Introduction to time series econometrics and VARs

Tom Holden

http://www.tholden.org/

PhD Macroeconomics, Semester 2

Outline of today's talk

- Discussion of the structure of the course.

- Some basics:
  - Vector spaces
  - Linear maps and the lag operator

- Concepts in time series econometrics.
  - Autocovariance, autocorrelation and partial-autocorrelation
  - Stationarity
  - AR and MA processes, and the Wold representation theorem.
  - VARs.
  - Cointegration and the consistency of VAR estimates under non-stationarity.
  - Bayesian VARs.

Structure of the course

- Broadly in two halves.

- First half: Empirical macroeconomics, including:
  - Time series econometrics, VARs, SVARs and identification.
  - Continuous time stochastic processes and the frequency domain.
  - Structural estimation of DSGE models, including filtering (linear and otherwise), and GMM/SMM.

- Second half: Topics in macroeconomics, including:
  - Discrete time optimal control.
  - Continuous time optimal control and models of growth.
  - Search and matching models.
  - Rational expectations and learning.

Reading for today

- Canova: “Methods for applied macroeconomic research”
  - Chapter 1 covers the basics.
  - Chapter 4 covers VARs.
  - Chapter 10 covers BVARs.

- Wikipedia, as needed for basic results in linear algebra.
  - Reading all of the pages in this category would be a good start: https://en.wikipedia.org/wiki/Category:Matrix_decompositions

- Hendry: “Dynamic econometrics” may also be useful for the cointegration material.
  - Worth buying just for the appendices.

Vector Spaces

- Let $\mathbb{F}$ be a field, such as $\mathbb{R}$ or $\mathbb{C}$.
  - A field is a set endowed with two operations "+" and "×", where both operations have identity elements and inverses.
- Let V be some set endowed with two operations, "+" and "×" where $+: V \times V \to V$ and $\times: \mathbb{F} \times V \to V$.
  - So you can add two elements of V, but we only define multiplication between an element of V and an element of $\mathbb{F}$.
  - As is standard, we will omit multiplication signs in practice.

- If "+" is associative ($v + (w + x) = (v + w) + x$), commutative ($v + w = w + v$), possesses an identity element, denoted $0 \in V$, ($v + 0 = v$), possesses an inverse ($v + (-v) = 0$), if "×" is compatible with multiplication in the field ($(\alpha\beta)v = \alpha(\beta v)$) and shares an identity element with the field, denoted $1 \in \mathbb{F}$ ($1v = v$), and if the distributive laws are satisfied ($\alpha(v + w) = \alpha v + \alpha w$, $(\alpha + \beta)v = \alpha v + \beta v$), then we call V a vector space.

More on Vector Spaces

- Canonical example of a vector space is $\mathbb{R}^n$, the space of n dimensional vectors.
  - May be thought of as the set: $\{x|x: \{1, \dots, n\} \to \mathbb{R}\}$ of functions mapping $\{1, \dots, n\}$ to $\mathbb{R}$.
  - Addition is defined componentwise: $(x + y)(i) = x(i) + y(i)$.
  - Likewise, multiplication: $(\alpha x)(i) = \alpha x(i)$.
  - Note that this means $\mathbb{R}^1 = \mathbb{R}$ is a vector space too (in fact all fields are vector spaces).

- The previous construction generalises.
  - In particular, let I be some set, and let V be a vector space, then the set $V^I \coloneqq \{x|x: I \to V\}$ is a vector space too, where addition and multiplication are again defined componentwise.
  - In time series econometrics, we are interested in infinite dimensional vector spaces which are indexed by "time": i.e. the spaces $(\mathbb{R}^n)^{\mathbb{N}}$ and $(\mathbb{R}^n)^{\mathbb{Z}}$.

The lag operator, and other linear maps

- Let V and W be vector spaces, and let $T: V \to W$, where:
  - $T(u+v) = Tu + Tv$ for all $u, v \in V$, and:
  - $T(\alpha v) = \alpha T v$
- Then we call T a linear map.
  - When $V = W$ we call T a linear operator.

- If $V = \mathbb{R}^n$ and $W = \mathbb{R}^m$, then T may be represented as a matrix.
  - In general, you may think of linear maps as being a generalisation of matrices.
- Just as matrices may be multiplied, so too linear maps may be "multiplied" through functional composition.
  - For linear maps S and T, $(ST)(v) = S(Tv)$.

- In time series econometrics, one linear map in particular will be very useful, namely the lag operator, $L: (\mathbb{R}^n)^{\mathbb{Z}} \to (\mathbb{R}^n)^{\mathbb{Z}}$, defined by:
  - $L(x)(i) = x(i-1)$ for all $x \in (\mathbb{R}^n)^{\mathbb{Z}}$ and $i \in \mathbb{Z}$.
  - In the standard notation of time series econometrics, this is often written as $Lx_t = x_{t-1}$ for all t.

Stochastic processes

- Stochastic processes are just random variables drawn from a vector space that is indexed by time.

- Important to keep in mind the distinction between the random variable (which is strictly a function of the state of nature), and the realisation of that random variable.
  - Some use capital letters for random variables and lower case for realisations e.g. $\Pr(X_t \leq x_t)$ is the probability that the random variable X_t was lower than some particular value x_t.
  - Others make the dependence on the state of nature explicit by writing e.g. $x_t(\chi)$ for the random variable, where χ is the state of nature.
  - Particularly in time series econometrics though, these conventions are rarely followed exclusively, so often x_t may denote a random variable in one place, and a realisation in another. Be careful!

Autocovariance

- The autocovariance function of a possibly vector valued stochastic process x_t is given by:

$$\mathrm{ACF}_t(\tau) = \mathbb{E}[(x_t - \mathbb{E}x_t)(x_{t-\tau} - \mathbb{E}x_{t-\tau})']$$

- I.e. its value at τ is the covariance between x_t and $x_{t-\tau}$.

- There is no reason in general why $\mathrm{ACF}_t(\tau)$ should equal $\mathrm{ACF}_s(\tau)$ if $t \neq s$.

- However, by definition: $\mathrm{ACF}_t(\tau) = \mathrm{ACF}_{t-\tau}(-\tau)$.

Autocorrelation and partial autocorrelation

- The autocorrelation of a scalar valued stochastic process x_t is defined as you would expect:

$$\mathrm{ACRF}_t(\tau) = \frac{\mathrm{ACF}_t(\tau)}{\sqrt{\mathrm{var}\, x_t \,\mathrm{var}\, x_{t-\tau}}}$$

- I.e. it is the correlation between x_t and $x_{t-\tau}$.

- The partial autocorrelation $\mathrm{PACRF}_t(\tau)$ is the correlation between $x_t - \hat{x}_{t|t-\tau+1,t-\tau+2,\ldots,t-1}$ and $x_{t-\tau} - \hat{x}_{t-\tau|t-\tau+1,t-\tau+2,\ldots,t-1}$, where $\hat{x}_{t|t-\tau+1,t-\tau+2,\ldots,t-1}$ and $\hat{x}_{t-\tau|t-\tau+1,t-\tau+2,\ldots,t-1}$ are the best linear predictors of x_t and $x_{t-\tau}$ given $x_{t-\tau+1}$, $x_{t-\tau+2}$, …, x_{t-1} (in the sense of minimal mean squared error).

Stationarity: definitions

- A stochastic process is called stationary if for any finite index set $I \subseteq \mathbb{Z}$, and any $\tau \in \mathbb{Z}$, the joint distribution of $\{x_t | t \in I\}$ is the same as that of $\{x_{t-\tau} | t \in I\}$.
  - I.e. shifting the process backwards or forwards will not change its distribution.
  - Consequently, for any stationary process, and any $t, s, \tau \in \mathbb{Z}$, $\mathbb{E}x_t^\tau = \mathbb{E}x_s^\tau$, $\mathrm{ACF}_t(\tau) = \mathrm{ACF}_s(\tau)$, $\mathrm{ACRF}_t(\tau) = \mathrm{ACRF}_s(\tau)$ and $\mathrm{PACRF}_t(\tau) = \mathrm{PACRF}_s(\tau)$.
  - As a result, for stationary processes we usually drop the t index to the A.C.F. and just write e.g. $\mathrm{ACF}(\tau)$.

- A process is called weakly or weak-sense or covariance stationary if for all $t, s, \tau \in \mathbb{Z}$, $\mathbb{E}x_t = \mathbb{E}x_s < \infty$ and $\mathrm{ACF}_t(\tau) = \mathrm{ACF}_s(\tau) < \infty$ (so $\mathbb{E}x_t^2 = \mathbb{E}x_s^2 < \infty$ as well).

Stationarity: relationship between the definitions

- Most (but not all) stationary processes are weakly stationary.
  - Can you spot which stationary processes are not weakly stationary?
  - Note that Canova repeats a common mistake and asserts that all stationary processes are weakly stationary, which is not correct.

- Not all weakly stationary processes are stationary.
  - Example in exercises.

Sample analogues of A.C.F. etc.

- For weakly stationary processes we can estimate the ACF etc.
  - Why can't we do this for non-stationary ones?

- Sample analogues are defined in the obvious way:

$$\widehat{\mathrm{ACF}}(\tau) = \frac{1}{T}\sum_{t=\tau+1}^{T}(x_t - \hat{\mu})(x_{t-\tau} - \hat{\mu})'$$

  - where $\hat{\mu}$ is the process's sample mean.

  - An alternative estimator uses $\frac{1}{T-\tau}$ in front instead, which is unbiased when the mean is known, but has higher MSE generally.

- For the partial auto correlation function the best linear predictors are just estimated by regression.

Moving average processes

- Let ε_t be i.i.d. with mean 0 and variance σ^2.
  - In future we will write $\varepsilon_t \sim \mathrm{WNIID}(0, \sigma^2)$.
- An $\mathrm{MA}(q)$ process is of the form:

$$\begin{aligned} x_t &= \mu + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q} \\ &= \mu + \left(I + \theta_1 L + \theta_2 L^2 + \cdots + \theta_q L^q\right)\varepsilon_t \\ &= \mu + \Theta_q(L)\varepsilon_t \end{aligned}$$

  - where $\Theta_q(z) = 1 + \theta_1 z + \theta_2 z^2 + \cdots + \theta_q z^q$.
- The process is called invertible if x_t may be written as a linear combination of $x_{t-1}, x_{t-2}, \ldots$ and ε_t, with the coefficients on the lags being absolutely summable.
- Let $\rho_{\Theta,1}, \ldots, \rho_{\Theta,q}$ be the q (possibly complex) roots of the polynomial Θ_q.
  - Then $x_t - \mu = \left[\prod_{k=1}^{q}\left(I - \rho_{\Theta,k}^{-1} L\right)\right]\varepsilon_t$.
  - Note that $\frac{1}{1-\rho^{-1}z} = 1 + \rho^{-1}z + \rho^{-2}z^2 + \cdots$ where the RHS is only well defined if $|\rho| > 1$.
  - Consequently (or at least, by this intuition), x_t is only invertible if $\left|\rho_{\Theta,k}\right| > 1$ for all k.

Autoregressive processes

- Let $\varepsilon_t \sim \mathrm{WNIID}(0, \sigma^2)$.
- An $\mathrm{AR}(p)$ process is of the form:
$$x_t = \mu + \phi_1 x_{t-1} + \phi_2 x_{t-2} + \cdots + \phi_p x_{t-p} + \varepsilon_t$$
  - Hence:
$$\left(I - \phi_1 L - \phi_2 L^2 - \cdots - \phi_p L^p\right) x_t = \Phi_p(L) x_t = \mu + \varepsilon_t$$
  - where $\Phi_p(z) = 1 - \phi_1 z - \phi_2 z^2 - \cdots - \phi_p z^p$.
- The process is called causal ("stable" in Canova) if it may be written as a linear combination of ε_t, ε_{t-1}, …, with the coefficients being absolutely summable.
  - Clearly any causal process is also stationary.

- Much as before, let $\rho_{\Phi,1}$, …, $\rho_{\Phi,p}$ be the roots of Φ_p.
  - Then $\left[\prod_{k=1}^{p}\left(I - \rho_{\Phi,k}^{-1} L\right)\right] x_t = \mu + \varepsilon_t$, so as long as $\left|\rho_{\Phi,k}\right| > 1$ for all k, the process will be causal.

ARMA processes

- Are what you expect them to be…
- An $\mathrm{ARMA}(p, q)$ has the form:

$$\Phi_p(L)x_t = \mu + \Theta_q(L)\varepsilon_t$$

- One nice motivation for using ARMA processes is that any linear combination of finite ARMA processes has a finite ARMA representation.
  - In general this will have both an AR and an MA part, unless all of the original processes were MA.

- If the roots of Φ_p are all outside the unit circle, then the process is stationary and admits an infinite MA representation, and if the roots of Θ_q are outside the unit circle it admits an infinite AR representation.

Wold's representation theorem

- Let x_t be an arbitrary weakly stationary stochastic process.

- Then Wold's representation theorem states that there exists a deterministic process η_t, an independently distributed white noise process ε_t and $\theta_0, \theta_1, \ldots \in \mathbb{R}$ such that:

$$x_t = \eta_t + \sum_{k=0}^{\infty} \theta_k \, \varepsilon_{t-k}$$

- I.e. all weakly stationary processes admit a representation as a sum of a deterministic process and an MA one.

- Since ARMA processes provide efficient representations of MA ones, this is further justification for using ARMA processes.

Lag selection

- ACF and PACF are invaluable for lag selection, as:
  - The PACF of an AR process has p non-zero entries.
  - The ACF of an MA process has q non-zero entries.

- See examples…

Spurious regression

- When x_t and y_t are non-stationary stochastic processes, regressing x_t on y_t (or vice versa) will result in spurious correlation (the estimated β is inconsistent) and high "R^2" values.
  - Generally high "R^2" values are a bad thing, not a good thing!

- Important to check stationarity first. (ADF, KPSS etc.)
  - ADF effectively tests the null that $\phi = 1$ in $x_t = \phi x_{t-1} + \text{other lags} + \varepsilon_t$.
  - KPSS tests the null that $\sigma = 0$ in $x_t = z_t + y_t$ where y_t is stationary and $z_t = z_{t-1} + \sigma\varepsilon_t$.

Vector auto-regressions

- Vector versions of AR, MA and ARMA processes are straightforward to define.
  - If $x_t \in \mathbb{R}^n$ for all t, then rather than being scalars, $\phi_1, \dots, \phi_p$ and $\theta_1, \dots, \theta_p$ are now matrices.
- E.g.: A $\mathrm{VAR}(p)$ has a representation:

$$x_t = \mu + \phi_1 x_{t-1} + \phi_2 x_{t-2} + \cdots + \phi_p x_{t-p} + \varepsilon_t$$

  - where $\varepsilon_t \sim \mathrm{WNIID}(0, \Sigma_\varepsilon)$, $x_t, \mu \in \mathbb{R}^n$,$\Sigma, \phi_1, \dots, \phi_p \in \mathbb{R}^{n \times n}$.

- The conditions for being causal and invertible are now conditions on the roots of $\det \Phi_p(z)$ and $\det \Theta_q(z)$ respectively (where as before the roots need to be outside of the unit circle).
  - Basically then, everything is the same.

Moments of a stationary $\mathrm{VAR}(1)$

- Taking unconditional expectations gives: $\mathbb{E}x_t - \phi_1 \mathbb{E}x_{t-1} = \mu$.
  - If the process is stationary, $\mathbb{E}x_t = \mathbb{E}x_{t-1}$, hence $(I - \phi_1)\mathbb{E}x_t = \mu$, i.e. $\mathbb{E}x_t = (I - \phi_1)^{-1}\mu$.
    - Note that the inverse on the RHS exists for stationary processes.

- Taking variances gives: $\operatorname{var} x_t = \phi_1 \operatorname{var} x_{t-1} \phi_1' + \Sigma_\varepsilon$, since x_{t-1} and ε_t are independent.
  - Under stationarity this simplifies to the Lyapuvnov equation $\operatorname{var} x_t = \phi_1 \operatorname{var} x_t \phi_1' + \Sigma_\varepsilon$.
  - Applying the vec operator to both sides then gives: $\operatorname{vec} \operatorname{var} x_t = (\phi_1 \otimes \phi_1) \operatorname{vec} \operatorname{var} x_t + \operatorname{vec} \Sigma_\varepsilon$, (as $\operatorname{vec} ABC = (C' \otimes A) \operatorname{vec} B$).
  - I.e. $\operatorname{vec} \operatorname{var} x_t = [I - (\phi_1 \otimes \phi_1)]^{-1} \operatorname{vec} \Sigma_\varepsilon$.
  - This is numerically unstable though, and not recommended in practice!
    - Matlab's dlyap command is preferable.

Companion form representation

- We have proven some nice theoretical results for a $\mathrm{VAR}(1)$.
- Luckily, any $\mathrm{VAR}(p)$ procedure may be transformed into a $\mathrm{VAR}(1)$ with an augmented state space.
  - The procedure is the same as that used when solving rational expectations models with multiple lags, or lagged expectations.
  - In particular, we define $y_t = [x_t' \quad x_{t-1}' \quad \cdots \quad x_{t-p+1}']'$.
  - Then:

$$y_t = \begin{bmatrix} [\phi_1 & \cdots & \phi_{p-1}] & \phi_p \\ I_{(p-1)n\times(p-1)n} & & & 0_{(p-1)n\times n} \end{bmatrix} y_{t-1} + \begin{bmatrix} \varepsilon_t \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

- The previously given condition for stationarity can be derived from this representation.

Cointegration

- Let x_t be an n dimensional stochastic process. If $x_t \sim I(1)$ (i.e. each component of x_t is non-stationary, and $\Delta x_t \sim I(0)$, i.e. each component of Δx_t is stationary), then x_t is described as "cointegrated" if there exists $\beta_1 \in \mathbb{R}^n$ such that $\beta_1' x_t \sim I(0)$.

- More generally, x_t is described as being cointegrated with rank $r \leq n$ if there exists a full-rank matrix $\beta = [\beta_1 \quad \cdots \quad \beta_r] \in \mathbb{R}^{n \times r}$ such that $\beta' x_t \sim I(0)$ (and this isn't true for any matrix of higher rank).

- The economic relevance of this concept comes from thinking about common trends.

From common trends to cointegration

- Suppose z_t is a k dimensional stochastic process, and $\gamma' z_t \sim I(1)$ for any $\gamma \in \mathbb{R}^k$, $\gamma \neq 0$.
- Suppose further that x_t is an $n \geq k$ dimensional stochastic process, with $x_t = Az_t + u_t$ where $u_t \sim I(0)$ and where A is full rank, so $A'A$ is invertible.
- Define $P \coloneqq A(A'A)^{-1}A'$. Then $Px_t = PAz_t + Pu_t = Az_t + Pu_t = x_t - u_t + Pu_t$, so $(I - P)x_t = (I - P)u_t$.
- Since A is full rank, by the singular value decomposition (SVD), there exists unitary matrices U $(n \times n)$ and V $(k \times k)$, and a diagonal matrix D $(k \times k)$, such that $A = [U_{\cdot 1} \quad U_{\cdot 2}] \begin{bmatrix} D \\ 0 \end{bmatrix} V' = U_{\cdot 1} D V'$, where $[U_{\cdot 1} \quad U_{\cdot 2}] = U$ is a partition of U with $U_{\cdot 1}$ $n \times k$.
- Then $P = U_{\cdot 1} DV'(VDU'_{\cdot 1} U_{\cdot 1} DV')^{-1} VDU'_{\cdot 1} = U_{\cdot 1} DV'(VDDV')^{-1} VDU'_{\cdot 1} = U_{\cdot 1} DV'VD^{-2}V'VDU'_{\cdot 1} = U_{\cdot 1} U'_{\cdot 1}$. Thus $I - P = I - U_{\cdot 1} U'_{\cdot 1} = U_{\cdot 2} U'_{\cdot 2}$.
- Hence, $U_{\cdot 2} U'_{\cdot 2} x_t = U_{\cdot 2} U'_{\cdot 2} u_t$, so $U'_{\cdot 2} x_t = U'_{\cdot 2} u_t \sim I(0)$. I.e. $U_{\cdot 2}$ is a cointegrating matrix.
- Now suppose δ is a cointegrating vector. Then $\delta'(Az_t + u_t) \sim I(0)$. By assumption, this implies $\delta' A = 0$, i.e. $\delta' U_{\cdot 1} = 0$, so $0 = \delta' U_{\cdot 1} U'_{\cdot 1} = \delta'(I - U_{\cdot 2} U'_{\cdot 2})$ meaning $\delta = U_{\cdot 2}(U'_{\cdot 2} \delta)$.
- Thus there is no larger rank cointegrating matrix, and x_t is cointegrated with rank $n - k$.

From cointegration to common trends

- Suppose x_t is cointegrated with a rank r cointegration matrix β.
- Given $\beta' x_t \sim I(0)$, there must exist some $I(0)$ stochastic process v_t such that $\beta' x_t = v_t$.
- Since β is full rank, by the SVD, there exists unitary matrices U ($n \times n$) and V ($r \times r$), and a diagonal matrix D ($r \times r$), such that $\beta = [U_{\cdot 1} \quad U_{\cdot 2}] \begin{bmatrix} D \\ 0 \end{bmatrix} V' = U_{\cdot 1} D V'$, where $[U_{\cdot 1} \quad U_{\cdot 2}] = U$ is a partition of U with $U_{\cdot 1}$ $n \times r$.
- Define $u_t \coloneqq U_{\cdot 1} D^{-1} V' v_t$, $A \coloneqq U_{\cdot 2}$ and $z_t \coloneqq U'_{\cdot 2} x_t$.
- Then, $u_t = U_{\cdot 1} D^{-1} V' V D U'_{\cdot 1} x_t = U_{\cdot 1} U'_{\cdot 1} x_t = (I - U_{\cdot 2} U'_{\cdot 2}) x_t$. I.e. $x_t = A z_t + u_t$.
  - Clearly A is full rank and $u_t \sim I(0)$.
- Suppose that there exists $\gamma \in \mathbb{R}^r$, $\gamma \neq 0$ such that $\gamma' z_t \sim I(0)$.
  - Then $(U_{\cdot 2}\gamma)' x_t \sim I(0)$.
  - But $U_{\cdot 2}\gamma$ is linearly independent of β (as $\beta = U_{\cdot 1} D V'$) hence the matrix $[\beta \quad U_{\cdot 2}\gamma]$ is full rank, which contradicts our assumption that x_t is cointegrated with rank r.
  - Hence $\gamma' z_t \sim I(1)$ for any $\gamma \in \mathbb{R}^r$.

“Error correcting” representation

- Given:

$$x_t = \mu + \phi_1 x_{t-1} + \phi_2 x_{t-2} + \cdots + \phi_p x_{t-p} + \varepsilon_t$$

- We have:

$$\begin{aligned}
\Delta x_t &= \mu + (\phi_1 - I)x_{t-1} + \phi_2 x_{t-2} + \cdots + \phi_p x_{t-p} + \varepsilon_t \\
&= \mu + (\phi_1 - I)x_{t-1} + \phi_2 x_{t-2} + \cdots + \phi_{p-2} x_{t-p+2} \\
&+ \left(\phi_{p-1} + \phi_p\right) x_{t-p+1} - \phi_p \Delta x_{t-p+1} + \varepsilon_t \\
&= \cdots \\
&= \mu + \left[\sum_{i=1}^{p} \phi_i - I\right] x_{t-1} - \left[\sum_{i=2}^{p} \phi_i\right] \Delta x_{t-1} - \left[\sum_{i=3}^{p} \phi_i\right] \Delta x_{t-2} - \cdots \\
&- \phi_p \Delta x_{t-p+1} + \varepsilon_t
\end{aligned}$$

  - You might recognise this representation from the ADF test.

Relationship between vector error correcting models (VECM) and cointegration

- Assuming $\Delta x_t \sim I(0)$, every term is stationary except perhaps $\left[\sum_{i=1}^{p} \phi_i - I\right] x_{t-1}$.
  - Hence this term too must be stationary.
- Let r be the rank of $\left[\sum_{i=1}^{p} \phi_i - I\right]$.
  - By (e.g.) the singular value decomposition, we may write $\left[\sum_{i=1}^{p} \phi_i - I\right] = \alpha\beta'$, where $\alpha, \beta \in \mathbb{R}^{n \times r}$ are full rank.
  - Then β must be a cointegration matrix for x_t, and so x_t is cointegrated with rank r.
  - If we choose β such that $\beta'\beta = I$, then we may interpret α as giving the rate of convergence to the long-run equilibrium determined by the cointegrating relationship.
- Suppose we knew β, then we could estimate α and the other parameters consistently via OLS.
  - This is the heart of the Engle and Granger (1987) two step-procedure.
  - If we know the number of cointegrating relationships, then β may be "super-consistently" pre-estimated using a simple regression in levels.

Consistency of a VAR with non-stationary variables

- The problem arises because although $\mathbb{E}\varepsilon_t x'_{t-1} = 0$, if x_{t-1} is non-stationary, then $T^{-1}\sum_{t=1}^{T}\varepsilon_t x'_{t-1}$ does not converge to 0 in probability.
  - Non-stationary regressors act as if they were endogenous, with all this entails (bias, non-Gaussian limiting distributions etc.).

- However, VARs in levels are asymptotically efficient, and any linear combination of parameters that has standard asymptotics in a VECM model with known cointegration matrix will also have standard asymptotics when the VAR in level is run (Sims, Stock and Watson 1990).
  - This includes alpha and the parameters on all lags except the first.
  - It does not include μ or any included trends.
- Standard asymptotics on the first lag may be produced via the Toda and Yamamoto (1995) trick.
  - Read a nice description here http://davegiles.blogspot.co.uk/2011/10/var-or-vecm-when-testing-for-granger.html .

Alternative approaches

- Pre-testing for cointegration following Johansen (1991), then estimating the VECM model via ML.
  - The danger is that this introduces a zero-one decision about the number of cointegration vectors.

- Running a so-called FM-VAR (Phillips 1995).
  - Like a VAR in levels, but requires no knowledge of the degree of cointegration.
  - Plus, it has lower finite sample bias, and (conservative) standard asymptotics for all lags.
  - However, it requires an estimate of variance at frequency 0, which is difficult in practice. (Read: impossible...)

Improving finite sample performance of VARs: For classical econometricians

- Use the boostrap!
- Idea: find a distribution that we can sample from which well approximates the distribution of the statistics of interest under the null.

- Two main ways of generating the sample:
  1. The sieve bootstrap: Estimate a VAR (possibly with more lags), but imposing the null hypothesis. Keep the residuals. Run a simulation of that VAR, where the shocks are drawn uniformly at random from the stored residuals, with length equal to that of the data, then run the test on that simulated series. This provides one sample. Repeat this many times to get the distribution under the null.
  2. The stationary block bootstrap: Difference source variables as necessary to induce stationarity in the data, unless running a test with the null of stationarity. Then construct a new stationary data series as follows. Pick a time at random from the data, and use this data point as the first observation. With some probability p, use the subsequent observation in the data as the subsequent observation in the new series (wrapping round if we currently have the final observation). With probability $1 - p$ pick another time at random and use that data point instead. Repeat until a series the length of the original data has been generated, then run the test on that sample. Again, repeat this many times to get the distribution of the test statistic.

Bayesian approaches

- Bayesian statistics has no problems with unit roots generating particularly funny distributions, asymptotic, or otherwise.

- Bayesian stats 101:
  - Assume we have some prior beliefs over the vector of parameters to be estimated, θ. These priors are captured in $p(\theta)$.
  - Suppose we observe the data X. Then we should update our posterior beliefs about θ via Bayes-rule: $p(\theta|X) = \frac{p(\theta \cap X)}{p(X)} = \frac{p(X|\theta)p(\theta)}{p(X)}$.
  - The optimal posterior point estimate of $f(\theta)$ (for any f we are interested in) is given by $\mathbb{E}[f(\theta)|X] = \int f(\theta) \frac{p(X|\theta)p(\theta)}{p(X)} d\theta$.
  - For some distributions. this integral will be analytically tractable.
  - More generally, we have to evaluate it numerically.
    - This may be done without knowing $p(X)$ using Markov Chain Monte Carlo (MCMC, aka the Metropolis-Hastings algorithm), which we will discuss later in the course.

Bayesian VARs

- A variety of priors have been proposed in the literature.
- Many use the conjugate-prior property of the Normal-Inverse-Wishart distribution.
  - If the prior is from this family, the posterior will be too.

- One prior of this type popular in macro is the "Minnesota prior" of Doan, Litterman and Sims (1984).
  - This has a distribution with prior mean equal to a random walk, and greater concentration in the prior at longer lags reflecting a belief that macro-time series have short memory.
  - This solves the curse of dimensionality of VAR models.
    - nT observations, but $\mathrm{O}(n^2T^{1+c})$ parameters.
  - It also tends to reduce MSE.
    - In fact this is true even in basic regression, OLS is the best linear *unbiased* estimator, but estimators that deliberately bias towards zero such as ridge regression have lower MSE.
- Normal-Inverse-Wishart type priors are probably over-used in the literature though.
  - They were developed at a time when MCMC was computationally infeasible. It isn't now.
  - More reasonable priors would (at a minimum) truncate tails to rule out explosive behaviour a priori.
  - It is also desirable to use priors with peaks at zero (i.e. a discontinuity in the first derivative), and possibly even mass there. These induce sparsity, which makes for more readily interpretable results. Appropriate choice of prior gives an "oracle property" that recovers the true DGP asymptotically.

Conclusion

- We've seen the basics of reduced form time series econometrics.
- This is all you need for macroeconomic forecasting.

- However, to make inferences about the macro-economy, we will need to be sure that we have identified structural parameters.
  - I.e. we want to be sure that with different policy we wouldn't have obtained different estimates.

- Next week we will start to look at structural approaches, including structural VARs.