proving

(72) $$\max_{t\in\{1,\dots,n\}}\left|\frac{M_{t-1}^{j}}{\sqrt{n}}\right| = o_p(1) \;\; in\ P_{\theta,n}, \;\; \forall j,$$

and

(73) $$\max_{t\in\{1,\dots,n\}}\left|\frac{X_{2,t-1}^{j}}{n}\right| = o_p(1) \;\; in\ P_{\theta,n}, \;\; \forall j.$$

To prove (72), we use the following representation:

$$M_{t-1}^{j} = \sum_{s=1}^{\infty} c_s e_{t-s}$$

where the e_t are iid innovations and $\sum_{s=1}^{\infty}|c_s| < \infty$. It follows that

(74) $$\max_{t\in\{1,\dots,n\}}\left|\frac{M_{t-1}^{j}}{\sqrt{n}}\right| \le \left(\sum_{s=1}^{\infty}|c_s|\right) \max_{t\in\{1,\dots,n\}}\left|\frac{e_t}{\sqrt{n}}\right| = o_p(1) \;\; in\ P_{\theta,n}.$$

To show that $\max_{t\in\{1,\ldots,n\}}\left|\frac{e_t}{\sqrt{n}}\right| = o_p(1)$ in $P_{\theta,n}$, we apply Lemma A.9, setting $\xi_{nt} = n^{-1/2}e_t$ and $G_{nt} = \sigma(\ldots,e_{t-1},e_t)$. By Lemma A.9,

$$P\left(\max_{t\in\{1,\ldots,n\}}\left|\frac{e_t}{\sqrt{n}}\right| > \tau\right) \le \omega + P\left(\sum_{t=1}^{n} E\left[\frac{e_t^2}{n}\cdot I\left(\left|\frac{e_t}{\sqrt{n}}\right| > \tau\right)\middle| G_{n,t-1}\right] > \tau^2\omega\right).$$

The desired result follows since

$$\frac{1}{n}\sum_{t=1}^{n} E\left[e_t^2 \cdot I\left(|e_t| > \tau\sqrt{n}\right)\right] = o(1).$$

To show (73), we use the Beveridge-Nelson (1981) decomposition:

$$X_{2,t-1}^{j} = C(1)\sum_{s=0}^{t-1} e_s + \tilde{e}_0 - \tilde{e}_{t-1}$$

where $C(1) < \infty$, $\{e_s\}$ are iid innovations, and $\{\tilde{e}_s\}$ is a stationary process defined in Phillips and Solo (1992). To prove (73), note that:

$$\max_{t\in\{1,\ldots,n\}}\left|\frac{X_{2,t-1}^{\ell}}{n}\right| \le C(1)\max_{t\in\{1,\ldots,n\}}\left|\frac{\sum_{s=0}^{t-1} e_s}{n}\right| + \max_{t\in\{1,\ldots,n\}}\left|\frac{\tilde{e}_0 - \tilde{e}_{t-1}}{n}\right|$$
$$= o_p(1) + o_p(1) = o_p(1) \quad in\ P_{\theta,n},$$

the second term being $o_p(1)$ by an argument analogous to that used to show (72) and the first term being $o_p(1)$ by result (10.10) on p.70 of Billingsley (1968). From verification of (72) and (73) we obtain (69) and therefore (68), completing our proof of (66). As noted, (66) implies (67), since the left-hand side of the latter is less than or equal to

$$\left(n^{-3}\sum_{t=1}^{n}\left|X_{2,t-1}'b_n'a_n'\right|^2\right)\int_0^1\left(\max_{t\in\{1,\ldots,n\}}\int\left|\psi^*\left(\varepsilon-\kappa\left(h_n'\delta_nH_{t-1}(\theta)-n^{-3/2}X_{2,t-1}'b_n'a_n'\right)\right)\right.\right.$$
$$\left.\left.-\psi^*(\varepsilon)\right|^2d\varepsilon\right)d\kappa$$
$$=o_p(1)\cdot\int_0^1 o_p(1)d\kappa=o_p(1)\ \ in\ P_{\theta,n}.$$

We have finished step (ii), so the Lemma will follow immediately from step (iii).

Step (iii): It remains to verify (58), whose left-hand side can be rewritten as:

$$\sum_{t=1}^{n}\left|h_n'\delta_nH_{t-1}(\theta)-n^{-3/2}X_{2,t-1}'b_n'a_n'\right|^2$$
$$\cdot E\left[\left|\psi(\varepsilon_t)\right|^2\cdot I\left(\left|\left(h_n'\delta_nH_{t-1}(\theta)-n^{-3/2}X_{2,t-1}'b_n'a_n'\right)\psi(\varepsilon_t)\right|>\omega\right)\middle|F_{t-1}\right]$$
$$\leq\sum_{t=1}^{n}\left|h_n'\delta_nH_{t-1}(\theta)-n^{-3/2}X_{2,t-1}'b_n'a_n'\right|^2$$
$$\cdot E\left[\left|\psi(\varepsilon_t)\right|^2\cdot I\left(\left|h_n'\delta_nH_{t-1}(\theta)-n^{-3/2}X_{2,t-1}'b_n'a_n'\right|\left|\psi(\varepsilon_t)\right|>\omega\right)\middle|F_{t-1}\right]$$
$$\leq\left(\sum_{t=1}^{n}\left|h_n'\delta_nH_{t-1}(\theta)-n^{-3/2}X_{2,t-1}'b_n'a_n'\right|^2\right)$$
$$\cdot E\left[\left|\psi(\varepsilon_t)\right|^2\cdot I\left(\left|\psi(\varepsilon_t)\right|>\frac{\omega}{\max_{t\in\{1,\ldots,n\}}\left|h_n'\delta_nH_{t-1}(\theta)-n^{-3/2}X_{2,t-1}'b_n'a_n'\right|}\right)\middle|F_{t-1}\right]$$
$$=O_p(1)\cdot o_p(1)\ \ in\ P_{\theta,n}\ \ \forall\omega>0.$$

The $o_p(1)$ result follows from the facts that

$$\max_{t\in\{1,\dots,n\}} \left| h_n{}' \delta_n H_{t-1}(\theta) - n^{-3/2} X_{2,t-1}{}' b_n{}' a_n{}' \right| = o_p(1) \;\; in\; P_{\theta,n}$$

and

$$E\left[|\psi(\varepsilon_t)|^2 | F_{t-1} \right] = \lambda^2 < \infty.$$

This completes our proof of step (iii) and therefore of Lemma A.10. •

Theorem 3.2 now follows immediately. •

Proof of Lemma 5.5: Condition 5.1 is satisfied by assumption, so we proceed in two steps, checking Conditions 5.2 and 5.3, respectively.

Step (i): To show that Condition 5.2 holds, it is sufficient to verify the following three equations (corresponding, respectively, to (21), (22), and (23)):

(75)
$$\sum_{t=1}^{n} \left[\left(h_n{}' \delta_n H_{t-1}^j(\theta) - n^{-3/2} X_{2,t-1}{}' (b_n{}' a_n{}')^j \right) - h_n{}' \delta_n H_{t-1}^j(\theta) \right]^2 = \sum_{t=1}^{n} \left[n^{-3/2} X_{2,t-1}{}' (b_n{}' a_n{}')^j \right]^2 = o_p(1) \;\; in\; P_{\theta,n},$$

(76)
$$\sum_{t=1}^{n} \left| h_n{}' \delta_n H_{t-1}^j(\theta) \right|^2 = O_p(1) \;\; in\, P_{\theta,n}, \text{ and,}$$

(77)
$$\max_{t\in\{1,\dots,n\}} \left| h_n{}' \delta_n H_{t-1}^j(\theta) \right|^2 = o_p(1) \;\; in\; P_{\theta,n}.$$

Denoting the vector $(b_n{}' a_n{}')^j$ by γ, the left-hand side of the second equality in (75) can be written as

$$n^{-3}\sum_{t=1}^{n}\gamma' X_{2,t-1}X_{2,t-1}'\gamma$$
$$= n^{-1}\gamma'\left(n^{-2}\sum_{t=1}^{n}X_{2,t-1}X_{2,t-1}'\right)\gamma = o_p(1) \;\; in\; P_{\theta,n}$$

because

$$\gamma'\left(n^{-2}\sum_{t=1}^{n}X_{2,t-1}X_{2,t-1}'\right)\gamma \Rightarrow \gamma'\left(\int_0^1 B_2B_2'\right)\gamma.$$

We now turn to (76), which can be rewritten as

$$(78) \qquad \sum_{t=1}^{n}\left|h_n'\begin{bmatrix}\dfrac{\iota^j\otimes M_{t-1}}{\sqrt{n}}\\ \dfrac{-A(j)'\otimes X_{2,t-1}}{n}\end{bmatrix}\right|^2 = O_p(1) \;\; in\; P_{\theta,n},$$

where ι^j denotes the j^{th} column of the identity matrix and $A(j)$ denotes the j^{th} row of the matrix A. Verification of (78) follows in the same manner as verification of (65).

To verify (77), we must show that

$$\max_{t\in\{1,\ldots,n\}} h_n'\begin{bmatrix}\dfrac{\iota^j\iota^{j\prime}\otimes M_{t-1}M_{t-1}'}{n} & \dfrac{-\iota^jA(j)\otimes M_{t-1}X_{2,t-1}'}{n^{3/2}}\\ \dfrac{-A(j)'\,\iota^{j\prime}\otimes X_{2,t-1}M_{t-1}'}{n^{3/2}} & \dfrac{A(j)'A(j)\otimes X_{2,t-1}X_{2,t-1}'}{n^2}\end{bmatrix}h_n = o_p(1) \;\; in\; P_{\theta,n},$$

which is easily shown using the facts that

$$\max_{t\in\{1,\ldots,n\}}\left|\frac{M_{t-1}^i}{\sqrt{n}}\right| = o_p(1) \;\; in\; P_{\theta,n} \quad \forall i = 1,\ldots,r+q(k-1), \text{ and}$$

$$\max_{t\in\{1,\dots,n\}}\left|\frac{X^i_{2,t-1}}{n}\right| = o_p(1) \ \ in\ P_{\theta,n} \quad \forall i=1,\dots,q-r.$$

Step (ii): It remains to verify Condition 5.3. To do so, we set w=m and $V^j_{nt}(\theta) = U^j_{nt}(\theta)$, so that (24) becomes

$$(79) \qquad \sum_{t=1}^{n}\left|u'\,\delta_n H^j_{t-1}(\theta_n) - u'\,\delta_n H^j_{t-1}(\theta)\right|^2 = o_p(1) \ \ in\ P_{\theta,n},$$

while, by setting $\delta = 0$, (25) becomes

$$(80) \qquad \max_{t\in\{1,\dots,n\}} n\left|\delta_n H^j_{t-1}(\theta)\right|^2 = O_p(1) \ \ in\ P_{\theta,n}.$$

We begin by checking (79), which can be rewritten as

$$\begin{aligned}
&u'\,\delta_n\left(\sum_{t=1}^{n} H^j_{t-1}(\theta_n)H^j_{t-1}(\theta_n)'\right)\delta_n u - u'\,\delta_n\left(\sum_{t=1}^{n} H^j_{t-1}(\theta)H^j_{t-1}(\theta_n)'\right)\delta_n u \\
&-u'\,\delta_n\left(\sum_{t=1}^{n} H^j_{t-1}(\theta_n)H^j_{t-1}(\theta)'\right)\delta_n u + u'\,\delta_n\left(\sum_{t=1}^{n} H^j_{t-1}(\theta)H^j_{t-1}(\theta)'\right)\delta_n u \\
&= o_p(1) \ \ in\ P_{\theta,n},
\end{aligned}$$

which holds because

$$\delta_n\left(\sum_{t=1}^{n} H_{t-1}^j(\theta_n)H_{t-1}^j(\theta_n)'\right)\delta_n \Rightarrow S^{jj}(\theta),$$

$$\delta_n\left(\sum_{t=1}^{n} H_{t-1}^j(\theta_n)H_{t-1}^j(\theta)'\right)\delta_n \Rightarrow S^{jj}(\theta),$$

$$\delta_n\left(\sum_{t=1}^{n} H_{t-1}^j(\theta)H_{t-1}^j(\theta_n)'\right)\delta_n \Rightarrow S^{jj}(\theta),$$

$$\delta_n\left(\sum_{t=1}^{n} H_{t-1}^j(\theta)H_{t-1}^j(\theta)'\right)\delta_n \Rightarrow S^{jj}(\theta),$$

where $S^{ij}(\theta)$ is defined by writing

$$S(\theta) = \sum_{i=1}^{q}\sum_{\ell=1}^{q}\omega_{i\ell}S^{i\ell}(\theta).$$

This proves (79). We verify (80) by writing its left-hand side as

$$\max_{t\in\{1,\dots,n\}} n\left|\begin{matrix}\dfrac{\iota^j\otimes M_{t-1}}{\sqrt{n}}\\ \dfrac{-A(j)'\otimes X_{2,t-1}}{n}\end{matrix}\right|^2$$

$$= \max_{t\in\{1,\dots,n\}} n\left[\frac{\iota^{j\prime}\iota^j\otimes M_{t-1}'M_{t-1}}{n} + \frac{A(j)A(j)'\otimes X_{2,t-1}'X_{2,t-1}}{n^2}\right]$$

$$= \max_{t\in\{1,\dots,n\}}\left[M_{t-1}'M_{t-1} + n^{-1}(A(j)A(j)')X_{2,t-1}'X_{2,t-1}\right]$$

$$\leq \max_{t\in\{1,\dots,n\}} M_{t-1}'M_{t-1} + \max_{t\in\{1,\dots,n\}} n^{-1}(A(j)A(j)')X_{2,t-1}'X_{2,t-1}$$

$$= O_p(1) + O_p(1) = O_p(1) \ \ in\ P_{\theta,n}.$$

This completes step (ii) and proves the lemma. •

Proof of Theorem 5.6: From (15) and (30), it follows that

(81) $$\delta_n^{-1}(\hat{\theta}_n - \tilde{\theta}_n) = \hat{S}_n^{-1}W_n^*(\theta_n^{**}) - \hat{S}_n^{-1}(\theta_n^{**})\hat{\Delta}_n(\theta_n^{**}).$$

Using (29), we have

(82) $$\hat{\Delta}_n(\theta_n^{**}) = \hat{\Delta}_n(\theta) - S_n(\theta)h_n + o_p(1) \;\; in \; P_{\theta,n}.$$

Combining (82) and (27) gives us

(82) $$\hat{\Delta}_n(\theta_n^{**}) = \Delta_n(\theta) - S_n(\theta)h_n + o_p(1) \;\; in \; P_{\theta,n},$$

so that the second term on the right-hand side of (81) becomes, using (28), (83), and the fact that $\Delta_n(\theta) = W_n(\theta)$,

(84) $$-S_n^{-1}(\theta)\left[W_n(\theta) - S_n(\theta)h_n\right] + o_p(1) \;\; in \; P_{\theta,n}.$$

By definition,

(85) $$\hat{S}_n = S_n(\theta) + o_p(1) \;\; in \; P_{\theta,n},$$

while (16) gives

(86) $$\begin{aligned} W_n^*(\theta_n^{**}) &= W_n(\theta) - \hat{S}_n h_n + o_p(1) \;\; in \; P_{\theta,n} \\ &= W_n(\theta) - S_n(\theta)h_n + o_p(1) \;\; in \; P_{\theta,n}, \end{aligned}$$

the second equality holding due to (85). Combining (85) and (86) gives

(87) $$\hat{S}_n^{-1} W_n^*(\theta_n^{**}) = S_n^{-1}(\theta)\left[W_n(\theta) - S_n(\theta)h_n\right] + o_p(1) \;\; in \; P_{\theta,n}.$$

Using (81), (84), and (87), the desired result (31) follows. •

REFERENCES

Ahn, S.K. and Reinsel, G.C. 1988. Nested reduced-rank autoregressive models for multiple time series. *Journal of the American Statistical Association* 83:849-856.

Ahn, S.K. and Reinsel, G.C. 1990. Estimation for partially nonstationary multivariate autoregressive models. *Journal of the American Statistical Association* 85:813-823.

Beran, R. 1974. Asymptotically efficient rank estimates in location models. *Annals of Statistics* 2:63-74.

Beveridge, S. and Nelson, C. R. 1981. A new approach to decomposition of economic time series into permanent and transitory components with particular attention to measurement of the "business cycle". *Journal of Monetary Economics* 7:151-174.

Bickel, P.J. 1982. On adaptive estimation. *Annals of Statistics* 10:647-671.

Billingsley, P. 1968. *Convergence of Probability Measures*. New York: John Wiley & Sons.

Engle, R.F. and Granger, C.W.J. 1987. Co-integration and error correction: representation, estimation and testing. *Econometrica* 55:251-276.

Fama, E.F. 1963. Mandelbrot and the stable Paretian hypothesis. *Journal of Business* 36:420-429.

Fama, E.F. 1965. The behaviour of stock market prices. *Journal of Business* 38:34-105.

Fang, K.T., Kotz, S., and Ng, K.W. 1990. *Symmetric Multivariate and Related Distributions*. London; Chapman and Hall.

Friedman, B.M. and Kuttner, K.N. 1992. Money, income, prices, and interest rates. *American Economic Review* 82:472-492.

Ghosh, J.K. 1985. Efficiency of estimates-part I. *Sankhya A* 47:310-325.

Hall, P. and Heyde, C.C. 1980. *Martingale Limit Theory and its Applications*. New York; Academic Press.

Hardle, W., Hart, J., Marron, J.S., and Tsybakov, A.B. 1992. Bandwidth choice for average derivative estimation. *Journal of the American Statistical Association* 87:218-226.

Hardle, W., Marron, J.S., and Wand, M.P. 1990. Bandwidth choice for density derivatives. *Journal of the Royal Statistical Society, Series B* 52:223-232.

Hsieh, D.A. and Manski, C.F. 1987. Monte Carlo evidence on adaptive maximum likelihood estimation of a regression. *Annals of Statistics* 15:541-551.

Jeganathan, P. 1988. Some aspects of asymptotic theory with applications to time series models. University of Michigan Department of Statistics, Technical Report #166, September.

Jeganathan, P. 1994. Some aspects of asymptotic theory with applications to time series models. Forthcoming in *Econometric Theory*.

Johansen, S. 1988. Statistical analysis of cointegration vectors. *Journal of Economic Dynamics and Control* 12:231-254.

Johansen, S. 1991. Estimation and hypothesis testing of cointegration vectors in Gaussian vector autoregressive models. *Econometrica* 59:1551-1580.

Johansen, S. 1992. Testing weak exogeneity and order of cointegration in UK money demand data. *Journal of Policy Modeling* 14:313-334.

Johansen, S. 1994. The role of the constant and linear terms in cointegration analysis of nonstationary variables. *Econometric Reviews* 13:205-229.

Johansen, S. and Juselius, K. 1990. Maximum likelihood estimation and inference on cointegration - with applications to the demand for money. *Oxford Bulletin of Economics and Statistics* 52:169-210.

Johansen, S. and Juselius, K. 1992. Testing structural hypotheses in a multivariate cointegration analysis of the PPP and the UIP for UK. *Journal of Econometrics* 53:211-244.

Kasa, K. 1992. Common stochastic trends in international stock markets. *Journal of Monetary Economics* 29:95-124.

Kreiss, J.-P. 1987a. On adaptive estimation in autoregressive models when there are nuisance functions. *Statistics & Decisions* 5:59-76.

Kreiss, J.-P. 1987b. On adaptive estimation in stationary ARMA processes. *Annals of Statistics* 15:112-133.

Kunst, R. and Neusser, K. 1990. Cointegration in a macroeconomic system. *Journal of Applied Econometrics* 5:351-365.

LeCam, L. and Yang, G.L. *Asymptotics in Statistics - Some Basic Concepts*. New York; Springer-Verlag.

Linton, O. 1993. Adaptive estimation in ARCH models. *Econometric Theory* 9:539-569.

Mandelbrot, B. 1963. The variation of certain speculative prices. *Journal of Business* 36:394-419.

Manski, C.F. 1984. Adaptive estimation of non-linear regression models. *Econometric Reviews* 3:145-210 (including commentary).

McDonald, J.B. and White, S.B. 1993. A comparison of some robust, adaptive, and partially adaptive estimators of regression models. *Econometric Reviews* 12:103-124.

McGuirk, A.M., Robertson, J., and Spanos, A. 1993. Modeling exchange rate dynamics: non-linear dependence and thick tails. *Econometric Reviews* 12:33-63.

Mitchell, A.F.S. 1989. The information matrix, skewness tensor and $\alpha -$connections for the general multivariate elliptic distribution. *Annals of the Institute of Mathematical Statistics* (Tokyo) 41:289-304.

Mittnik, S. and Rachev, S.T. 1993. Modeling asset returns with alternative stable distributions. *Econometric Reviews* 12:261-330.

Phillips, P.C.B. 1991. Optimal inference in cointegrated systems. *Econometrica* 59:283-306.

Phillips, P.C.B. 1993. Forward market unbiasedness: The case of the Australian dollar since 1984. Cowles Foundation Discussion Paper #1055, August.

Phillips, P.C.B. and Park, J. 1988. Asymptotic equivalence of ordinary least squares and generalized least squares in regressions with integrated regressors. *Journal of the American Statistical Association* 83:111-115.

Phillips, P.C.B. and Solo, V. 1992. Asymptotics for linear processes. *Annals of Statistics* 20:971-1001.

Silverman, B. 1986. *Density Estimation for Statistics and Data Analysis.* London; Chapman and Hall.

Stein, C. 1956. Efficient nonparametric estimation and testing. *Proc. Third Berkeley Symp. Math. Statist Prob.* 1:187-196.

Stone, C. 1975. Adaptive maximum likelihood estimation of a location parameter. *Annals of Statistics* 3:267-284.

Velu, R.P., Reinsel, G.C., and Wichern, D.W. 1986. Reduced rank models for multiple series. *Biometrika* 73:105-118.

White, H. 1984. *Asymptotic Theory for Econometricians*. Orlando; Academic Press.