ASYMPTOTIC NORMALITY OF COEFFICIENTS IN A VECTOR AUTOREGRESSION WITH UNIT ROOTS

by

Christopher A. Sims

Discussion Paper No. 229, June, 1986

Center for Economic Research
Department of Economics
University of Minnesota
Minneapolis, Minn 55455

Research for this paper was supported in part by NSF grant SES 8309329.

ASYMPTOTIC NORMALITY OF COEFFICIENTS IN A VECTOR AUTOREGRESSION WITH UNIT ROOTS

by Christopher A. Sims

In a 1978 discussion paper I pointed out that in a second order univariate autoregression with one unit root the coefficients were asymptotically normal, with a singular covariance matrix, when normalized by $T^{.5}$. Earlier Fuller [1976] had obtained the nonsingular asymptotic distribution for a univariate autoregression of arbitrary order with a unit root, from which it was easy to deduce the results of my discussion paper for the case where a scalar $T^{.5}$ normalizing factor replaces the normalizing matrix required to get the nonsingular limiting distribution. (My 1978 paper used a different argument which allowed more general assumptions on the residuals than Fuller's.) The earlier paper proposed that its results extended to vector autoregressions and to the case where roots were on the unit circle but not equal to one. To this point, apparently no one has displayed such extensions.

The extension to complex roots seems quite difficult, but it appears that the case of a vector autoregression (VAR) with roots equal to one is in fact a fairly simple extension of the argument in the 1978 paper. This case has recently come to the fore as the class of "co-integrated" VAR models. (See Engle and Granger [1985], e.g.) have pointed out that such models can be handled with a two step procedure, in which the co-integrating vector is estimated first and used to form a reduced, stationary model. The asymptotic distribution theory for the reduced model is as if the co-integrating vector coefficients were known exactly. This paper shows that such two-step procedures are unnecessary. If the vector autoregression is estimated on the original data, the asymptotic distribution for the coefficients normalized by $T^{.5}$ is again singular normal and identical with that for a model in which the co-integrating vector is known exactly a priori. This result is important, because the two-step procedures have so far been justified only by assuming the number of co-integrating vectors is known a priori. This paper shows that so long as we are not interested in testing hypotheses concerning those linear combinations of coefficients which have degenerate limiting distributions when normalized by $T^{.5}$, we can avoid two-step procedures and ignore the need to guess the number of co-integrating vectors.

The next section considers the case of a first order VAR with non-repeated unit roots and a condition on the disturbance covariance matrix which rules out exact collinearity and deterministic polynomial trends. By the usual expansion of the state vector, this case includes univariate autoregressions of arbitrary order with a single unit root, some bivariate VAR's with two unit roots, some trivariate VAR's with up to three unit roots, etc. So long as a k-variate autoregressive model has no components which need more than first differencing to become stationary, the argument in section 1 below applies. A later paper or a later version of this one will extend these results to the case of repeated roots, which covers VAR's of arbitrary vector length and order of lag with arbitrarily many unit roots and allows for polynomial trend terms as well.

1. Non-repeated Roots

We suppose

1) $$Y(t) = AY(t-1) + u(t) \quad ,$$

with

2) $$E[u(t)|Y(t-s), \text{ all } s>1] = 0$$
$$E[u(t)u(t)'|Y(t-s), \text{ all } s>0] = \Gamma, \text{ and}$$

u stationary with finite fourth moments.

We suppose that the matrix A has Jordan canonical form

$$A = W J W^{-1} ,$$

where J has the eigenvalues of A down its diagonal, in descending absolute value from upper left to lower right. We define the transformed variable vector

$$Z(t) = W^{-1} Y(t) .$$

It is easy to see that (1) implies

3) $Z(t) = J Z(t-1) + v(t)$,

where $v(t)=W^{-1}u(t)$ has zero expectation conditional on all data (Z or Y) up to time t-1 and fixed conditional covariance matrix $W^{-1}\Gamma W^{-1\prime}$. We suppose that there are no roots of A larger than one in absolute value and that all the roots which are of absolute value one are equal to one. Let m be the number of unit roots and n=k-m be the number of other roots. Then the last n rows of W^{-1}, the left eigenvectors of A corresponding to stable roots, are co-integrating vectors for Y. Finally we suppose that for any subvector of the diagonal of J such that all elements are equal, the corresponding submatrix of Ω=Var(v(t)) is nonsingular. (This means that none of the elements of Z are exactly collinear.)[fn]

[fn] In general, A may have complex eigenvalues and eigen vectors, in conjugate pairs, even though all elements of A are real. Since we are assuming all roots on the unit circle are one, none of these are complex, but there could be complex roots less than one in absolute value. Though we take no explicit account of this below, the arguments all go through, with "′"

being interpreted as implying both complex conjugation and transposition of the matrix to which it is applied.

This system can be estimated in two steps, if we know n and we know an n-element subset of the elements of Y such that each has a non-zero coefficient in at least one of the co-integrating vectors. First we estimate a set of regressions of these n variables on the m other variables in the system. The coefficients in these equations will be consistent estimates of a set of vectors which span the same space as the last n rows of W^{-1}. The residuals from these equations will be approximately stationary, and we can estimate a VAR for them as if they did not depend on estimated coefficients. The coefficients of the overall VAR for Y and their asymptotic distribution can be deduced by using the estimated co-integrating vectors and the estimated VAR for the residuals to transform back to the original A matrix. In this process we can ignore the sampling variation in the estimated cointegrating vectors.

We now consider what happens if instead we simply estimate the coefficients of (1) by ordinary least squares. In doing so it is convenient to start by considering OLS estimation of (3), since the estimated coefficients in this set of regressions will just be linear transformations of those for (1) itself. Note that here we are considering OLS estimation of (3) in which the estimated J is unrestricted, so that the diagonality of J is not used in estimation.

The first m components of Z, the nonstationary ones, we will call G, and the remaining n stationary components we will call H. The corresponding decomposition of v is g,h. We partition J conformably as J = [K,L]. Each element of G can be written

$$G_i(t) = \sum_{s=1}^{t} g_i(s) + G_i(0) \quad ,$$

and each element of H as

$$H_i(t) = \sum_{s=0}^{\infty} \tau_i^s h_i(t-s) \quad .$$

We will need the following lemmas.

Lemma 1: $M(T) = (1/T^2) \sum_{t=1}^{T} G(t)G(t)'$

converges in distribution to the distribution of

$$\int_0^1 W(t)W(t)' \, dt \quad ,$$

where W(t) is a vector Wiener process with variance matrix of innovations equal to E[g(t)g(t)')].

Proof: The assumptions we have made imply that $g_i(t)$, $g_j(t)$ is a two-dimensional stationary martingale difference process, which in turn guarantees that

$$S_T(a)/\sqrt{T} = \sum_{t=1}^{[aT]} g(t)/\sqrt{T} = (G(t) - G(0))/\sqrt{T}$$

converges weakly to a vector Wiener process W on (0,1) with the required innovation covariance matrix. (The notation "[x]" is used to mean "the largest integer less than x".) In particular this implies that the integral over (0,1) of $S_T(t)/\sqrt{T}$ and $S_T(t)S_T(t)'/T$ converge in distribution to the same integrals of W(t) and W(t)W(t)'. Since

$$T\,M(T) = \int_0^1 \{S_T(t)S_T(t)' + G(0)S_T(t)' + S_T(t)G(0)'\}dt$$
$$- G(0)G(0)',$$

M(T) converges in distribution to

$$\int_0^1 W(t)W(t)' \, dt \quad .$$

Lemma 2:

$M(T) = T^{-1}\sum^{T} G(t)H(t)'$ is bounded in probability.

$t=1$

Proof: Consider

$$4)\ E[M(T)M(T)'] = T^{-2} \sum_{t=1}^{T} \sum_{s=1}^{t} \sum_{w=1}^{T} \sum_{r=1}^{w} E[g(s)H(t)'H(w)g(r)'] \ .$$

Using the moving average representation of H as

$$H(t) = \sum_{t=1}^{\infty} L^{m}h(t-m)$$

we can write (4) as

$$T^{-2} \sum_{t=1}^{T} \sum_{s=1}^{t} \sum_{m=0}^{\infty} \sum_{w=1}^{T} \sum_{r=1}^{w} \sum_{n=0}^{\infty} E[g(s)L^{m}h(t-m)h(w-n)'L'^{n}g(r)'$$

Every term in this summation is zero, except those in which

i) $t-m=s$ and $w-n=r$ or

ii) $t-m=w-n$ and $s=r$ or

iii) three of the four subscripts are the same and the fourth differs.

This assertion depends on the first two parts of assumption (2). That all terms are finite depends also on the third part of (2). Stationarity guarantees that all these terms are uniformly bounded. For each fixed pair of values of m and n there are fewer than T^2 terms of each of types (i) and (ii), and fewer than 4T of type (iii). Thus the presence of the T^{-2} factor in front of (4) guarantees that for each m, n the sum over the remaining indexes is bounded. But L is a diagonal matrix with diagonal elements all less than one in absolute value, so that when we sum over m and n the result remains bounded. Thus M(T) has bounded second moment and therefore is bounded in probability Q.E.D.

Lemma 3: $M(T) = T^{-1}\sum_{t=1}^{T} G(t)v(t+1)'$ is bounded in probability.

Proof: Follows the proof of Lemma 2, except that the argument is simpler and does not require the bounded fourth moment assumption from (2).

Lemma 4: $M(T) = T^{-1}\sum_{t=1}^{T} H(t)H(t)'$ converges almost surely to a constant matrix, M_H.

Proof: An immediate consequence of ergodicity, the finite variance of H, and our assumption limiting collinearity in Ω.

Now consider $\sqrt{T}(J_T-J)$, where J_T is the ordinary least squares estimator of J in (3). We can write

$$5) \quad \sqrt{T}(K_T-K) = \sqrt{T}\begin{bmatrix} g_T^* \\ h_T^* \end{bmatrix} G_T^{*\prime}(G_T^* G_T^{*\prime})^{-1}$$

$$6) \quad \sqrt{T}(L_T-L) = \sqrt{T}\begin{bmatrix} g_T^* \\ h_T^* \end{bmatrix} H_T^{*\prime}(H_T^{*\prime} H_T^*)^{-1} ,$$

where

$$g_T^* = g_T - g_T H_T'(H_T H_T')^{-1} H_T \text{ , with}$$

$$g_T = [g(2),\ldots,g(T)], \quad H_T = [H(1),\ldots,H(T-1)] ,$$

and

$$h_T^* = h_T - h_T G_T'(G_T G_T')^{-1} G_T \text{ with}$$

$$h_T = [h(2),\ldots,h(T)], \; G_T = [G(1),\ldots,G(T-1)] .$$

G_T^* and H_T^* are defined correspondingly. What we are doing here is simply expressing the multivariate least squares estimators for, say, K, the coefficient of G(t-1), as the univariate least squares estimator in the regression of the sample residuals of Z on the sample residuals of lagged G, after each has been regressed on lagged H. (Alternatively one can think of these expressions as implementations of the usual formulas for partitioned matrix inversion.)

Our next task is to show that (5) converges in probability to zero, while (6) has a limiting normal distribution, so that (5) and (6) jointly have a limiting normal distribution with singular covariance matrix.

$$G_T^* G_T^{*\prime} = G_T G_T' - G_T H_T'(H_T H_T')^{-1} H_T G_T'.$$

Multiplying the above expression through by T^{-2}, we have by Lemma 1 above that the first term on the right-hand side has a limiting distribution which makes it nonsingular with probability one, by our assumption limiting collinearity in Ω. From Lemma 2, we have that $T^{-1}G_T H_T'$ is bounded in probability and from Lemma 4 we have $T^{-1}H_T H_T'$ converging a.s. to a nonsingular constant matrix. This guarantees that the second term on the right-hand side converges to zero in probability when scaled by T^{-2}, so that $T^{-2}G_T^* G_T^{*\prime}$ has the same limiting distribution as $T^{-2}G_T G_T'$. A similar argument shows that the other term in (5), $v_T^* G_T^{*\prime}$, converges to zero in probability when scaled by $T^{-1.5}$, which guarantees (5) itself converges in probability to zero.

Turning now to (6),

$$H_T^* H_T^{*\prime} = H_T H_T' - H_T G_T'(G_T G_T')^{-1} G_T H_T'.$$

Again applying our Lemmas, we can see that the second term on the right converges to zero when scaled by 1/T, so that $(1/T)H_T^* H_T^{*\prime}$ like $(1/T)H_T^* H_T^{*\prime}$ converges in probability to a nonsingular constant matrix. The other term in (6) is

$$v_T^* H_T^{*\prime} = v_T H_T' - v_T G_T'(G_T G_T')^{-1} G_T H_T'.$$

Our Lemmas allow us to conclude that the second term in this expression converges in probability to zero when scaled by $1/\sqrt{T}$. The first term on the right-hand side has a limiting normal distribution when scaled by $1/\surd T$, by the ergodic martingale central limit theorem. (See e.g. Hall and Heyde [1980].) Thus (6) as a whole has a limiting normal distribution. Letting σ_{ij} be the covariance of $v_i(t)$ with $v_j(t)$, the limiting covariance matrix of the coefficients in the i'th and j'th equations, the i_{th} and j^{th} rows of J_T, is given by

$$\begin{bmatrix} 0 & 0 \\ 0 & \sigma_{ij} M_H^{-1} \end{bmatrix} ,$$

where $M_H = E[H(t)H(t)']$. Calling this limiting covariance matrix Ω, we can conclude immediately that the limiting covariance matrix for the least squares estimators of the i'th and j'th rows of A (the i'th and j'th equations of (1)) is

$$\Gamma_{ij}\ W^{-1}{}'\Omega\ W^{-1} ,$$

where Γ_{ij} is the typical element of $\Gamma = E[u(t)u(t)']$. Furthermore, it is easy to check that $W^{-1}{}'\Omega W^{-1}$ is consistently estimated by

$$\left[(1/T) \sum_{t=1}^{T-1} Y(t)Y(t)' \right]^{-1}$$

so that the usual covariance matrix as printed out by an OLS regression program based on the standard normal model is asymptotically justified.

We can summarize these results in a theorem.

Theorem: If A_T is the least squares estimator of A in (1) from a sample of size T and assumptions (2) are satisfied, then $\surd T(A_T - A)$ is asymptotically normally distributed with a singular covariance matrix. The covariance of coefficients in equations i and j is consistently estimated by

$$s_{ij} \, (Y_T Y_T')^{-1},$$

where s_{ij} is a consistent estimator of $\mathrm{cov}(u_i(t),u_j(t))$ and $Y_T = [Y(1),\ldots,Y(T)]$.

2. Implications.

These results show that, in a situation where nonstationarity is a possibility, we can rely on standard estimation and testing procedures to allow properly for that possibility. The common presumption that the usual asymptotic distribution theory does not apply in the presence of nonstationarity is in this sense incorrect. It remains true, of course, that the asymptotic distribution theory we display here is not useful for testing hypotheses about linear combinations of elements of a row of A which have coefficients which lie in the space spanned by the right eigenvectors of A which correspond to unit roots, i.e. the first few columns of W. Since these are also all orthogonal to the last rows of W^{-1}, which are the cointegrating vectors, we can characterize the forbidden linear combinations as those which are orthogonal to all the cointegrating vectors. Thus individual coefficients in the estimated autoregressive equations are asymptotically normal with the usual limiting variance unless they are coefficients of a variable which is nonstationary and does not appear in any of the system's stationary linear combinations.

Thus in a system in which logs of nominal quantities are non-stationary, while real quantities and relative prices are not, the usual asymptotic distribution will apply to all coefficients on individual variables so long as there is more than one nominal variable in the system. We can be sure of this because all differences of logged nominal variables will be stationary, so there cannot be any nominal variables which fail to be part of some stationary linear combination of variables, while of course the real variables are each in themselves a stationary linear combination of variables.

In this example there can be only one unit root, because the claim that all differences of nominal variables and all real variables are stationary leaves room for only one nonstationary linear combination orthogonal to all stationary linear combinations. It also determines the form of that nonstationary linear combination: equal weights on all nominal variables, zero weight on real variables. The usual asymptotic distribution theory fails in such a model just when we try to test an hypothesis about the sum of coefficients of nominal variables. The common sense of it is that the nonstationarity of the level of prices enforces the neutrality condition that the sum of coefficients on nominal variables be zero in the parts of the system which determine relative prices and real quantities. It enforces the condition so firmly that least squares estimates of the sum of coefficients converge to zero faster than $1/\sqrt{T}$, too fast to allow asymptotic normality to emerge.

Classical statistical theory for models like these, asymptotic or not and including that developed in this paper, ought to be regarded as a dubious guide to practice. We can devise test statistics for the null hypothesis of a unit root; but because the asymptotic distribution theory changes character discontinuously we cannot use the coefficient estimates as pivotal statistics to derive confidence regions once we admit the possibility of a unit root. In practice, assessing the evidence about the character of the dynamic model must involve considering both the possibility of unit roots and the possibility of no unit roots, and classical hypothesis testing here is a poor guide.

A simpler and more instructive way to carry out inference in these models is to explore the shape of the likelihood function, using prior notions about what are reasonable dynamics either informally, or formally as a Bayesian prior.[fn] Bayesian analysis with Gaussian disturbances, conditioned on initial values of the process, is of course very convenient, since it leads to Gaussian posteriors. Conditioning on initial values is reasonable when all roots are close to one, since then the process takes a long time to converge to its stationary distribution. Otherwise, parameter values have implications about the marginal distribution of initial observations, which in principle ought to be used. This leads to messy computations even with a Bayesian approach. Nonetheless the Bayesian approach does suggest how in principle evidence about model form ought to blend unit-root distribution theory with stationary-model distribution theory, which a classical approach can do at best clumsily.[fn]

3. Extensions to systems with repeated roots.

[fn] In recent unpublished work J.-M. Dufour of the University of Montreal has explored the use of classical hypothesis testing in autoregressive models, working from exact finite-sample Gaussian distribution theory. His results suggest that the classical approach is more flexible and computationally manageable than one might have guessed.

[fn] Bayesian methods do, in their more convenient forms, condition on a class of distributions for disturbances, but under roughly the same conditions which yield asymptotic normality of estimators independent of the distribution of disturbances, one can show that the likelihood asymptotically takes the shape of a normal p.d.f. (see Hartigan [1983], e.g.). It would be interesting to see whether there is a corresponding limiting result for the shape of the likelihood for these models with unit roots.

There are analogues to Lemmas 1-4 for general VAR systems which include polynomial trends as well as repeated unit roots. The case considered above makes every variable in the transformed system either a random walk or a stationary process. In the general case each variable is a linear combination of random walks integrated up to p times with a q'th order polynomial. Each such term is dominated by the term of highest degree, with the polynomial term dominating when q=p. The estimated coefficients of polynomial-dominated transformed variables are asymptotically normal when properly normalized, while those of random-walk dominated terms are not. This means that the presence of polynomial trend in the system can make the applicability of normal asymptotic theory wider, though assessing which linear combinations of coefficients have the "usual" distribution becomes more complicated. To give a taste of what happens in these cases, consider

$$y(t) = c + by(t-1) + e(t)$$

with the usual stochastic specification and b=1, c nonzero. This model is equivalent to the bivariate singular autoregression

$$y(t) = cz(t-1) + by(t-1) + e(t)$$

$$z(t) = z(t-1),$$

which makes y a linear combination of a linear trend and a random walk, thus dominated by the linear trend. The usual asymptotic distribution theory applies to estimates of b and c even with b=1. There are no linear combinations of b and c which require special treatment. Obviously this is very different from the situation in the same model with c=0, b=1.

REFERENCES

Engle, R.F. and C.W.J. Granger [1985]. "Dynamic Model Specification with Equilibrium Constraints: Co-Integration and Error-Correction." University of California, San Diego discussion paper.

Fuller, W. A. [1976]. Introduction to Statistical Time Series. New York, London, Sydney, Toronto: John Wiley and Sons.

Hall, P. and C.C. Heyde [1980]. Martingale Limit Theory and its Application. New York, London, Toronto, Sydney, San Francisco: Academic Press.

Hartigan, J.A. [1983]. Bayes Theory. New York, Berlin, Heidelberg, Tokyo: Springer-Verlag.