Econometric Theory, **14**, 1998, 44–69. Printed in the United States of America.

ADAPTIVE ESTIMATION OF ERROR CORRECTION MODELS

DOUGLAS J. HODGSON
University of Rochester

This paper considers 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–maximum likelihood estimator are evaluated for elliptically symmetric innovations.

1. 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 because the reduced rank matrix can be decomposed into a matrix of cointegrating vectors and a matrix of error correction coefficients, or factor load-

This paper is based on Chapter 2 of my Ph.D. dissertation at Yale University. I am grateful to Peter Phillips and Oliver Linton for their advice and assistance and to the anonymous referees, Stephane Gregoire, and seminar participants at Yale, Rochester, Ohio State, Columbia, York, CREST at INSEE, the Department of Finance Canada, Chicago GSB, Iowa State, Northwestern, NYU, Michigan, Boston College, and the 1995 Canadian Econometrics Study Group for their comments. I also acknowledge financial assistance from the Alfred P. Sloan Foundation and the Social Sciences and Humanities Research Council of Canada. A GAUSS program implementing the adaptive estimator developed in this paper is available from the author upon request. Address correspondence to: Douglas J. Hodgson, Department of Economics, University of Rochester, Rochester, NY 14627-0156, USA; e-mail: dshn@troi.cc.rochester.edu.

  0266-4666/98 $9.50

ings, 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 derived 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 empirical cointegration analysis. For example, Johansen (1992), Johansen and Juselius (1990), and Friedman and Kuttner (1992) estimated monetary models, Johansen and Juselius (1992) estimated exchange rate models, Kunst and Neusser (1990) estimated real business cycle models, and Kasa (1992) estimated 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) reported a simulation study comparing their estimator of the cointegrating parameter in a Gaussian bivariate model with ordinary least squares (OLS). They found 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 has been provided by Phillips (1991).

Although the MLE's discussed previously 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) simulations show, it matters in the estimation of cointegrating vectors whether or not an efficient estimator is used. Some of the applied studies cited earlier (Johansen and Juselius, 1990, 1992; Kasa, 1992) test and reject the Gaussianity hypothesis for 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, e.g., Mandelbrot, 1963; Fama, 1963, 1965; Mittnik and Rachev, 1993; 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) and Phillips, McFarland, and McMahon (1996) 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 estimator 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) found that adaptive estimators compare quite favorably with OLS, least absolute deviations (LAD), generalized method of moments (GMM), and *M*-estimators in the estimation of a (noncointegrating) non-Gaussian linear regression model. Further evidence on the finite-sample behavior of adaptive estimators in stationary models was provided by Hsieh and Manski (1987) and Steigerwald (1992).

There is a growing literature addressing the problem of robust and efficient estimation in non-Gaussian cointegrated models. Phillips (1995) developed robust LAD and *M*-estimators for triangular models, and adaptive estimators in triangular models were derived by Jeganathan (1995, 1997) and Hodgson (in press). Simulation and empirical results obtained by Hodgson (1995, in press) illustrate the good finite-sample properties of the adaptive estimators and of the estimator developed in the present paper.

We analyze adaptive estimation of reduced rank regression in cointegrated error correction models. We extend the work of Jeganathan (1995) and Hodgson (in press), who analyzed the adaptive estimation of linear cointegrating regressions. In Section 2, the model and notation are introduced. In Section 3, 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 4, we define exactly what we mean by the term *efficient estimator* and describe the optimality properties of these estimators in LAN/LAMN models. We first show how to compute efficient estimators when the density function of the innovations is known, and then we show how to construct estimators that are asymptotically equivalent to efficient estimators, thus sharing their optimality properties, but that do not require knowledge of the shape of the density of the innovations. These estimators, termed *adaptive*, use nonparametric density estimators 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 5 discusses possible extensions of this research. The Appendix contains proofs of all lemmas and theorems.

The following notation is used throughout the paper. The term 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 $\mathrm{vec}(X)$ stacks the transposed rows of the matrix X, and the inequalities $X > Y$ and $X \geq Y$, when applied to matrices, signify that the difference $X - Y$ is positive definite and positive semidefinite, respectively. We simplify notation by writing $\int_0^1 Z$ in place of $\int_0^1 Z(r)dr$ when $Z(r)$ is a Brownian motion process defined on the interval [0,1]. The expression $L(X|P)$ 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 $\Rightarrow$.

2. 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 $1 \le r < q$, 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:

$$X_t = \pi_0 + \Pi_1 X_{t-1} + \cdots + \Pi_k X_{t-k} + \varepsilon_t, \tag{1}$$

where π_0 is a $q \times 1$ intercept vector. 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 the models of Johansen (1988) and Ahn and Reinsel (1990). Where we differ from these authors is in our assumptions regarding the distribution of the independent and identically distributed (i.i.d.) innovation process $\{\epsilon_t\}$. Whereas they assumed 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, $p(\varepsilon)$, which is absolutely continuous with respect to Lebesgue measure and symmetric about the origin in the sense that $p(\varepsilon) = p(-\varepsilon)$, and which satisfies the moment condition $\mathrm{cov}(\varepsilon) = \Sigma_\varepsilon < \infty$. The symmetry assumption is important for our purposes because it implies that the q-vector of partial derivatives (which we assume to exist for every ε) is antisymmetric, 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 $\psi(\varepsilon) = (\partial p(\varepsilon)/\partial\varepsilon)/p(\varepsilon)$, is also antisymmetric 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, $\Omega = \Sigma_\varepsilon^{-1}$. If Gaussianity does not hold, then $\Omega > \Sigma_\varepsilon^{-1}$.

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

$$\Delta X_t = \pi_0 + ABX_{t-1} + \sum_{j=1}^{k-1} \Phi_j \Delta X_{t-j} + \varepsilon_t, \tag{2}$$

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. We also assume that $\pi_0 = -AB_1$, where B_1 is the r-vector of intercepts in the cointegrating vectors (see (3)). Because the VAR intercept vector π_0 belongs to the subspace spanned by the columns of A, we do not allow for the presence of linear time trends in the data $\{X_t\}$.

Thus, our model belongs to the class denoted as $H_1^*(r)$ in Johansen's (1994) discussion of cointegrated models with deterministic components. This model is analyzed by Johansen (1988) and Ahn and Reinsel (1990) under the Gaussianity assumption.

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.[2] We can then rewrite (2) as

$$\Delta X_t = A[X_{1,t-1} - B_1 - B_0 X_{2,t-1}] + \Phi Y_{t-1} + \varepsilon_t, \qquad \textbf{(3)}$$

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

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, and 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, B_1, 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 subsequently that this can be done for B_1 and B_0 but not for A. In fact, we will show that B_1 and B_0 can be efficiently estimated by 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-à-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 $\alpha = \text{vec}(A)$, $\varphi = \text{vec}(\Phi)$, and $\beta = \text{vec}(B_0)$, of dimensions qr, $q^2(k-1)$, and $r(q-r)$, respectively. The respective parameter vectors are gathered into the m-dimensional full parameter vector, $\theta = [\alpha', \varphi', B_1', \beta']' = [\eta', B_1', \beta']'$, where $\eta \equiv (\alpha', \varphi')'$, $m = r + 2qr - r^2 + q^2(k-1)$, 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' is not of full column 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 $\delta_n = \text{diag}[n^{-1/2} I_s, n^{-1/2} I_r, n^{-1} I_{m-s-r}]$. 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

https://doi.org/10.1017/S0266466698141026 Published online by Cambridge University Press

so at different rates in different directions of the parameter space. In directions associated with transitory dynamics and with the cointegrating intercepts, the rate of convergence is $\sqrt{n}$, whereas in those associated with the cointegrating slope parameters, the rate is n.

In the remainder of the paper, we shall denote by $P_{\theta,n}$ the distribution of the sample of size n with parameter θ. All convergence statements, unless otherwise noted, should be read as occurring under $P_{\theta,n}$.

3. LOCAL ASYMPTOTIC (MIXED) NORMALITY

In this section, we analyze the asymptotic behavior of the log-likelihood ratio,

$$\Lambda_n(\theta_n,\theta) = \log(dP_{\theta_n,n}/dP_{\theta,n}).$$

We find that the behavior of $\Lambda_n(\theta_n,\theta)$ is such that our model falls within the LAN/LAMN family, which is important because a theory of optimal estimation applies to such models. In Section 4, optimal estimators for this family are characterized.

We establish in Theorem 3.2 that the component of the model associated with the parameters describing the long-run relationships in the model (i.e., the cointegrating coefficients B_1 and B_0) belongs to the LAMN family and that the component associated with parameters describing short-run dynamics (i.e., A and Φ) belongs to the LAN family. 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 obtain our asymptotic results regarding $\Lambda_n(\theta_n,\theta)$, we couch our model in terms of the framework of a general nonlinear model, as described by Jeganathan (1995, equation (37)). We assume that the initial observations $\underline{X}_0 = (X_{1-k},\dots,X_0)$ have a density (with respect to a σ-finite measure) denoted by $f_0(\underline{X}_0,\theta)$ that has the property that $f_0(\underline{X}_0,\theta_n) - f_0(\underline{X}_0,\theta) = o_p(1)$ when $\theta_n \to \theta$. We define $\underline{X}_t = (\underline{X}_0, X_1,\dots,X_t)$ and denote by F_t the σ-field generated by $\underline{X}_t$. We can then write our model as

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

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

$$\begin{aligned} g_t(\underline{X}_{t-1},\theta) &= X_{t-1} - AB_1 + ABX_{t-1} + \sum_{j=1}^{k-1} \Phi_j \Delta X_{t-j} \\ &= X_{t-1} - AB_1 + ABX_{t-1} + \Phi Y_{t-1}. \end{aligned}$$

The following result is derived within the preceding nonlinear framework and is useful in our subsequent development of the LAN/LAMN 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*

$$d_t(\theta_n,\theta)' = h_n'\delta_n H_{t-1}(\theta) - (n^{-3/2}X_{2,t-1}'b_n'a_n' + n^{-1}b_{1n}'a_n'), \quad \textbf{(4)}$$

https://doi.org/10.1017/S0266466698141026 Published online by Cambridge University Press

where $H_{t-1}(\theta) = [(I_q \otimes V_{t-1})', (I_q \otimes Y_{t-1})', -A, (-A' \otimes X_{2,t-1})']'$, $V_{t-1} = X_{1,t-1} - B_1 - B_0 X_{2,t-1}$, *and* $\{a_n\}$, $\{b_{1n}\}$, *and* $\{b_n\}$ *are bounded sequences of matrices of dimensions* $q \times r$, $r \times 1$, *and* $r \times (q-r)$, *respectively*.

Remark. The q-vector $d_t(\theta_n, \theta)$ plays an important role in the theory developed later. Its first component is particularly important because it enters into our expressions for the (scaled) score function and information matrix of the sample. We write the sample score function, scaled by δ_n, as

$$W_n(\theta) = -\sum_{t=1}^{n} \delta_n H_{t-1}(\theta)\psi(\varepsilon_t), \tag{5}$$

and the sample information, pre- and postmultiplied by δ_n, as

$$S_n(\theta) = \sum_{t=1}^{n} \delta_n H_{t-1}(\theta)\Omega H_{t-1}(\theta)'\delta_n. \tag{6}$$

Note that in these expressions for the sample score vector and information matrix[3] 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 4, we show how this problem can be addressed through the estimation of $\psi(\varepsilon)$ and Ω using nonparametric density estimates.

The following theorem, proved in the Appendix, is the central result of this section.

THEOREM 3.2. *The likelihood ratios* $\Lambda_n(\theta_n, \theta)$ *have the following asymptotic quadratic approximation for every bounded sequence of m-vectors* $\{h_n\}$:

$$\Lambda_n(\theta_n, \theta) = h_n' W_n(\theta) - (\tfrac{1}{2}) h_n' S_n(\theta) h_n + o_p(1). \tag{7}$$

Furthermore, *we have the following weak convergence result*:

$$L(W_n(\theta), S_n(\theta) | P_{\theta,n}) \Rightarrow L(MN(0, S(\theta)), S(\theta)), \tag{8}$$

where

$$S(\theta) = \begin{bmatrix} \Omega \otimes E[M_t M_t'] & 0 & 0 \\ 0 & A'\Omega A & A'\Omega A \otimes \bar{Z}_2' \\ 0 & A'\Omega A \otimes \bar{Z}_2 & A'\Omega A \otimes \int_0^1 Z_2 Z_2' \end{bmatrix}, \tag{9}$$

$M_{t-1} = [V_{t-1}', Y_{t-1}']'$, Z_2 *is a Brownian motion process with covariance matrix* $lr\operatorname{cov}(\Delta X_{2t}) = E[\Delta X_{2t}\Delta X_{2t}'] + \sum_{j=1}^{\infty} E[\Delta X_{2t}\Delta X_{2,t+j}'] + \sum_{j=1}^{\infty} E[\Delta X_{2t}\Delta X_{2,t+j}']'$, *and* $\bar{Z}_2 = \int_0^1 Z_2$.

Remarks. (a) Equations (7) and (8) together imply that our model belongs to the LAMN family, as defined, for example, by LeCam and Yang (1990) and

https://doi.org/10.1017/S0266466698141026 Published online by Cambridge University Press

Jeganathan (1995), the latter of whom showed, in a result similar to ours, that triangular cointegrated models belong to the LAMN family. The LAN family is a special case of the LAMN in which the matrix $S(\theta)$ is nonrandom. We can see from (9) that, for our model, $S(\theta)$ is block diagonal between components that are, respectively, nonrandom and random. It follows that we can decompose the quadratic on the right-hand side of (7) into the sum of two asymptotically independent quadratic functions, one of which has an asymptotically *nonrandom* quadratic term and represents the asymptotic approximation of the likelihood ratios of a model belonging to the LAN family, and the other of which has an asymptotically *random* quadratic term and represents the asymptotic approximation of the likelihood ratios of a model belonging to the LAMN family. The first component corresponds to the parameters η of the transitory dynamics of the model, while the second corresponds to the parameters $(B_1', \beta')'$ of the cointegrating relationships.

The block diagonality of our asymptotic information matrix $S(\theta)$ implies that short-run and long-run dynamics are asymptotically independent of one another, so that in the analysis of one component the other can be treated as known. Of particular importance is the fact that we can estimate the cointegrating parameters as efficiently when treating the remaining parameters as unknown nuisance parameters as we could if the latter were known. The information matrix is not, however, block diagonal between the components associated with the cointegrating slope and intercept parameters, implying that we can estimate the slopes more efficiently if we know the intercepts than if we do not.

(b) One important consequence of this theorem is that the sequences of probability measures $\{P_{\theta,n}\}$ and $\{P_{\theta_n,n}\}$ are contiguous and therefore have the property that a 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 subsequently 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 $P_{\theta_n,n}$.

4. EFFICIENT AND ADAPTIVE ESTIMATION

In this section, we are concerned with the efficient estimation of our model. 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. We begin by formally defining the notion of an efficient estimator for our model and then briefly discuss the optimality properties of such an estimator. We then show how to compute an efficient estimator of the full parameter vector θ when the innovation density $p(\varepsilon)$ is known to the investigator. Finally, we consider adaptive estimation, deriving efficient estimators even when the innovation density is unknown.

We begin with the following definition.

https://doi.org/10.1017/S0266466698141026 Published online by Cambridge University Press

DEFINITION 4.1. *If the model is LAMN or LAN at* θ, *we call a sequence* $\{\hat{\theta}_n\}$ *of estimators* efficient *if*

$$\delta_n^{-1}(\hat{\theta}_n - \theta) - S_n^{-1}(\theta)W_n(\theta) = o_p(1), \tag{10}$$

where $W_n(\theta)$ *and* $S_n(\theta)$ *are defined in* (5) *and* (6).

Remark. We can see from this definition that an efficient 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 efficient estimator has the same asymptotic distribution as the MLE and thus shares the latter's optimality properties. Efficient estimators are optimal according to the locally asymptotically minimax (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, e.g., Ghosh, 1985, pp. 318–320). These estimators were termed *asymptotically centering* by Jeganathan (1995).

The computation of efficient estimators through a one-step iteration in the LAN case was analyzed by LeCam (1960), and the extension of the method to LAMN models was described by Jeganathan (1995). We begin with some sequence of preliminary estimators $\{\theta_n^*\}$ with the property that

$$\delta_n^{-1}(\theta_n^* - \theta) = O_p(1).$$

From a practical standpoint, there are several possible ways to obtain such estimators. We may, for example, employ the Gaussian maximum likelihood estimator of Ahn and Reinsel (1990). A simpler alternative would be to compute estimates (B_{1n}^*, B_{0n}^*) through an OLS regression of X_{1t} on X_{2t} and a constant vector, substitute these estimates into the right-hand side of (3), and then use OLS in (3) to compute (A_n^*, Φ_n^*).

We define θ_n^{**}, the *discretized* version of θ_n^*, as follows (cf. LeCam and Yang, 1990; Jeganathan, 1995, 1996).

DEFINITION 4.2. *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 nonempty because* $\theta_n^* \in \Theta$.

Remark. The θ_n^{**} constructed in this way preserve the properties of θ_n^* in the sense that $\theta_n^{**} \in \Theta$ and $\delta_n^{-1}(\theta_n^{**} - \theta) = O_p(1)$ for every $\theta \in \Theta$. The "trick" of using a discretized preliminary estimator is due to LeCam (1960) and allows us to prove that iterative estimators are efficient under quite general conditions.

The following lemma gives an expression for an efficient one-step iterative estimator computed using a discretized preliminary estimator.

LEMMA 4.3. *Suppose that* θ_n^{**} *is a* δ_n^{-1}*-consistent*, *discretized preliminary estimator*, *as defined earlier*. *Define* $\hat{S}_n$ *as any consistent estimator of* $S_n(\theta)$. *Then*

https://doi.org/10.1017/S0266466698141026 Published online by Cambridge University Press

$\hat{\theta}_n$ *as defined in equation* (11) *is an efficient estimator*:

$$\hat{\theta}_n = \theta_n^{**} + \delta_n \hat{S}_n^{-1} W_n(\theta_n^{**}). \tag{11}$$

Furthermore, *we have*

$$W_n(\theta_n) = W_n(\theta) - \hat{S}_n h_n + o_p(1) \tag{12}$$

for every bounded $\{h_n\}$ *and* $\theta_n = \theta + \delta_n h_n$.

Remark. We can see from (11) that our model can be efficiently estimated following the usual Newton–Raphson iteration. The efficiency of such estimators in LAN models was shown by LeCam (1960) and was extended to LAMN models by Jeganathan (1995). The result given by (12) follows from a similar result in Jeganathan (1995) and is important for our proof of adaptivity in Theorem 4.5.

The asymptotically efficient estimator derived in (11) is of no immediate use to us because of the assumption that the density $p(\varepsilon)$ is known to the investigator, an assumption we wish to avoid. We now 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 efficient estimator given by (11). 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 (5) and (6) to give us consistent estimators of the sample score and information, with which we can construct a one-step Newton–Raphson estimator similar in form to that given by (11).

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 i.i.d. 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 misbehavior 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 (1995), among others, and are also employed here.

Our first step in this section is to formulate a nonparametric kernel estimator of the score $\psi(\varepsilon_t)$. Our construction is a multivariate generalization of that developed by Kreiss (1987b) for univariate autoregressive moving average (ARMA) models with symmetric innovations. We introduce the following notation:

$$\pi(x,\sigma) = (\sigma\sqrt{2\pi})^{-q} \exp(-|x|^2/2\sigma^2),$$

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

and let $\hat{p}^{j}_{\sigma,t}(x,\theta)$ be the partial derivative of $\hat{p}_{\sigma,t}(x,\theta)$ with respect to the jth element of x, for all $j = 1,\ldots,q$. In these expressions, $\pi(x,\sigma)$ is a q-dimensional

Gaussian kernel with smoothing parameter σ. The larger is σ, the smoother is the estimate of p. We further define

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

where $c^j_n \to \infty$, $\alpha^j_n \to \infty$, $\sigma \to 0$, and $m^j_n \to 0$. Because $\hat{p}^j_{\sigma,t}$ is an estimator of the jth element of the vector of partial derivatives of $p(\varepsilon)$, $\hat{\psi}^j_{n,t}$ is an estimator of $\psi^j(\varepsilon)$, the jth element of the score vector $\psi(\varepsilon)$. The trimming parameters m^j_n, α^j_n, and c^j_n serve to omit extreme outlying observations that would distort the behavior of our score estimator. Our derivation of an adaptive estimator hinges on showing that the mean integrated squared difference between $\hat{\psi}^j_{n,t}$ and $\psi^j(\varepsilon)$ converges to zero. The preceding conditions on the trimming parameters and on the smoothing parameter σ are used to prove this consistency.

With this notation, we can define our q-dimensional score estimator as $\hat{\psi}_{n,t}(x,\theta) = (\hat{\psi}^1_{n,t}(x,\theta), \ldots, \hat{\psi}^q_{n,t}(x,\theta))'$. Assuming that $p(\varepsilon)$ is symmetric about the origin implies that $\psi(\varepsilon)$ is antisymmetric about the origin. Furthermore, $\hat{\psi}_{n,t}(x,\theta)$ is antisymmetric about the origin in x by construction. We can use this error density score estimator to derive the following consistent estimator of the score function of the model:

$$\hat{W}_n(\theta) = -\sum_{t=1}^{n} \delta_n H_{t-1}(\theta) \hat{\psi}_{n,t}(\varepsilon_t(\theta), \theta), \tag{13}$$

where we recall that the score of the model is

$$W_n(\theta) = -\sum_{t=1}^{n} \delta_n H_{t-1}(\theta) \psi(\varepsilon_t(\theta)). \tag{14}$$

The key step in obtaining an adaptive estimator for our model is to show that $\hat{W}_n(\theta)$ is a consistent estimator of $W_n(\theta)$. The proof of this consistency result will be facilitated by the introduction of the following notation:

$$W^j_n(\theta) = -\sum_{t=1}^{n} \delta_n H^j_{t-1}(\theta) \psi^j(\varepsilon_t(\theta))$$

and

$$\hat{W}^j_n(\theta) = -\sum_{t=1}^{n} \delta_n H^j_{t-1}(\theta) \hat{\psi}^j_{n,t}(\varepsilon_t(\theta), \theta),$$

where $H^j_{t-1}(\theta)$ is the jth column of $H_{t-1}(\theta)$ and we recall that $\psi^j(\varepsilon_t(\theta))$ is the jth element of $\psi(\varepsilon_t(\theta))$. Now, we seek to prove that

$$\hat{W}_n(\theta_n) - W_n(\theta_n) = o_p(1), \tag{15}$$

https://doi.org/10.1017/S0266466698141026 Published online by Cambridge University Press

for which it is sufficient to show that

$$\hat{W}_n^j(\theta_n) - W_n^j(\theta_n) = o_p(1) \quad \forall j = 1,\ldots,q \tag{16}$$

because

$$\hat{W}_n(\theta_n) - W_n(\theta_n) = \sum_{j=1}^{q} (\hat{W}_n^j(\theta_n) - W_n^j(\theta_n)).$$

Our desired consistency result (16), and hence (15), is a consequence of the following theorem, which follows from Proposition 15 in Jeganathan (1988) and is similar to Theorem 17 in Jeganathan (1995).

THEOREM 4.4. *For every* $j = 1,\ldots,q$, *assume that* $c_n^j \to \infty$, $\alpha_n^j \to \infty$, $m_n^j \to 0$, $\sigma \to 0$, $\sigma c_n^j \to 0$, *and* $n^{-1}\alpha_n^j \sigma^{-(4+q)} \to 0$. *Then, for every bounded* $\{h_n\}$ (*where we recall that* $h_n = \delta_n^{-1}(\theta_n - \theta)$), *it follows that, for every* $j = 1,\ldots,q$,

$$\hat{W}_n^j(\theta_n) - W_n^j(\theta_n) = o_p(1). \tag{17}$$

Furthermore, we have

$$\hat{W}_n(\theta_n) = \hat{W}_n(\theta) - S_n(\theta)h_n + o_p(1). \tag{18}$$

It follows from (17) that consistent estimation of the score for the sample is possible, i.e., that (15) holds.

With a consistent score estimator in hand, we need only find a consistent estimator of the scaled sample information matrix $S_n(\theta)$ to be able to derive an adaptive estimator. Recall that

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

where $S_n^{jl}(\theta) = \sum_{t=1}^{n} \delta_n H_{t-1}^j(\theta) H_{t-1}^l(\theta)'\delta_n$. For the moment, we shall assume the existence of a consistent estimator $\hat{\omega}_{jl}$ of ω_{jl} (the construction of such an estimator is discussed later). Now, given our sequence $\{\theta_n\}$ of local approximations to θ, we have $S_n^{jl}(\theta_n) - S_n^{jl}(\theta) = o_p(1)$ for every $j, l = 1,\ldots,q$. We now define the estimator

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

From the preceding discussion, it follows that

$$\hat{S}_n(\theta_n) - S_n(\theta) = o_p(1). \tag{19}$$

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

https://doi.org/10.1017/S0266466698141026 Published online by Cambridge University Press

THEOREM 4.5. *The estimator* $\tilde{\theta}_n$ *given in* (20) *is adaptive for our model*:

$$\tilde{\theta}_n = \theta_n^{**} + \delta_n \hat{S}_n(\theta_n^{**})^{-1} \hat{W}_n(\theta_n^{**}). \quad \textbf{(20)}$$

In other words,

$$\delta_n^{-1}(\hat{\theta}_n - \tilde{\theta}_n) = o_p(1), \quad \textbf{(21)}$$

where $\hat{\theta}_n$ *is given by* (11).

Remarks.

(a) In deriving our estimator $\hat{S}_n(\theta_n)$ of $S_n(\theta)$, we assumed the existence of a consistent estimator of ω_{jt}. The argument of Kreiss (1987b, p. 123) can be used to show that $\hat{\omega}_{jl}$ as defined in (22) is such an estimator (the proof uses the fact that $\{P_{\theta,n}\}$ and $\{P_{\theta_n,n}\}$ are contiguous):

$$\hat{\omega}_{jl} = n^{-1} \sum_{t=1}^{n} \hat{\psi}_{n,t}^{j}(\varepsilon_t(\theta_n^{**}), \theta_n^{**}) \hat{\psi}_{n,t}^{l}(\varepsilon_t(\theta_n^{**}), \theta_n^{**}). \quad \textbf{(22)}$$

(b) It may seem odd that the convergence results (17)–(19) are useful in our proof of Theorem 4.5 because these results prove consistency for statistics evaluated at a *deterministically* consistent sequence $\{\theta_n\}$, whereas it would seem that we require such results for these statistics evaluated at the *stochastically* consistent sequence $\{\theta_n^{**}\}$. However, this difficulty disappears as a result of the fact that $\{\theta_n^{**}\}$ is discretized. According to Lemma 4.4 of Kreiss (1987b, p. 120), the sequence of statistics $\{T_n(\theta_n^{**})\} = o_p(1)$ if the sequence $\{T_n(\theta_n)\} = o_p(1)$.

(c) We now derive the asymptotic distribution of the adaptive estimator $\tilde{\theta}_n$. Because $\tilde{\theta}_n$ is an efficient estimator, we can use the definition of the latter in (10), combined with (8), to obtain

$$L(\delta_n^{-1}(\tilde{\theta}_n - \theta)|P_{\theta,n}) \Rightarrow L(MN(0, S(\theta)^{-1})), \quad \textbf{(23)}$$

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

$$L(\sqrt{n}(\tilde{\eta}_n - \eta)|P_{\theta,n}) \Rightarrow L(N(0, \Omega^{-1} \otimes E[M_t M_t']^{-1})), \quad \textbf{(24)}$$

$$L(n(\tilde{\beta}_n - \beta)|P_{\theta,n}) \Rightarrow L\left(MN\left(0, (A'\Omega A)^{-1} \otimes \left(\int_0^1 Z_2 Z_2' - \bar{Z}_2 \bar{Z}_2'\right)^{-1}\right)\right), \quad \textbf{(25)}$$

and

$$L(n^{1/2}(\tilde{B}_{1n} - B_1)|P_{\theta,n}) \Rightarrow L\left(MN\left(0, (A'\Omega A)^{-1} \otimes \left(1 + \bar{Z}_2'\left(\int_0^1 Z_2 Z_2' - \bar{Z}_2 \bar{Z}_2'\right)\bar{Z}_2\right)^{-1}\right)\right). \quad \textbf{(26)}$$

(d) 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 the results in the preceding remark. The co-

https://doi.org/10.1017/S0266466698141026 Published online by Cambridge University Press

variance matrix of the scaled and centered Gaussian pseudo-MLE of the slope parameters β is

$$(A'\Sigma_\varepsilon^{-1}A)^{-1} \otimes \left(\int_0^1 Z_2 Z_2' - \bar{Z}_2 \bar{Z}_2' \right)^{-1}. \quad \textbf{(27)}$$

If $p(\varepsilon)$ actually is Gaussian, then $\Sigma_\varepsilon^{-1} = \Omega$, so that (27) is identical to the covariance matrix in (25). However, if Gaussianity fails, then $\Omega > \Sigma_\varepsilon^{-1}$, so that $(A'\Sigma_\varepsilon^{-1}A)^{-1} \geq (A'\Omega A)^{-1}$ and the Gaussian estimator is inefficient. The degree of inefficiency can be measured using the following ratio of generalized variances (cf. Mitchell, 1989), where AER is the mnemonic for asymptotic efficiency ratio:

$$\begin{aligned} \text{AER} &= \frac{\left[\det\left[(A'\Omega A)^{-1} \otimes \left(\int_0^1 Z_2 Z_2' - \bar{Z}_2 \bar{Z}_2'\right)^{-1}\right]\right]^{1/r(q-r)}}{\left[\det\left[(A'\Sigma_\varepsilon^{-1}A)^{-1} \otimes \left(\int_0^1 Z_2 Z_2' - \bar{Z}_2 \bar{Z}_2'\right)^{-1}\right]\right]^{1/r(q-r)}} \\ &= \frac{[\det(A'\Sigma_\varepsilon^{-1}A)]^{(q-r)/r(q-r)} \left[\det\left(\int_0^1 Z_2 Z_2' - \bar{Z}_2 \bar{Z}_2'\right)\right]^{r/r(q-r)}}{[\det(A'\Omega A)]^{(q-r)/r(q-r)} \left[\det\left(\int_0^1 Z_2 Z_2' - \bar{Z}_2 \bar{Z}_2'\right)\right]^{r/r(q-r)}} \\ &= \left[\frac{\det(A'\Sigma_\varepsilon^{-1}A)}{\det(A'\Omega A)}\right]^{1/r}. \end{aligned} \quad \textbf{(28)}$$

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

$$p(\varepsilon) = |\det \Sigma|^{-1/2} f^*(\varepsilon'\Sigma^{-1}\varepsilon). \quad \textbf{(29)}$$

Then, as shown by Mitchell (1989),

$$\Omega = 4a_p k_\varepsilon \Sigma_\varepsilon^{-1}, \quad \textbf{(30)}$$

where a_p is defined in Mitchell (1989, p. 296). (In the Gaussian case, $k_\varepsilon = 1$ and $a_p = \frac{1}{4}$, giving us the familiar result that $\Omega = \Sigma_\varepsilon^{-1}$.) Substituting (30) into the last line of (28) gives us

$$\begin{aligned} \text{AER} &= \left[\frac{\det(A'\Sigma_\varepsilon^{-1}A)}{\det(4a_p k_\varepsilon A'\Sigma_\varepsilon^{-1}A)}\right]^{1/r} \\ &= (4a_p k_\varepsilon)^{-1}. \end{aligned}$$

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 i.i.d. observations is 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 (we can see from (23)–(26) that a similar argument can be made with respect to any of the subvectors of $\tilde{\theta}_n$, including $\tilde{\theta}_n$ itself).

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

$$\text{AER} = (1 - 2/\nu)\cdot(1 + 2/(\nu + q)).$$

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

(e) The preceding remark suggests that the adaptive estimator will deliver superior relative performance the larger is the system. At least for Student t errors, the asymptotic efficiency ratio is decreasing in the dimensionality q. However, this improvement in asymptotic performance as q increases is likely to be significantly mitigated by a deterioration in finite-sample performance as a result of the notorious "curse of dimensionality" that is common in nonparametric estimation problems. We are using a multivariate kernel to estimate the innovation density $p(\varepsilon)$. The rate of convergence of multivariate kernel density estimators is well known to be negatively related to the dimensionality of the data. This consideration suggests that our procedure is only useful in the estimation of fairly small systems.

However, there is a way to avert the curse of dimensionality if we are willing to assume that our innovation density $p(\varepsilon)$ belongs to the elliptically symmetric family.[4] This is because, as can be seen from (29), the multivariate density $p(\varepsilon)$ can then be written in terms of a function f^* whose argument is a *scalar* transformation of the data. We can use this fact to reduce our kernel density estimation problem to a univariate one. Stute and Werner (1991) showed that kernel estimators that correctly impose elliptical symmetry restrictions have a rate of convergence that is independent of dimension. Such dimensionality reduction was employed by Bickel (1982) in the estimation of a multivariate location model and by Hodgson, Linton, and Choo (1997) in the estimation of seemingly unrelated regression (SUR) models. Extension of this work to our model and to the triangular model estimator of Hodgson (in press) should be feasible and is a topic for future research.

(f) Our trimmed, leave-one-out, Gaussian kernel score estimator $\hat{\psi}^j_{n,t}(x,\theta)$ is just one among several valid possibilities. Jeganathan (1995) used a similar estimator that does not require the omission of an observation, and Schick (1987) developed a logistic kernel estimator that requires neither trimming nor the omission of an observation and that could presumably be shown to apply to our model. Of greater interest would be the investigation of the applicability to nonstationary models of the estimators of Kreiss (1987a) and Drost and Klaassen (in press), which are applied respectively to AR and GARCH models and allow for asymmetric error densities.

(g) Some evidence on the finite-sample performance of the estimator developed in this paper is provided in Hodgson (1995), which reported the results of Monte Carlo simulations and the application of our method to the estimation of a forward exchange market unbiasedness model. The simulation study compares the performances of our adaptive estimator and Johansen's (1988) Gaussian pseudo-MLE in bivariate models for various sample sizes and error densities. The adaptive estimator exhibits significant improvements in a truncated mean-squared error criterion for all non-Gaussian densities (Student's t, variance-contaminated mixed normal, and bimodal mixed normal) and sample sizes (100, 250, and 500) considered. The magnitude of the gains is found to be fairly insensitive to smoothing and trimming parameter variation, with the Silverman (1986) rule-of-thumb bandwidth generally yielding good results.

The empirical study reported in Hodgson (1995) evaluates a variant of the forward exchange market unbiasedness hypothesis according to which the k-period–ahead forward exchange rate between two currencies in period $t - k$ should be an unbiased forecast of the spot rate between them in period t. This hypothesis is generally tested by estimating a cointegrating relationship between the logarithms of the spot rate and the lagged forward rate, with the unbiasedness hypothesis positing zero intercept and unit slope parameters (cf. Baillie and Bollerslev, 1989; Baillie, Lippens, and McMahon, 1983; Barnhart and Szakmary, 1991; Corbae, Lim, and Ouliaris, 1992; Hakkio and Rush, 1989; Phillips, 1993; Phillips et al., 1996). Hodgson (1995) estimated such a model using a sample of 650 daily observations on the Canada–U.S. spot and 90-day forward exchange rates using an error correction model and a maintained hypothesis of zero intercept. The Johansen (1988) methodology yields a slope estimate of 0.937 with an estimated standard error of 0.033. The adaptive estimator, using this as the preliminary estimate and using the Silverman (1986) rule-of-thumb bandwidth, yields an estimate of 0.995 with asymptotic standard error of 0.026, a result that is highly insensitive to smoothing and trimming parameter variation.

5. 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 estimators 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. We have also cited the simulation and empirical results reported in Hodgson's (1995) practical implementation of the procedure.

As they stand, the theory and methods developed here have the potential to be of substantial value to practitioners. Nevertheless, further developments seem desirable. The relaxation of the symmetry assumption (possibly along the lines of Kreiss, 1987b) 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 could 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.

We might also want to extend the model to allow for the presence of higher order dependence, such as ARCH effects, in the innovations. Indeed, the innovations to the error correction model of the daily exchange rate data analyzed in

https://doi.org/10.1017/S0266466698141026 Published online by Cambridge University Press

Hodgson (1995) appear to be subject to volatility clustering. However, even if unmodeled conditional or unconditional heterogeneity is present in the innovations, results derived in Hodgson (1996) suggest that the adaptive estimator constructed under the false assumption that the innovations are i.i.d. will still have desirable robustness properties and will be asymptotically more efficient than the Gaussian pseudo-MLE.

NOTES

1. It follows from this assumption that $0 < \lambda^2 \equiv \int |\psi(\varepsilon)|^2 p(\varepsilon)\, d\varepsilon < \infty$, because $\lambda^2 = \mathrm{tr}(\Omega)$.

2. Note that B_0 is the coefficient matrix of the reduced form representation of the cointegrating vectors that is analogous to the reduced form representation of a simultaneous equations model. This representation is valid under the assumption that the coefficient matrix on X_{1t} in the structural form of the cointegrating vectors is nonsingular. Our identification assumption implies that the parameters of this structural form can be uniquely recovered from B_0.

3. Note that $\sum_{t=1}^{n} \delta_n H_{t-1}(\theta)\Omega H_{t-1}(\theta)'\delta_n$ is not, strictly speaking, the sample information (as pointed out by a referee), because it contains the population quantity Ω. However, for ease of exposition we shall continue to refer to it as the sample information.

4. There is an intimate relationship between the assumption of elliptical symmetry of asset returns and mean-variance asset pricing theory. See Chamberlain (1983), Owen and Rabinovitch (1983), and Ingersoll (1987, Ch. 4).

REFERENCES

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

Ahn, S.K. & G.C. Reinsel (1990) Estimation for partially nonstationary multivariate autoregressive models. *Journal of the American Statistical Association* 85, 813–823.

Baillie, R.T. & T. Bollerslev (1989) Common stochastic trends in a system of exchange rates. *Journal of Finance* 44, 167–181.

Baillie, R.T., R.E. Lippens, & P.C. McMahon (1983) Testing rational expectations and efficiency in the foreign exchange market. *Econometrica* 51, 553–563.

Barnhart, S.W. & A.C. Szakmary (1991) Testing the unbiased forward rate hypothesis: Evidence on unit roots, co-integration, and stochastic coefficients. *Journal of Financial and Quantitative Analysis* 26, 245–267.

Beran, R. (1974) Asymptotically efficient rank estimates in location models. *Annals of Statistics* 2, 63–74.

Bickel, P.J. (1982) On adaptive estimation. *Annals of Statistics* 10, 647–671.

Chamberlain, G. (1983) A characterization of the distributions that imply mean variance utility functions. *Journal of Economic Theory* 29, 185–201.

Corbae, D., K.-G. Lim, & S. Ouliaris (1992) On cointegration and tests of forward market unbiasedness. *Review of Economics and Statistics* 74, 728–732.

Drost, F.C. & C.A.J. Klaassen (in press) Efficient estimation in semiparametric GARCH models. *Journal of Econometrics.*

Engle, R.F. & C.W.J. Granger (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.

Friedman, B.M. & K.N. Kuttner (1992) Money, income, prices, and interest rates. *American Economic Review* 82, 472–492.

Ghosh, J.K. (1985) Efficiency of estimates—part I. *Sankhyā* 47A, 310–325.

Hajek, J. & Z. Sidak (1967) *Theory of Rank Tests.* New York: Academic Press.

Hakkio, C.S. & M. Rush (1989) Market efficiency and cointegration: An application to the sterling and deutschemark exchange markets. *Journal of International Money and Finance* 8, 75–88.

Hodgson, D.J. (in press) Adaptive estimation of cointegrating regressions with ARMA errors. *Journal of Econometrics*.

Hodgson, D.J. (1995) Adaptive Estimation of Cointegrated Models: Simulation Evidence and an Application to the Forward Exchange Market. RCER Working paper 409, University of Rochester.

Hodgson, D.J. (1996) Robust Semiparametric Estimation in the Presence of Heterogeneity of Unknown Form. RCER Working paper 416, University of Rochester.

Hodgson, D.J., O. Linton, & E. Choo (1997) Estimation in Multivariate Time Series Regression Models with Elliptically Symmetric Errors. Unpublished manuscript.

Hsieh, D.A. & C.F. Manski (1987) Monte Carlo evidence on adaptive maximum likelihood estimation of a regression. *Annals of Statistics* 15, 541–551.

Ingersoll, J.E. (1987) *Theory of Financial Decision Making.* Savage, Maryland: Rowman & Littlefield.

Jeganathan, P. (1988) Some Aspects of Asymptotic Theory with Applications to Time Series Models. Technical report 166, University of Michigan.

Jeganathan, P. (1995) Some aspects of asymptotic theory with applications to time series models. *Econometric Theory* 11, 818–887.

Jeganathan, P. (1997) On asymptotic inference in linear cointegrated time series systems. *Econometric Theory* 13, 692–745.

Johansen, S. (1988) Stochastic 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. & K. Juselius (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. & K. Juselius (1992) Testing structural hypotheses in a multivariate cointegration analysis of the PPP and the UIP for the 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. & K. Neusser (1990) Cointegration in a macroeconomic system. *Journal of Applied Econometrics* 5, 351–365.

LeCam, L. (1960) Locally asymptotically normal families of distributions. *University of California Publications in Statistics* 3, 37–98.

LeCam, L. & G.L. Yang (1990) *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. & S.B. White (1993) A comparison of some robust, adaptive, and partially adaptive estimators of regression models. *Econometric Reviews* 12, 103–124.

McGuirk, A.M., J. Robertson, & A. Spanos (1993) Modelling 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 α-connections for the general multivariate elliptic distribution. *Annals of the Institute of Mathematical Statistics* (Tokyo) 41, 289–304.

Mittnik, S. & S.T. Rachev (1993) Modelling asset returns with alternative stable distributions. *Econometric Reviews* 12, 261–330.

Owen, J. & R. Rabinovitch (1983) On the class of elliptical distributions and their applications to the theory of portfolio choice. *Journal of Finance* 38, 745–752.

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, Yale University.

Phillips, P.C.B. (1995) Robust nonstationary regression. *Econometric Theory* 11, 912–951.

Phillips, P.C.B. & S.N. Durlauf (1986) Multiple time series regression with integrated processes. *Review of Economic Studies* 53, 473–495.

Phillips, P.C.B., J.W. McFarland, & P.C. McMahon (1996) Robust tests of forward exchange market efficiency with empirical evidence from the 1920's. *Journal of Applied Econometrics* 11, 1–22.

Phillips, P.C.B. & J. Park (1988) Asymptotic equivalence of ordinary least squares and generalized least squares in regressions with integrated processes. *Journal of the American Statistical Association* 83, 111–115.

Silverman, B. (1986) *Density Estimation for Statistics and Data Analysis.* London: Chapman & Hall.

Steigerwald, D. (1992) On the finite sample behaviour of adaptive estimators. *Journal of Econometrics* 54, 371–400.

Stein, C. (1956) Efficient nonparametric estimation and testing. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics & Probability* 1, 187–196.

Stone, C. (1975) Adaptive maximum likelihood estimation of a location parameter. *Annals of Statistics* 3, 267–284.

Stute, W. & U. Werner (1991) Nonparametric estimation of elliptically contoured densities. In G. Roussas (ed.), *Nonparametric Functional Estimation and Related Topics*, pp. 173–190. Amsterdam: Kluwer Academic Publishers.

Velu, R.P., G.C. Reinsel, & D.W. Wichern (1986) Reduced rank models for multiple series. *Biometrika* 73, 105–118.

White, H. (1984) *Asymptotic Theory for Econometricians.* New York: Academic Press.

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. By writing $h_n = (h'_{\alpha n}, h'_{\phi n}, h'_{B_1 n}, h'_{\beta n})'$, we decompose h_n into components of respective dimension qr, $q^2(k-1)$, r, and $r(q-r)$. These components can be thought of as vectorizations of the matrices a_n, ϕ_n, b_{1n}, and b_n, whose respective dimensions are $q \times r$, $q \times q(k-1)$, $r \times 1$, and $r \times (q-r)$, and where $\{\phi_n\}$ is a bounded sequence. Using this notation and (3), we have

$$g_{t-1}(\theta) = X_{t-1} + A[X_{1,t-1} - B_1 - B_0 X_{2,t-1}] + \Phi Y_{t-1} \quad \textbf{(A.1)}$$

https://doi.org/10.1017/S0266466698141026 Published online by Cambridge University Press

and

$$g_{t-1}(\theta_n) = X_{t-1} + (A + n^{-1/2}a_n)[X_{1,t-1} - (B_1 + n^{-1/2}b_{1n}) - (B_0 + n^{-1}b_n)X_{2,t-1}]$$
$$+ (\Phi + n^{-1/2}\phi_n)Y_{t-1}$$
$$= g_{t-1}(\theta) + n^{-1/2}a_n[X_{1,t-1} - B_1 - B_0X_{2,t-1}] - n^{-1}Ab_nX_{2,t-1}$$
$$- n^{-1/2}Ab_{1n} - n^{-3/2}a_nb_nX_{2,t-1} + n^{-1/2}\phi_nY_{t-1} - n^{-1}a_nb_{1n}. \quad \textbf{(A.2)}$$

Subtracting (A.1) from (A.2) yields

$$d_t(\theta_n,\theta) = n^{-1/2}(a_nV_{t-1} + \phi_nY_{t-1} - Ab_{1n}) - n^{-3/2}a_nb_nX_{2,t-1}$$
$$- n^{-1}(Ab_nX_{2,t-1} + a_nb_{1n}). \quad \textbf{(A.3)}$$

Noting that $d_t(\theta_n,\theta)$ is a vector, and applying the formula for vectorizing products of matrices, we can rewrite (A.3) as follows:

$$d_t(\theta_n,\theta)' = h'_{\alpha n}n^{-1/2}(I_q \otimes V_{t-1}) + h'_{\varphi n}n^{-1/2}(I_q \otimes Y_{t-1}) - h'_{B_1 n}n^{-1/2}A'$$
$$- h'_{\beta n}n^{-1}(A' \otimes X_{2,t-1}) - n^{-3/2}(a_nb_nX_{2,t-1})' - n^{-1}(a_nb_{1n})',$$

from which (4) immediately follows. ■

Proof of Theorem 3.2. We begin by quoting Condition A.1 and Proposition A.2, which follow, as stated by Jeganathan (1995, p. 848). According to Proposition A.2, the likelihood ratios $\Lambda_n(\theta_n,\theta)$ have the asymptotic quadratic approximation given by (7) if Condition A.1 holds. We therefore proceed to verify that Condition A.1 is satisfied for our model. To save space, we only report the proof of (7) for the case where it is known that $B_1 = 0$. The extension to our more general model with nonzero intercepts follows very similar lines. In the course of checking Condition A.1, we shall prove the second part of the weak convergence result (8). The proof of the first part of (8), for the general model, is reported immediately following our verification of Condition A.1.

Condition A.1 and Proposition A.2 are as follows.

Condition A.1. 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

$$\sum_{t=1}^{n} |d_t(\theta_n,\theta) - h'_n\delta_nH_{t-1}(\theta)|^2 = o_p(1) \quad \textbf{(A.4)}$$

such that

$$\sum_{t=1}^{n} |h'_n\delta_nH_{t-1}(\theta)|^2 = O_p(1) \quad \textbf{(A.5)}$$

and

$$\max_{t\in\{1,\dots,n\}} |h'_n\delta_nH_{t-1}(\theta)|^2 = o_p(1). \quad \textbf{(A.6)}$$

PROPOSITION A.2. *Under our assumptions*, *Condition A*.1 *implies the quadratic approximation given by* (7).

Verification of Condition A.1. We now proceed to verify that equations (A.4)–(A.6) are satisfied for our model.

To verify (A.4), we must check that

$$n^{-3}\sum_{t=1}^{n}|X'_{2,t-1}b'_n a'_n|^2 = o_p(1). \tag{A.7}$$

But (A.7) will hold as a result of the fact that

$$\begin{aligned} n^{-2}\sum_{t=1}^{n}|X'_{2,t-1}b'_n a'_n|^2 &= n^{-2}\sum_{t=1}^{n} a_n b_n X_{2,t-1}X'_{2,t-1}b'_n a'_n \\ &= O_p(1) \end{aligned}$$

(cf. Phillips and Durlauf, 1986).

To establish (A.5), we shall verify (A.8),

$$h'_n\left[\sum_{t=1}^{n}\delta_n H_{t-1}(\theta)\Omega H_{t-1}(\theta)'\delta_n\right]h_n = O_p(1), \tag{A.8}$$

in the process of which we also prove the second part of (8). The following calculations will prove useful:

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

from which it follows that

$$\begin{aligned} &\sum_{t=1}^{n}\delta_n H_{t-1}(\theta)\Omega H_{t-1}(\theta)'\delta_n \\ &\quad = \sum_{t=1}^{n}\begin{bmatrix} n^{-1}\Omega \otimes M_{t-1}M'_{t-1} & -n^{-3/2}\Omega A \otimes M_{t-1}X'_{2,t-1} \\ -n^{-3/2}A'\Omega \otimes X_{2,t-1}M'_{t-1} & n^{-2}A'\Omega A \otimes X_{2,t-1}X'_{2,t-1} \end{bmatrix}. \end{aligned} \tag{A.9}$$

We proceed to show that the matrix on the right-hand side of (A.9) is $O_p(1)$. To achieve this, we show that the following results hold:

$$n^{-1}\sum_{t=1}^{n}\Omega \otimes M_{t-1}M'_{t-1} = \Omega \otimes E[M_t M'_t] + o_p(1), \tag{A.10}$$

$$n^{-3/2}\sum_{t=1}^{n}\Omega A \otimes M_{t-1}X'_{2,t-1} = o_p(1), \tag{A.11}$$

and

$$n^{-2}\sum_{t=1}^{n}A'\Omega A \otimes X_{2,t-1}X'_{2,t-1} \Rightarrow A'\Omega A \otimes \int_0^1 Z_2 Z'_2. \tag{A.12}$$

https://doi.org/10.1017/S0266466698141026 Published online by Cambridge University Press

We obtain (A.10) as a consequence of Lemma 1(iv) of Ahn and Reinsel (1990, p. 815), while (A.11) and (A.12) can be shown to follow from Lemmas 3.1(e) and (b), respectively, of Phillips and Durlauf (1986, p. 477). In consequence of (A.10)–(A.12), we have

$$\sum_{t=1}^{n} \delta_n H_{t-1}(\theta)\Omega H_{t-1}(\theta)'\delta_n \Rightarrow \begin{bmatrix} \Omega \otimes E[M_t M_t'] & 0 \\ 0 & A'\Omega A \otimes \int_0^1 Z_2 Z_2' \end{bmatrix}. \tag{A.13}$$

Incidentally, (A.13) indicates that $S_n(\theta) \Rightarrow S(\theta)$, so that the second part of the convergence result (8) in the theorem is established.

To complete our proof of (7), we must verify (A.6), which can be rewritten as

$$\max_{t\in\{1,\ldots,n\}} h_n' \begin{bmatrix} n^{-1} I \otimes M_{t-1}M_{t-1}' & -n^{-3/2} A \otimes M_{t-1}X_{2,t-1}' \\ -n^{-3/2} A' \otimes X_{2,t-1}M_{t-1}' & n^{-2}A'A \otimes X_{2,t-1}X_{2,t-1}' \end{bmatrix} h_n = o_p(1). \tag{A.14}$$

We verify (A.14) by checking the following three conditions:

$$\max_{t\in\{1,\ldots,n\}} |n^{-1} M_{t-1}^j M_{t-1}^l| = o_p(1)\ \forall j,l,$$

$$\max_{t\in\{1,\ldots,n\}} |n^{-3/2} M_{t-1}^j X_{2,t-1}^l| = o_p(1)\ \forall j,l,$$

$$\max_{t\in\{1,\ldots,n\}} |n^{-2} X_{2,t-1}^j X_{2,t-1}^l| = o_p(1)\ \forall j,l,$$

where the superscripts j and l represent the jth and lth elements of the respective vectors. Using the inequality

$$\max_{t\in\{1,\ldots,n\}} |x_t y_t| \le \max_{t\in\{1,\ldots,n\}} |x_t| \max_{t\in\{1,\ldots,n\}} |y_t|$$

for scalar random variables x_t and y_t, (A.14) follows because $\max_{t\in\{1,\ldots,n\}}|M_t^j| = O_p(n^{1/2})\ \forall j$ and $\max_{t\in\{1,\ldots,n\}}|X_t^j| = O_p(n)\ \forall j$.

We now complete our proof of Theorem 3.2 by verifying the first part of the convergence result (8) for the general model (we can verify the second part of (8) for the general model along the same lines as we derived (A.13) earlier). To analyze the limit distribution of the score $W_n(\theta)$, we write it as follows:

$$W_n(\theta) = \sum_{t=1}^{n} \begin{bmatrix} -n^{-1/2}\psi(\varepsilon_t) \otimes M_{t-1} \\ n^{-1/2} A'\psi(\varepsilon_t) \\ n^{-1} A'\psi(\varepsilon_t) \otimes X_{2,t-1} \end{bmatrix}. \tag{A.15}$$

We begin with an analysis of the first component of (A.15),

$$-n^{-1/2} \sum_{t=1}^{n} \psi(\varepsilon_t) \otimes M_{t-1}.$$

We can use a central limit theorem for stationary and ergodic processes (e.g., White, 1984, p. 118) to show that

$$L\left(-n^{-1/2} \sum_{t=1}^{n} \psi(\varepsilon_t) \otimes M_{t-1} | P_{\theta,n}\right) \Rightarrow L(N(0, \Omega \otimes E[M_t M_t'])). \tag{A.16}$$

As for the second and third components of (A.15), we use Lemma 3.1(e) of Phillips and Durlauf (1986) to obtain

$$\tilde{\delta}_n \sum_{t=1}^{n} A'\psi(\varepsilon_t) \otimes \begin{pmatrix} 1 \\ X_{2,t-1} \end{pmatrix} \Rightarrow \int_0^1 A'\, dZ_1 \otimes \begin{pmatrix} 1 \\ Z_2 \end{pmatrix},$$

where Z_1 is the Brownian motion with covariance matrix Ω generated by the scaled partial sums of the $\{\psi(\varepsilon_t)\}$ and $\tilde{\delta}_n = \text{diag}[n^{-1/2}I_r, n^{-1}I_{m-s-r}]$. As shown in Lemma A.3, which follows, the Brownian motion processes $A'Z_1$ and Z_2 are independent, from which it follows that

$$L\left(\int_0^1 A'\, dZ_1 \otimes \begin{pmatrix} 1 \\ Z_2 \end{pmatrix}\right) = L\left(MN\left(0, \begin{bmatrix} A'\Omega A & A'\Omega A \otimes \bar{Z}_2' \\ A'\Omega A \otimes \bar{Z}_2 & A'\Omega A \otimes \int_0^1 Z_2 Z_2' \end{bmatrix}\right)\right), \tag{A.17}$$

as shown by Phillips and Park (1988). But (A.17) will hold as a result of the following lemma.

LEMMA A.3. *The Brownian motion processes $A'Z_1$ and Z_2 are independent*.

Proof of Lemma A.3. Denote the Brownian motion process generated by scaled partial sums of the innovations $\{\varepsilon_t\}$ by Z_ε, with covariance matrix Σ_ε. We begin by showing that the covariance matrix between Z_1 and Z_ε is the negative of an identity matrix. Because $\{\varepsilon_t\}$ and $\{\psi(\varepsilon_t)\}$ are both i.i.d. zero-mean processes, the covariance between Z_1 and Z_ε is equal to $E[\psi(\varepsilon)\varepsilon']$, so we must prove that

$$E[\psi^j(\varepsilon)\varepsilon^{-j}] = 0 \tag{A.18}$$

and

$$E[\psi^j(\varepsilon)\varepsilon^j] = -1 \tag{A.19}$$

for every $j = 1,\ldots,q$, where the superscripts denote the jth elements of the respective vectors, and $\varepsilon^{-j} = (\varepsilon_1,\ldots,\varepsilon_{j-1},\varepsilon_{j+1},\ldots,\varepsilon_q)$. We denote the marginal density of ε^{-j} by $\tilde{p}(\varepsilon^{-j})$.

As in Jeganathan (1995), we can use the law of iterated expectations to show that (A.18) is implied by

$$E[\psi^j(\varepsilon)|\varepsilon^{-j}] = 0$$

or

$$\int \psi^j(\varepsilon)(p(\varepsilon)/\tilde{p}(\varepsilon^{-j}))\, d\varepsilon^j = 0,$$

which holds by Lemma (a) on p. 19 of Hajek and Sidak (1967).

By the law of iterated expectations, (A.19) will hold if we can show that

$$E[\psi^j(\varepsilon)\varepsilon^j|\varepsilon^{-j}] = -1. \tag{A.20}$$

We can write the left-hand side of (A.20) as

$$\int \frac{p^j(\varepsilon)}{p(\varepsilon)}\, \varepsilon^j\, \frac{p(\varepsilon)}{\tilde{p}(\varepsilon^{-j})}\, d\varepsilon^j = \int \varepsilon^j\, \frac{p^j(\varepsilon)}{\tilde{p}(\varepsilon^{-j})}\, d\varepsilon^j.$$

But our earlier moment conditions on ε and $\psi(\varepsilon)$ allow us to apply Lemma (b) on p. 20 of Hajek and Sidak (1967) to show that

$$\int \varepsilon^j \frac{p^j(\varepsilon)}{\tilde{p}(\varepsilon^{-j})}\, d\varepsilon^j = -1.$$

The lemma then follows through a straightforward application of the argument in the first column of p. 818 of Ahn and Reinsel (1990). ■

Continuing our proof that $L(W_n(\theta)|P_{\theta,n}) \Rightarrow L(MN(0,S(\theta)))$, we verify the following equation:

$$n^{-3/2}\sum_{t=1}^{n} A'\psi(\varepsilon_t)\psi(\varepsilon_t)' \otimes X_{2,t-1}M'_{t-1} = o_p(1). \tag{A.21}$$

To this end, we rewrite the left-hand side of (A.21) as

$$n^{-3/2}\sum_{t=1}^{n} A'(\psi(\varepsilon_t)\psi(\varepsilon_t)' - \Omega) \otimes X_{2,t-1}M'_{t-1} + n^{-3/2}\sum_{t=1}^{n} A'\Omega \otimes X_{2,t-1}M'_{t-1}$$
$$= o_p(1) + o_p(1).$$

The second term is $o_p(1)$ by (A.11), and, as pointed out by a referee, the first term can be shown to be $o_p(1)$ by a law of large numbers for martingale difference sequences (e.g., White, 1984, p. 58), using the facts that $E[M_tM_t'] < \infty$ and $n^{-1}E[X_{2t}X_{2t}'] < \infty\ \forall t = 1,\dots,n$.

To complete our proof of (8), we consider the asymptotic behavior of the statistic $\gamma' W_n(\theta)$, for any real m-vector $\gamma = (\gamma_1', \gamma_2')'$, where the subvectors γ_1 and γ_2 have respective dimensions of s and $m - s$. It follows from (A.16), (A.17), and (A.21) that

$$L(\gamma' W_n(\theta)|P_{\theta,n}) \Rightarrow L\left(N(0,\gamma_1'[\Omega \otimes E[M_tM_t']]\gamma_1) + MN\left(0,\gamma_2'\begin{bmatrix} A'\Omega A & A'\Omega A \otimes \bar{Z}_2' \\ A'\Omega A \otimes \bar{Z}_2 & A'\Omega A \otimes \int_0^1 Z_2Z_2' \end{bmatrix}\gamma_2\right)\right)$$
$$= L(MN(0,\gamma' S(\theta)\gamma)),$$

from which (8) follows. This completes the proof of Theorem 3.2. ■

Proof of Lemma 4.3. This follows directly from Proposition 3 and Theorem 2 of Jeganathan (1995), using the facts that Θ is open, $\delta_n \to 0$, and δ_n is independent of θ. ■

Proof of Theorem 4.4. We obtain (18) from (12) and (17), while equation (17) follows from Proposition 15 of Jeganathan (1988), Condition A.1 as given earlier, and verification of the following condition (Condition (B.3) in Jeganathan, 1988).

Condition A.4. Verify that, for every $j = 1,\dots,q$, for every bounded $\{h_n\}$, and for every u, it holds that

$$\sum_{t=1}^{n} |u'\delta_n H_{t-1}^j(\theta_n) - u'\delta_n H_{t-1}^j(\theta)|^2 = o_p(1), \tag{A.22}$$

https://doi.org/10.1017/S0266466698141026 Published online by Cambridge University Press

and

$$\max_{t\in\{1,\ldots,n\}} n|\delta_n H^j_{t-1}(\theta)|^2 = O_p(1). \tag{A.23}$$

We begin by checking (A.22), which can be rewritten as

$$u'\delta_n\left(\sum_{t=1}^n H^j_{t-1}(\theta_n)H^j_{t-1}(\theta_n)'\right)\delta_n u - u'\delta_n\left(\sum_{t=1}^n H^j_{t-1}(\theta)H^j_{t-1}(\theta_n)'\right)\delta_n u$$
$$- u'\delta_n\left(\sum_{t=1}^n H^j_{t-1}(\theta_n)H^j_{t-1}(\theta)'\right)\delta_n u - u'\delta_n\left(\sum_{t=1}^n H^j_{t-1}(\theta)H^j_{t-1}(\theta)'\right)\delta_n u = o_p(1), \tag{A.24}$$

which holds because each of the four terms on the left-hand side of (A.24) converges weakly to $u'S^{jj}(\theta)u$, where $S^{jj}(\theta)$ is defined by writing

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

We verify (A.23) by writing its left-hand side as

$$\max_{t\in\{1,\ldots,n\}} n\left|\begin{array}{c} n^{-1/2}\iota^j \otimes M_{t-1} \\ -n^{-1}A(j)' \otimes X_{2,t-1}\end{array}\right|^2, \tag{A.25}$$

where ι^j denotes the jth column of the identity matrix and $A(j)$ denotes the jth row of the matrix A. We can rewrite (A.25) as

$$\begin{aligned}
&\max_{t\in\{1,\ldots,n\}} n[n^{-1}\iota^{j\prime}\iota^j \otimes M'_{t-1}M_{t-1} + n^{-2}A(j)A(j)' \otimes X'_{2,t-1}X_{2,t-1}] \\
&\quad = \max_{t\in\{1,\ldots,n\}} [M'_{t-1}M_{t-1} + n^{-1}(A(j)A(j)')X'_{2,t-1}X_{2,t-1}] \\
&\quad \le \max_{t\in\{1,\ldots,n\}} M'_{t-1}M_{t-1} + \max_{t\in\{1,\ldots,n\}} n^{-1}(A(j)A(j)')X'_{2,t-1}X_{2,t-1} \\
&\quad = O_p(1) + O_p(1) = O_p(1),
\end{aligned}$$

completing our proof of the theorem. ■

Proof of Theorem 4.5. From (11) and (20), it follows that

$$\delta_n^{-1}(\hat\theta_n - \tilde\theta_n) = \hat S_n^{-1} W_n(\theta_n^{**}) - \hat S_n^{-1}(\theta_n^{**})\hat W_n(\theta_n^{**}). \tag{A.26}$$

Using (18), we have

$$\hat W_n(\theta_n^{**}) = \hat W_n(\theta) - S_n(\theta)h_n + o_p(1) \quad \text{in } P_{\theta,n}. \tag{A.27}$$

Combining (A.27) and (15) gives us

$$\hat W_n(\theta_n^{**}) = W_n(\theta) - S_n(\theta)h_n + o_p(1) \quad \text{in } P_{\theta,n}, \tag{A.28}$$

so that the second term on the right-hand side of (A.26) becomes, using (19) and (A.28),

$$-S_n^{-1}(\theta)[W_n(\theta) - S_n(\theta)h_n] + o_p(1) \quad \text{in } P_{\theta,n}. \tag{A.29}$$

By definition,

$$\hat{S}_n = S_n(\theta) + o_p(1) \quad \text{in } P_{\theta,n}, \tag{A.30}$$

while (12) gives

$$\begin{aligned} W_n(\theta_n^{**}) &= W_n(\theta) - \hat{S}_n h_n + o_p(1) \quad \text{in } P_{\theta,n} \\ &= W_n(\theta) - S_n(\theta) h_n + o_p(1) \quad \text{in } P_{\theta,n}, \end{aligned} \tag{A.31}$$

the second equality holding because of (A.30). Combining (A.30) and (A.31) gives

$$\hat{S}_n^{-1} W_n(\theta_n^{**}) = S_n^{-1}(\theta)[W_n(\theta) - S_n(\theta) h_n] + o_p(1) \quad \text{in } P_{\theta,n}. \tag{A.32}$$

Using (A.26), (A.29), and (A.32), the desired result (21) follows. ■

https://doi.org/10.1017/S0266466698141026 Published online by Cambridge University Press