UNIT ROOT NOTES

CHRISTOPHER A. SIMS
PRINCETON UNIVERSITY
SIMS@PRINCETON.EDU

1. UNIT ROOT DISTRIBUTIONS

- To study this topic we need to be able to state and understand a **functional central limit theorem**.
- This in turn requires the notion of a **stochastic process** as a probability on a space of functions.
- We also require the notion of a **metric** on a space of functions and of
- a continuous function sometimes called a "functional" on a space of functions.
- Because we have little time, we will go over these notions roughly. If you have the math background, you can fill in details yourself or look them up. An excellent reference that is newly available (published 2002, but for some reason already in the Princeton bookstore in December 2001) is Pollard (2002). It is written by a statistician who has been teaching the material to economics graduate students.

2. STOCHASTIC PROCESS

- Just as a random variable is a mapping from a probability space $\mathcal{S}$ to the real line $\mathbb{R}$, and a random vector is a mapping $\mathcal{S} \to \mathbb{R}^k$ for finite k, a stochastic process is a mapping $\mathcal{S} \to \mathbb{R}^{\mathbb{R}}$ or $\mathcal{S} \to \mathbb{R}^{\mathbb{Z}}$. And just as a r.v. induces probabilities on sets of real numbers, e.g. intervals, a stochastic process induces probabilities on sets of functions.
- We need to be able to define distance between functions. Examples are
  - the L^p norms on the space of functions $[0,1] \to \mathbb{R}$:

$$\|f-g\|_p = \left(\int_0^1 |f_n(t) - f(t)|^p \, dt \right)^{1/p} .$$

  - the L^∞ or sup norm:

$$\|f-g\|_\infty = \max_{t \in [0,1]} |f(t) - g(t)| .$$

Date: December 30, 2001.

– the Skorohod metric:

$$\rho^0(f,g) \leq \|f-g\|_\infty \,.$$

Of course this is only a property of the Skorohod metric. Its definition is below.

In Euclidean space, a very wide range of reasonable metrics all lead to the same conclusions as to which sequences converge.

$$\sum_{j=1}^{k} \left|x_j(n) - x_j\right| \xrightarrow[n\to\infty]{} 0 \Leftrightarrow \sqrt{\sum_{j=1}^{k}(x_j(n)-x_j)^2} \xrightarrow[n\to\infty]{} 0\,,$$

for example. This is no longer true in spaces of functions or infinite sequences. If $k=\infty$ in the formulas above, the equivalence asserted no longer holds. (E.g. if

$$x_j(n) = \begin{cases} 1/n, & j = 1,\ldots,n \\ 0 & j<0,\, j>n\,, \end{cases}$$

then the sequence does not converge to $x_j \equiv 0$ in the first metric above (called ℓ_1) but does converge to it in the second (called ℓ_2). We are not going to be using technical convergence arguments, but you need to know why the results we quote always specify a metric instead of (as with Euclidean space arguments) just a space.

3. The Wiener Process W

(It's also called a Brownian Motion.)

- It's a stochastic process on the space of continuous functions $[0,1] \to \mathbb{R}$.
- $W(0)=0$; for any $\{t_1,\ldots,t_n\}$, $\{W(t_1),\ldots,W(t_n)\}$ is jointly normal; $\mathrm{Cov}(W(t)-W(s), W(u)-W(v))$, with $t>s$, $u>v$, is just the length of the interval $[s,t]\cap[v,u]$.
- The time paths of W are with probability one continuous and nowhere differentiable.

4. Convergence in Distribution Revisited

For Euclidean-space random variables, convergence in distribution is often defined in elementary courses as "pointwise convergence of the distribution function at all points of continuity". There is no useful generalization of this definition to infinite-dimensional random variables. A definition that is equivalent in Euclidean space (and sometimes used, or at least noted as equivalent, even in elementary courses) is, if X_n takes values in

$\mathcal{S}$ which is endowed with the metric ρ,

$$X_n \xrightarrow[n\to\infty]{\mathcal{D},\rho} X \quad \Leftrightarrow$$
$$\text{for every bounded, } \rho\text{-continuous } f : \mathcal{S} \to \mathbb{R},$$
$$E[f(X_n)] \xrightarrow[n\to\infty]{} E[f(X)]\,.$$

When defined this way, convergence in distribution is often called "weak convergence". Usually the ρ above the convergence arrow is left implicit.

5. The Functional Central Limit Theorem

Theorem 1. *Suppose $\{U_t\}$ form a stationary process with finite-variance martingale increments (i.e., $E_t[U_{t+1}] = 0$). Then if*

$$W_T(r) = \frac{1}{\sqrt{T}} \sum_{t=1}^{[rT]} U_t\,,$$

$$W_T \xrightarrow{\mathcal{D},\rho^0} W\,.$$

The notation $[rT]$ here refers to the smallest integer greater than or equal to rT.

Theorem 2 (Continuous Mapping)**.** *If $f : \Omega \to \Gamma$ is continuous, if the stochastic processes X_n take values in the space of functions Ω, and if $X_n \xrightarrow{\mathcal{D}} X_\infty$ in Ω, then $f(X_n) \xrightarrow{\mathcal{D}} f(X_\infty)$ in Γ.*

6. The Skorhod Metric

While for the purposes of this course you do not need to know it, you may find it comforting to see the explicit definition of the Skorohod metric. The basic idea is to define a metric that agrees with L_∞ for continuous functions, but that has $\rho^0(f_n, f) \to 0$ when $f(t) = \mathbf{1}(t \geq .5)$ and $f_n(t) = \mathbf{1}(t \geq .5 - n^{-1})$. That is, functions that are close to each other except for discontinuities at slightly different values of t are to be treated as near each other.

Let Λ be the class of time-distortion functions, i.e. monotone increasing functions $\lambda : [0,1] \to [0,1]$ with $\lambda(0) = 0$, $\lambda(1) = 1$. The degree of distortion impied by λ is

$$\delta(\lambda) = \max_{t,s\in[0,1],t\neq s} \log\left(\frac{\lambda(t) - \lambda(s)}{t - s}\right)$$

Then

$$\rho^0(f,g) = \inf_{\lambda\in\Lambda} \max\left\{\|f(\cdot) - g(\lambda(\cdot))\|_\infty\,,\ \delta(\lambda)\right\}$$

This metric is clearly "weaker" than L_∞, which means that it will accept as convergent some sequences of functions that L_∞ does not accept as convergent. This is guaranteed by the fact that $\|f - g\|_\infty \geq \rho^0(f,g)$. (This inequality should be close to obvious. Be sure you see why.)

7. THE SIMPLEST UNIT ROOT AR

- $y_t = \rho y_{t-1} + \varepsilon_t$
- $$\hat{\rho} = \rho + \frac{\sum_1^T y_{t-1}\varepsilon_t}{\sum_1^T y_{t-1}^2} .$$
- If $\rho = 1$,
$$\begin{bmatrix} \frac{1}{T^2}\sum_1^T y_{t-1}^2 \\ \frac{1}{T^2}\sum_1^T \varepsilon_t y_{t-1} \end{bmatrix} \xrightarrow{\mathcal{D}} \begin{bmatrix} \int_0^1 W_t^2\, dt \\ W(1)^2 - 1 \end{bmatrix} .$$
- Therefore $T(\hat{\rho}_{OLS} - \rho)$ has a limiting distribution that is non-normal.

Note that of course some more regularity conditions are needed here: for example that the ε sequence is stationary, finite-variance, and has zero expectation conditional on all past y's.

8. IMPLICATIONS

- Discontinuous confidence regions based on asymptotics
- Power/significance level relationship is different from pure location case
- Helicopter tour (Sims and Uhlig, 1991)

9. SPURIOUS REGRESSION

(i) It is not unit-root non-stationarity alone that generates the non-standard behavior. If the model is instead
$$\begin{aligned} y_t &= \alpha x_t + \varepsilon_t \\ x_t &= x_{t-1} + \nu_t \\ \mathrm{Cov}(\varepsilon_t, \nu_s) &= 0\,, \text{ all } t, s\,, \end{aligned}$$
with ε_t still stationary, finite-variance, martingale differences and ν the stationary innovation sequence for x, then the usual OLS Normal asymptotics for the OLS estimate of α are correct (though the sum of squares of the rhs variable grows faster than T, so the *speed* of convergence is faster than usual).

(ii) There is a widely known result, labeled "spurious regression" that might seem to contradict (i). It states that when
$$\begin{aligned} y_t &= y_{t-1} + \varepsilon_t \\ x_t &= x_{t-1} + \nu_t \\ \mathrm{Cov}(\varepsilon_t, \nu_s) &= 0, \text{ all } t, s\,, \end{aligned}$$
then if we try to estimate as a regression equation
$$y_t = \alpha x_t + \xi_t\,,$$
we will, by applying the usual OLS test statistics, get incorrect results.

(iii) The reconciliation is that in (ii) we are attempting to estimate a mis-specified model. There is no choice of α that can make that equation's residual a martingale difference. The lesson is therefore not that regression equations involving non-stationary variables always require non-standard distribution theory, but that special care is required to be sure that residuals have the usual stationarity and serial uncorrelatedness properties and that the exogeneity assumptions that justify standard distribution theory are satisfied.

10. MULTIVARIATE CASES: READ (SIMS, STOCK, AND WATSON, 1990)

General multivariate time series regression:

$$(*) \qquad y(t) = Bx(t) + \varepsilon(t)\,.$$

y and x are both components of a longer vector w, and with

$$E[\varepsilon(t) \mid \{w(s-1),\, s \le t\}] = 0, \quad \mathrm{Var}(\varepsilon_t) = \sigma\,.$$

$$w(t) = Aw(t-1) + \nu(t)$$

with $\nu(t)$ the innovation in w and $\mathrm{Var}(\nu(t))$ possibly singular. This allows for the possibility that some elements of w are in fact pure deterministic polynomial trends. However, we assume that all elements of the w vector are distinct — that is, that no linear combination of w's is identically zero.

$$\nu(t) = \gamma\varepsilon(t) + \xi(t)\,,$$

with $\xi(t) \perp \varepsilon(t)$.

11. USE JORDAN DECOMPOSITION

$$A = P\Lambda P^{-1}\,, \quad z(t) = P^{-1}w(t)$$

to get

$$z(t) = \Lambda z(t-1) + P^{-1}c + P^{-1}\nu_t\,,$$

with Λ a Jordan matrix, meaning it has the eigenvalues of A down the main diagonal, and is block diagonal with all diagonal blocks in the form

$$\begin{bmatrix} \lambda & 1 & 0 & 0 & \dots & 0 \\ 0 & \lambda & 1 & 0 & \dots & 0 \\ 0 & 0 & \lambda & 1 & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 & \lambda & 1 \\ 0 & \dots & \dots & \dots & 0 & \lambda \end{bmatrix}$$

Each element z_i of the z vector is of one four types, for our purposes:

(i) It corresponds to an eigenvalue of A (diagonal element of Λ) that is less than one in absolute value, and thus is stationary (at least in the long run).

(ii) It is a component of a subvector of z that corresponds to a unit-root Jordan block for which the component of $P^{-1}\nu(t)$ corresponding to the lower right corner of the Jordan block is identically zero. In this case we say z_i is **dominated by polynomial trend**. Note that, because of our assumption that no linear combination of of w's (and hence of z's) is identically zero, there can be at most one unit root Jordan block of this kind.

(iii) It is a component of a subvector of z that corresponds to a unit-root Jordan block in which all elements of $P^{-1}\gamma$ corresponding to z_i or to lower elements of the same Jordan block are zero and the element of $P^{-1}\nu$ corresponding to the lower right corner of the block is non-zero. In this case we say z_i is **dominated by exogenous stochastic trend**.

(iv) It is a component of a subvector of z that corresponds to a unit-root Jordan block and does not fit any of the previous categories. In this case we say z_i is **dominated by endogenous stochastic trend**.

z_i's in these categories can be ordered as follows. Any z_i in the p'th postion from the bottom of a unit root Jordan block dominates any other z_j that is stationary or in a lower position in any unit root Jordan block. We will say such a z is at **position p**. If z_i and z_j are at the same position p, then the blocks that are of higher-numbered categories in the scheme above dominate lower-numbered categories. It is useful to extend this partial ordering on the z_i's by treating any linear combination of z_i's as having the dominance characteristics of the "strongest" of its components.

Now $x(t) = \theta z(t)$ for some θ. Suppose we order the z vector so that the highest ordered components are at the top, lowest ordered at the bottom. Then we form the QR decomposition $QR = \theta$ of θ, where Q is orthonormal and R has its lower left triangle zero. Define $B^* = BQ$ and $x^* = Rz$. Then the elements of x^* are ordered from bottom to top of the vector, with upper elements dominating lower.

12. Implications of Sims, Stock, and Watson (1990)

OLS estimates of B^* have these properties:

(i) Estimated coefficients of stationary components of $x^*(t)$ have standard normal limiting distributions given by the usual OLS formulas, and converge at the rate $1/\sqrt{T}$.

(ii) If there are no components of x^* dominated by endogenous stochastic trend, then all estimated coefficients have standard normal limiting distributions, conditional on the right-hand-side variable matrix.

(iii) If there are any components of x^* that are dominated by endogenous stochastic trend, then their coefficients and also coefficients of all components of x^* other than the stationary components have non-standard limiting distributions and converge at faster than $1/\sqrt{T}$.

Since we are usually interested in B itself, not B^*, we have to translate these results back into results about OLS estimates $\hat{B} = \hat{B}^* Q$. Since convergence of estimated coefficients of

stationary variables in $\hat{B}^*$ is slower than convergence of any of the other categories. Any element of or linear combination of $\hat{B} = \hat{B}^* Q$ that puts non-zero weight on the components of $\hat{B}^*$ corresponding to stationary x^*'s itself has a standard limiting normal distribution, convergent at rate $1/\sqrt{T}$.

13. THE EYEBALL METHOD

The algorithm just described can be thought of as transforming the right-hand-side variable vector $x(t)$ into a new vector $z(t)$ in which there as many as possible stationary variables, then as many as possible position 1, category (i) variables, then given that as many as possible position 1, (ii) variables, etc., then repeating for position 2, etc. The transformation must be nonsingular, and it must keep all the variables dominated by deterministic trend distinct, in the sense that there can be only one variable dominated by p'th order deterministic trend for each p. In small models it is often possible to construct such a transformation "by eye", in which case the numerical apparatus of the Jordan and QR decompositions is not necessary. All that is needed is the transformation matrix that plays the role of Q above, and its inverse, so we can write $\hat{B} = \hat{B}^* Q$, where $\hat{B}^*$ is the vector of coefficients on z.

14. EXAMPLES:

$$y(t) = ay(t-1) - by(t-2) + \varepsilon(t)$$

(i) $a = 2$, $b = 1$
(ii) $a = 1.7$, $b = .7$

$$y(t) = c + ay(t-1) + \varepsilon(t)$$

(iii) $c = .02$, $a = 1$
(iv) $c = 0$, $a = 1$

In cases (i) and (iv) all individual coefficients have non-standard distributions. In cases (ii) and (iii) none do. In case (ii) the sum of the two coefficients has a non-standard distribution, even though each individually does not. In case (iii) the whole joint distribution (and thus also any linear combination of coefficient estimates) has a standard limiting distribution.

Here are the details. We use the conventional terminology, in which a position-p variable dominated by stochastic trend is called an $I(p)$ variable:

(i) Here the eyeball method tells us that $y(t)$ itself is $I(2)$, meaning that the lowest category and order we can obtain by taking linear combinations of the right-hand-side variables $y(t-1)$ and $y(t-2)$ is $\Delta y(t-1)$, which being $I(1)$ is order 1, dominated by endogenous stochastic trend. So every linear combination of the two right-hand-side variables is dominated by stochastic trend and all coefficients

and linear combinations of them have non-standard distributions.

$$z(t) = w(t) = \begin{bmatrix} y(t) \\ y(t) - y(t-1) \end{bmatrix}, \quad z(t) = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} z(t-1) + \begin{bmatrix} \varepsilon(t) \\ \varepsilon(t) \end{bmatrix}.$$

Thus A is already in the form of a Jordan block with no constant term in the lowest equation of the block. Both elements of x are thus in category (iv) and the joint distribution of OLS coefficients is thus nonstandard.

(ii) In this system the characteristic roots are 1 and .7. The stationary linear combination is $\Delta y(t-1)$. Thus we can write

$$Bx(t) = B \begin{bmatrix} y(t-1) \\ y(t-2) \end{bmatrix} = B \begin{bmatrix} 1 & 0 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} y(t-1) \\ \Delta y(t-1) \end{bmatrix}.$$

From this we can see that the sum of the two elements of B corresponds to the coefficient of a variable dominated by stochastic trend, while any other linear combination of coefficients will have some weight on the coefficient of a stationary variable and will therefore have a standard limiting distribution.

(iii) In case (iii) and (iv) we can take $x(t) = w(t) = \begin{bmatrix} y(t) & \gamma_t \end{bmatrix}'$, where γ_t is constant. This gives us

$$A = \begin{bmatrix} a & c \\ 0 & 1 \end{bmatrix}.$$

In both cases (iii) and (iv) $a = 1$, so there are two eigenvalues of 1. In case (iii), with $c = .02$, the two unit roots are in a single Jordan block, corresponding to the fact that y consists of a linear trend plus a random walk and is thus dominated by its linear trend component. In case (iv), with $c = 0$, there is no linear trend component to y, so it is dominated by endogenous stochastic trend, and the two coefficients have a non-standard limiting joint distribution.

15. Implications for Practice

It is rare for the non-standard classical distribution theory for multivariate AR's to be applied in practice to systems of any substantial size. The asymptotic distribution theory depends on the number and types of unit roots and cointegrating vectors, and generally we do not know these in advance. Furthermore, producing confidence regions requires considering the distribution theory for regions in parameter space that usually include points for which the numbers of unit roots and cointegrating vectors are different.

So the most important practical implication of these results is that for many purposes the non-standard theory is not necessary, even from a classical perspective. However, some of the results in this direction need to be used with caution. Most VAR systems are estimated with constant terms included, and the constant terms are not expected to be zero. Economic variables often are well modeled as including both linear trend terms and pure $I(1)$ stochastic components. Thus it might be expected that we would often find systems in which all variables are dominated by deterministic trend. But recall that not more than one z-variable in a system can be dominated by linear trend. Furthermore,

even in univariate models, the component of variation explained by linear trend is usually modeled as not much different in size from the component of variation explained by the $I(1)$ component. The asymptotic theory that says the $I(1)$ component can be ignored relies on the fact that for very large sample sizes, the component of variation attributable to the $I(1)$ component must be negligible relative to that explained by the trend component. And in fact it is true that the standard Gaussian distribution theory for OLS is a bad approximation, despite its asymptotic validity, in models like example (iii), when T, c, and $\text{Var}(\varepsilon(t))$ have values typical for economic data.

The difficulties surrounding inference in VAR models with unit roots are completely different, and arguably much more manageable, from a Bayesian perspective. The likelihood with Gaussian disturbance terms, conditional on initial conditions, is Gaussian regardless of the presence or absence of unit roots. Non-normal disturbance terms still lead to asymptotically Gaussian likelihood shape, even in the presence of unit roots (Kim, 1994). This is not to say there are no difficulties for Bayesian inference associated with unit roots. As we discussed earlier, a flat prior on the coefficients of an AR with a constant term, together with conditioning on initial conditions, implies what will in most applications be an inappropriately large prior weight on parameter values that imply a large part of sample variation could have been predicted at the start of the sample. This problem will affect inference strongly mainly when unit roots (or explosive non-stationarity) are present.

Non-Bayesian asymptotics implies a maze of technical econometric issues must be addressed in order to test hypotheses or construct confidence regions when unit roots may be present. Bayesian inference also implies that there are issues special to the presence of unit (or explosive) roots. But the Bayesian difficulties are of a different nature. They suggest that there is a special need, when non-stationarity may be present, to think carefully about the substance of the problem. Do we believe that the initial conditions should be treated as if generated from a long run of the estimated model? Do we believe that low frequency oscillations in the model's variables are possibly predictable long in advance? How we report the shape of the likelihood depends on our answers to these questions.

REFERENCES

KIM, J. (1994): "Bayesian Asymptotic Theory in a Times Series Model with a Possible Nonstationary Process," *Econometric Theory*, 10(3), 764–773.

POLLARD, D. (2002): *A User's Guide to Measure Theoretic Probability*, Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

SIMS, C. A., J. STOCK, AND M. WATSON (1990): "Inference in Linear Time Series Models with Some Unit Roots," *Econometrica*, 58, 113–144.

SIMS, C. A., AND H. D. UHLIG (1991): "Understanding Unit Rooters: A Helicopter Tour," *Econometrica*, 59.