Lecture III

Harvard Econ 2416

Professor Gabriel Chodorow-Reich

Spring 2021

Outline

Outline

Time series properties

- Series have trends.
- Serial correlation.
- We see one realization of events.
- Stationarity and ergodicity required for valid inference.

STATIONARITY

Let $\{y_t\}$ denote a sequence of random variables. The process y_t is **strictly stationary** if the joint distribution of $(y_t, y_{t+1}, \ldots, y_{t+k})$ is the same for all t. For example, (y_1, y_5) has the same joint distribution as (y_{12}, y_{16}).

Define the following moments:

- Conditional mean: $\mu_t = E[y_t]$;
- Conditional variance: $\gamma_t(0) = var(y_t) = E[y_t - \mu_t]^2$;
- Autocovariance at lag k:
  $\gamma_t(k) = cov(y_t, y_{t-k}) = E[(y_t - \mu_t)(y_{t-k} - \mu_{t-k})]$;
- Autocorrelation at lag k: $\rho_t(k) = \gamma_t(k)/\gamma_t(0)$.

A stochastic process $\{y_t\}$ is **covariance stationary** if the first and second moments do not depend on t and are finite. That is, for all t,

- $E[y_t] = \mu < \infty$;
- $\gamma_t(0) = \gamma(0) < \infty$;
- $\gamma_t(k) = \gamma(k)$.

Ergodicity

A stationary process $\{y_t\}$ is said to be **ergodic** if for any two bounded functions f and g:

$$\lim_{T\to\infty} \left| E\left[f\left(y_t,\ldots,y_{t+k}\right) g\left(y_{t+T},\ldots,y_{t+T+k}\right)\right]\right| = \left| E\left[f\left(y_t,\ldots,y_{t+k}\right)\right]\right| \left| E\left[g\left(y_{t+T},\ldots,y_{t+T+k}\right)\right]\right|$$

- In words, ergodicity says that if two sequences of y are "far enough" apart, then one can treat them as independent (the covariance of any two functions of the sequences is zero).
- Excludes permanent path dependence. For example:

$$x_t \sim iidN(0,1), z \sim N(0,1), y_t = x_t + z \Rightarrow Cov(y_t, y_{t-j}) = var(z) = 1.$$

- If we have one long history, ergodicity and stationarity mean we can make inference (LLN, CLT) about other potential histories.

INNOVATIONS

y_t is a **serially uncorrelated process** if $\gamma(k) = 0$ for all $k \geq 1$.

- Type 1 (**white noise**): $E[e_t] = 0$, $E[e_t e_s] = 0$ for $t \neq s$.
- Type 2 (**independent white noise**): $e_t \sim iid(0, \sigma^2)$. Note that
  - $E[e_t] = 0$,
  - $E[e_t e_s] = 0$ if $t \neq s$,
  - $E[e_t^2] = \sigma^2$.
- Type 3 (**martingale difference sequence**): Let $y_t = y_{t-1} + e_t$. e_t is a MDS if $E[e_t | e_{t-1}, e_{t-2}, \ldots, y_{t-1}, y_{t-2}, \ldots] = 0$.
  - The process y_t is a martingale if $E[y_t | y_{t-1}, y_{t-2}, \ldots] = y_{t-1}$.
- $IWN \Rightarrow MDS \Rightarrow WN$.

Lag operator

- Definition:

$$Ly_t = y_{t-1}.$$

- Commutative with multiplication:

$$L(\alpha y_t) = \alpha L y_t.$$

- Distributive over addition:

$$L(x_t + y_t) = Lx_t + Ly_t.$$

- "Powers up":

$$L^p y_t = y_{t-p}.$$

AUTOREGRESSIVE MOVING-AVERAGE (ARMA) OVERVIEW

- Let e_t be a white noise process.
- General ARMA(p,q) process:

$$\begin{aligned} y_t = \alpha_1 y_{t-1} + \ldots + \alpha_p y_{t-p} + e_t + \theta_1 e_{t-1} + \ldots \theta_q e_{t-q} \\ \Leftrightarrow (1 - \alpha_1 L - \ldots - \alpha_p L^p)\, y_t = (1 + \theta_1 L + \ldots + \theta_q L^q)\, e_t \\ \Leftrightarrow \alpha(L) y_t = \theta(L) e_t \\ \Leftrightarrow y_t = \frac{\theta(L)}{\alpha(L)} e_t = \psi(L) e_t. \end{aligned}$$

- Note: y_t is mean zero (why?). But this is WLOG, because we can always demean a series before analyzing it.

Moving average

$$y_t = \mu + e_t + \theta_1 e_{t-1} + \ldots + \theta_q e_{t-q} = \mu + \theta(L) e_t.$$

Properties:

- Mean:

$$E[y_t] = \mu.$$

- Variance:

$$\gamma(0) = E[(y_t - \mu)^2] = \left(1 + \theta_1^2 + \ldots + \theta_q^2\right)\sigma^2.$$

- Auto-covariance:

$$\gamma(j) = \begin{cases} E\left[(y_t - \mu)(y_{t-j} - \mu)\right] = \left(\theta_j + \theta_{j+1}\theta_1 + \ldots + \theta_q \theta_{q-j}\right)\sigma^2, \ j \leq q \\ 0, j > q. \end{cases}$$

$\Rightarrow$ Serial correlation dies out after q lags.

AUTO REGRESSIVE

$$\begin{aligned} y_t &= c + \alpha_1 y_{t-1} + \ldots + \alpha_p y_{t-p} + e_t \\ \Leftrightarrow \alpha(L)(y_t - \mu) &= e_t, \\ \mu &= c/\alpha(1) = c/[1 - \alpha_1 - \alpha_2 - \ldots - \alpha_p]. \end{aligned}$$

- If the roots of $1 - \alpha_1 z - \alpha_2 z^2 - \ldots - \alpha_p z^p = 0$ all lie outside the unit circle, then can invert to obtain covariance stationary MA representation:

$$y_t = \mu + \theta(L) e_t.$$

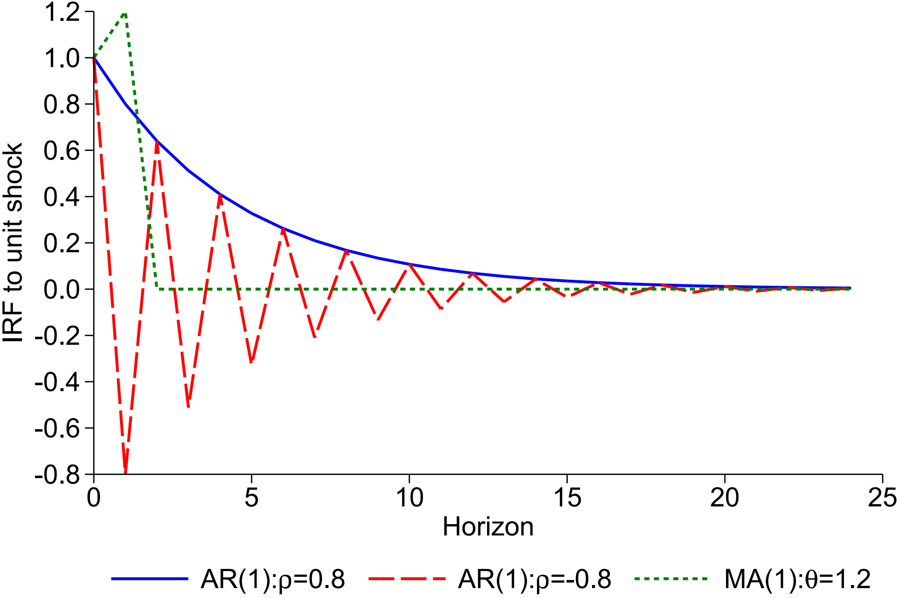
- **Impulse response** of y_t to e_{t-j} is θ_j.

AR(1) EXAMPLE

$$y_t = c + \alpha y_{t-1} + e_t.$$

- Mean:
$$E[y_t] = \mu = \frac{c}{1-\alpha}.$$
- Variance:
$$\gamma(0) = \frac{\sigma^2}{1-\alpha^2}.$$
- Autocorrelation:
$$\rho(j) = \alpha^j.$$
- Stability criterion: $|\alpha^{-1}| > 1 \Longrightarrow |\alpha| < 1$.
- MA(∞) representation:
$$y_t = \mu + \sum_{j=0}^{\infty} \alpha^j e_{t-j}.$$

IMPULSE RESPONSE EXAMPLES



Model selection

- Parsimony: use smallest number of parameters for adequate fit.
- **Information criterion**:

$$\hat{k} = argmin_{k\in[0,k_{\max}]} \ln \hat{\sigma}^2 + \frac{1}{T} C_T(k).$$

  - $k = p + q + d$, d is number of deterministic terms.
  - $IC = \ln \hat{\sigma}^2 + \frac{1}{T} C_T(k)$ is information criterion.
  - $\hat{\sigma}^2 = \frac{1}{T}\sum_{t=1}^{T} \hat{e}_t^2$ non-increasing in k, $\frac{1}{T} C_T(k)$ is penalty.
- AIC : $C_T(k) = 2k$.
- BIC : $C_T(k) = k \log(T)$.
- Finite order $AR(p)$ models: BIC consistently estimates p, AIC over-parameterizes with positive probability.
- Models with MA components: BIC tends to under-parameterize.
- Implementation: fix sample based on maximum lag length considered.
- Other alternatives in machine-learning literature.

Outline

Overview

$$x_t = G(L)y_t, \; c_t = y_t - x_t = [1 - G(L)]y_t.$$

- For business cycle analysis, we want to remove long-term trends and short-term noise. Filtered series can induce stationarity and ergodicity.
- For a time series y_t, c_t is the cycle.
- Deterministic filters: linear trend, quadratic trend, etc.
- Two popular filters: **Hodrick-Prescott** and **Baxter-King** filters.
- The HP filter uses a penalty function to extract a smooth trend. The residual component is the cycle.
- The BK filter estimates the spectrum and retains components with frequencies within a specified interval.

HP filter (JME 1999)

$$\min_{\{x_t\}} \sum_{t=1}^{T} (y_t - x_t)^2 + \lambda \sum_{t=2}^{T-1} [(x_{t+1} - x_t) - (x_t - x_{t-1})]^2 .$$

- Objective: minimize cycle $c_t = y_t - x_t$, but penalize changes in $\tau_t = \Delta x_t$.
- λ parameterizes the cost of changes in the trend τ.
  - $\lambda \to 0$: trend is original series.
  - $\lambda \to \infty$: trend is least squares linear trend.
- If $c_t = y_t - x_t \sim iidN(0, \sigma_c^2), \Delta\tau_t \sim iidN(\sigma_\tau^2)$, then $\lambda = \sigma_c^2 / \sigma_\tau^2$ will consistently recover the true $\{x_t\}$.
- HP heuristic: at quarterly frequency, $\sigma_c / \sigma_\tau \approx 40 \Longrightarrow \lambda = 1600$.
  - Ravn and Uhlig (RESTAT 2002) suggest scaling to 4th power of frequency ratio for alternative frequencies: 6.25 annual, 129,000 monthly, etc.

HP FILTER

$$\min_{\{x_t\}} \sum_{t=1}^{T} (y_t - x_t)^2 + \lambda \sum_{t=2}^{T-1} [(x_{t+1} - x_t) - (x_t - x_{t-1})]^2 .$$

Interior FOC:

$$\begin{aligned} y_t &= x_t + \lambda [(x_t - x_{t-1}) - (x_{t-1} - x_{t-2})] \\ &\quad - 2\lambda [(x_{t+1} - x_t) - (x_t - x_{t-1})] \\ &\quad + \lambda [(x_{t+2} - x_{t+1}) - (x_{t+1} - x_t)] \\ &= [1 + \lambda \left(L^{-2} - 4L^{-1} + 6 - 4L + L^2\right)] x_t \\ &= H(L) x_t . \end{aligned}$$

Boundary FOC:

$$\begin{aligned} x_1 &: y_1 = x_1 + \lambda [(x_3 - 2x_2 + x_1)] \\ x_2 &: y_2 = x_2 + \lambda [-2(x_3 - 2x_2 + x_1) + (x_4 - 2x_3 + x_2)] \\ x_{T-1} &: y_{T-1} = x_{T-1} + \lambda [(x_{T-1} - 2x_{T-2} + x_{T-3}) - 2(x_T - 2x_{T-1} + x_{T-2})] \\ x_T &: y_T = x_T + \lambda [(x_T - 2x_{T-1} + x_{T-2})] . \end{aligned}$$

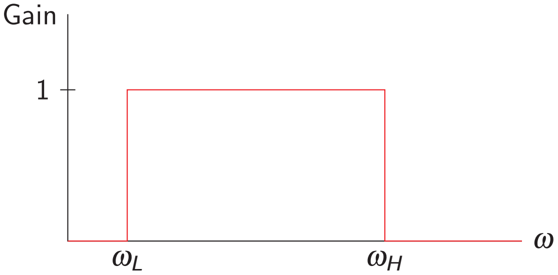
HP FILTER

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ \vdots \\ y_{T-3} \\ y_{T-2} \\ y_{T-1} \\ y_T \end{pmatrix} = \begin{pmatrix} 1+\lambda & -2\lambda & \lambda & 0 & 0 & 0 & \dots & 0 \\ -2\lambda & 1+5\lambda & -4\lambda & \lambda & 0 & 0 & \dots & 0 \\ \lambda & -4\lambda & 1+6\lambda & -4\lambda & \lambda & 0 & \dots & 0 \\ 0 & \lambda & -4\lambda & 1+6\lambda & -4\lambda & \lambda & 0 & \dots \\ \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \\ 0 & \dots & \lambda & -4\lambda & 1+6\lambda & -4\lambda & \lambda & 0 \\ 0 & \dots & 0 & \lambda & -4\lambda & 1+6\lambda & -4\lambda & \lambda \\ 0 & \dots & 0 & 0 & \lambda & -4\lambda & 1+5\lambda & -2\lambda \\ 0 & \dots & 0 & 0 & 0 & \lambda & -2\lambda & 1+\lambda \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ \vdots \\ x_{T-3} \\ x_{T-2} \\ x_{T-1} \\ x_T \end{pmatrix}$$

- $H(L)$ is a matrix. Invert it to obtain x_t.
- $c_t = y_t - x_t$.

BK filter (RESTAT 1999)

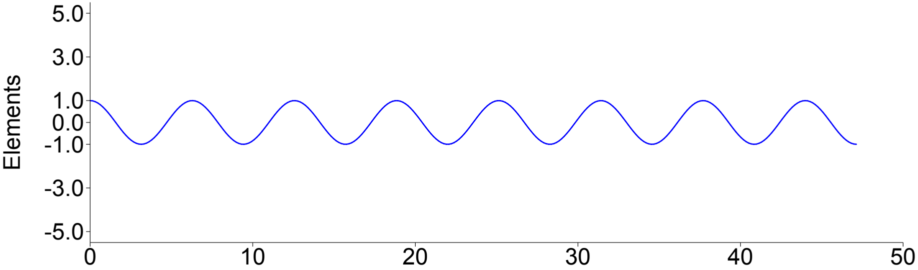
- Ideal bandpass filter

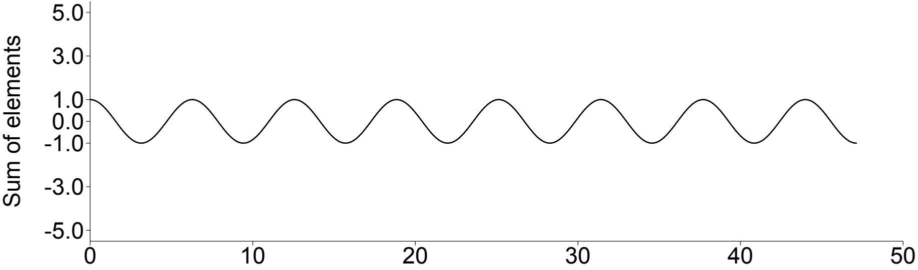


- Basic idea: decompose y_t into low frequency, medium frequency, and high frequency movements, and retain the medium frequency movements as the business cycle component of the series.

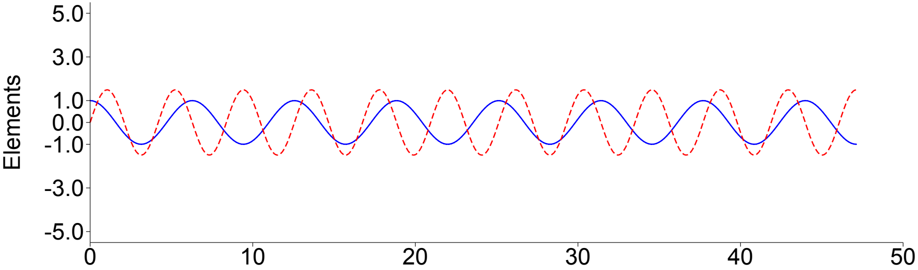
- To flesh this out a bit, need to digress onto the frequency domain.

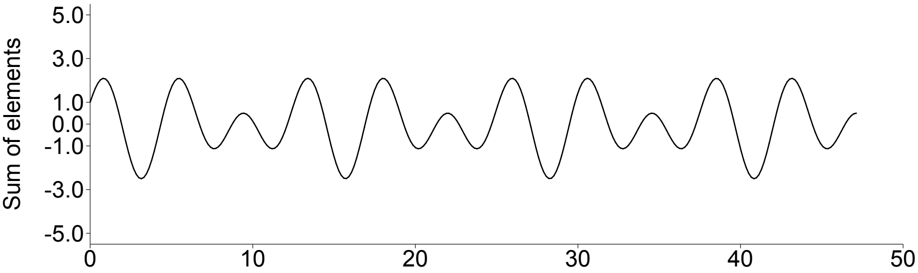
Sinusoidal process: 1 element



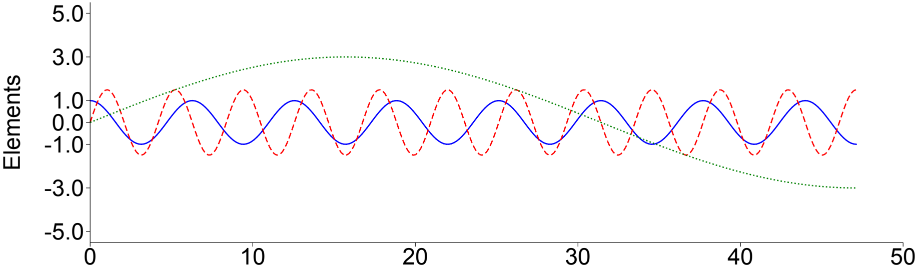


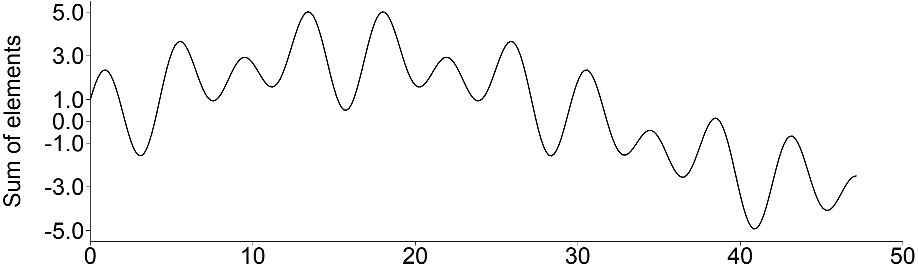
SINUSOIDAL PROCESS: 2 ELEMENTS



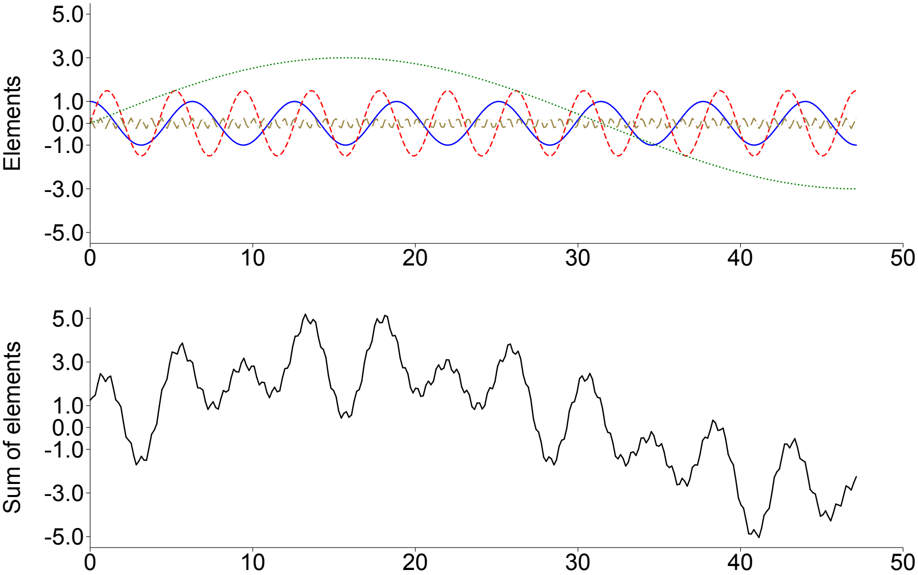


Sinusoidal process: add low frequency

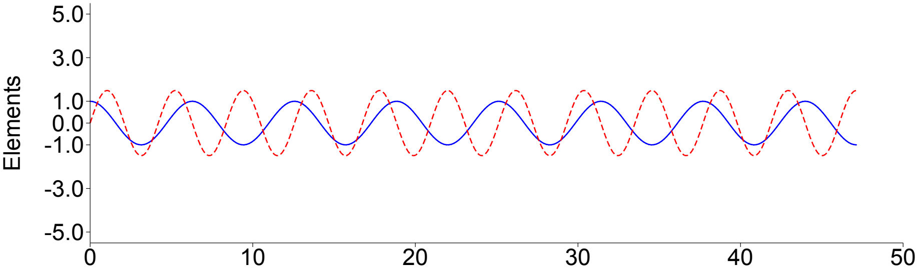


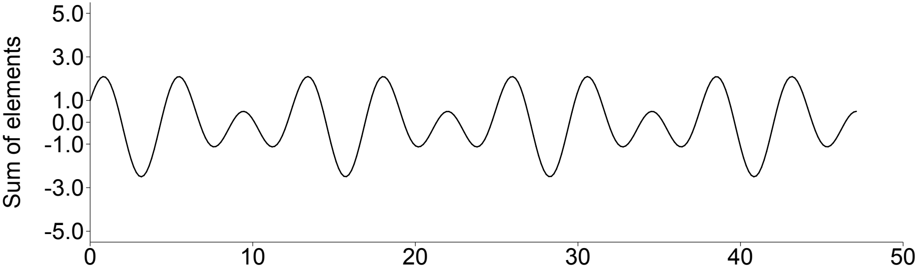


Sinusoidal process: add high frequency



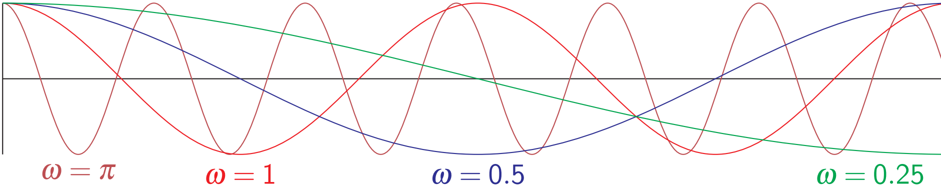
Bandpass filter: remove high & low frequency





Deterministic processes

$$Y_t = \cos(\omega t)$$



- The period is equal to $\frac{2\pi}{\omega}$, $Y_0 = 1$, and the amplitude equals 1.
- Next consider $Y_t = a\cos(\omega t) + b\sin(\omega t)$. The period remains equal to $\frac{2\pi}{\omega}$, $Y_0 = a$, and the amplitude is now $\sqrt{a^2 + b^2}$.

STOCHASTIC PROCESSES

- Now let $Y_t = a\cos(\omega t) + b\sin(\omega t)$, a and b mutually-uncorrelated, mean-zero, random variables with common variance σ^2:

$$
\begin{aligned}
E[Y_t] &= 0 \\
Var[Y_t] &= Var[a\cos(\omega t) + b\sin(\omega t)] \\
&= \cos^2(\omega t)\, Var[a] + \sin^2(\omega t)\, Var[b] \\
&= \sigma^2 \\
Cov[Y_t, Y_{t-k}] &= Cov[a\cos(\omega t) + b\sin(\omega t), a\cos(\omega[t-k]) + b\sin(\omega[t-k])] \\
&= \sigma^2 [\cos(\omega t)\cos(\omega[t-k]) + \sin(\omega t)\sin(\omega[t-k])] \\
&= \sigma^2 \{ \cos(\omega t)[\cos(\omega t)\cos(\omega k) + \sin(\omega t)\sin(\omega k)] \\
&+ \quad \sin(\omega t)[\sin(\omega t)\cos(\omega k) - \cos(\omega t)\sin(\omega k)] \} \\
&= \sigma^2 \cos(\omega k).
\end{aligned}
$$

- The way to think about this is drawing a, b once and for all, and then generating the periodic series.

Stochastic processes, more elements

- Next consider $Y_t = \sum_{j=1}^{N} a_j \cos(\omega_j t) + b_j \sin(\omega_j t)$, where a_j and b_j are mutually-uncorrelated mean-zero random variables with variance σ_j^2.
- Y_t is now the sum of multiple series each with its own periodicity:

$$
\begin{aligned}
E[Y_t] &= 0 \\
Var[Y_t] &= \sum_{j=1}^{N} \sigma_j^2 \\
Cov[Y_t, Y_{t-k}] &= \sum_{j=1}^{N} \sigma_j^2 \cos(\omega_j k).
\end{aligned}
$$

- The variance of Y_t can be decomposed into the variance of elements with different frequencies.
- Spectral analysis describes the relationship between frequency and the contribution to the variance.

Stochastic processes, even more elements

- Now add even more components:

$$Y_t = \int_0^{\pi} \cos(\omega t)\, da(\omega) + \int_0^{\pi} \sin(\omega t)\, db(\omega).$$

- As before, $da(\omega)$ and $db(\omega)$ are mutually-uncorrelated mean-zero random variables, with common variance that depends on frequency.

- The variance function $S(\omega)$ is called the spectrum.

- The Cramer representation theorem states that any covariance-stationary process has this representation.

Change of notation

- Recall the Euler formula $e^{i\omega t} = \cos(\omega t) + i\sin(\omega t)$.
- Scalar representation:

$$\begin{aligned}
Y_t &= a\cos(\omega t) + b\sin(\omega t) \\
&= \frac{1}{2}[\cos(\omega t) + i\sin(\omega t)][a - ib] + \frac{1}{2}[\cos(\omega t) - i\sin(\omega t)][a + ib] \\
&= \frac{1}{2}e^{i\omega t}[a - ib] + \frac{1}{2}e^{-i\omega t}[a + ib] = e^{i\omega t}Z + e^{-i\omega t}\bar{Z},
\end{aligned}$$

  $Z \equiv \frac{1}{2}[a - ib]$ and $\bar{Z} \equiv \frac{1}{2}[a + ib]$ is the complex conjugate of Z.
- Integral representation:

$$\begin{aligned}
Y_t &= \int_0^{\pi} \cos(\omega t)\, da(\omega) + \int_0^{\pi} \sin(\omega t)\, db(\omega) \\
&= \int_0^{\pi} e^{i\omega t} dZ(\omega) + \int_0^{\pi} e^{-i\omega t} d\bar{Z}(\omega) = \int_{-\pi}^{\pi} e^{i\omega t} dZ(\omega),
\end{aligned}$$

  $dZ(\omega) = \frac{1}{2}[da(\omega) - idb(\omega)]$ if $\omega \geq 0$,
  $dZ(-\omega) = \frac{1}{2}[da(\omega) + idb(\omega)] = d\bar{Z}(\omega)$ if $\omega < 0$.

Spectrum

- Note:

$$\begin{aligned} E\left[dZ(\omega)\right] &= 0 \\ Var\left[dZ(\omega)\right] &= E\left[dZ(\omega)\,d\bar{Z}(\omega)\right] \\ &\equiv S(\omega)\,d\omega. \end{aligned}$$

- $S(\omega)$ is the spectrum and relates frequency ω to its variance.
- Autocovariance function (use $E[dZ(\omega)\,dZ(\omega')] = 0$ for $\omega \neq \omega'$):

$$\begin{aligned} \gamma_k &\equiv E\left[Y_t Y_{t-k}\right] = E\left[Y_t \bar{Y}_{t-k}\right] \\ &= E\left[\int_{-\pi}^{\pi} e^{i\omega t} dZ(\omega) \int_{-\pi}^{\pi} e^{-i\omega(t-k)} d\bar{Z}(\omega)\right] \\ &= \int_{-\pi}^{\pi} e^{i\omega t} e^{-i\omega(t-k)} E\left[dZ(\omega)\,d\bar{Z}(\omega)\right] = \int_{-\pi}^{\pi} e^{i\omega k} S(\omega)\,d\omega. \end{aligned}$$

- Spectral representation theorem:

$$S(\omega) = \frac{1}{2\pi}\left[\sum_{k=-\infty}^{\infty} \gamma_k e^{-i\omega k}\right] = \frac{1}{2\pi}\left[\gamma_0 + 2\sum_{k=1}^{\infty} \gamma_k \cos(\omega k)\right].$$

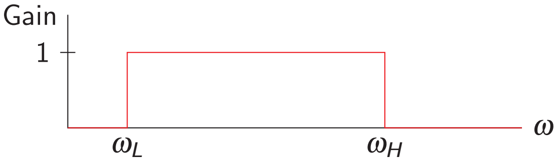
Frequency domain filter

- Idea is to amplify certain frequencies and mute other frequencies.
- Can also do phase shift, but less common.
- We call the gain the amplitude multiplicand at frequency ω.
- A linear filter is just the sum of the gain and phase shift applied to each frequency.
- Final task is to translate frequency domain filter into time domain weights:

$$c_t = [1 - G(L)]y_t = \sum_j a_j y_{t-j}.$$

BK filter

- Ideal bandpass filter



- With infinite observations, exact time domain weights are

$$j \neq 0: \;\; a_j = \frac{\sin(\omega_L j) - \sin(\omega_H j)}{\pi j},$$

$$j = 0: \;\; a_0 = \frac{\omega_L - \omega_H}{\pi}.$$

- In finite sample, truncate at q observations, and apply correction:

$$b_j = a_j - \theta, \; j = 1, 2, \ldots q,$$

$$\theta \equiv \frac{\sum_{j=-q}^{q} a_j}{2q+1}.$$

Practical details

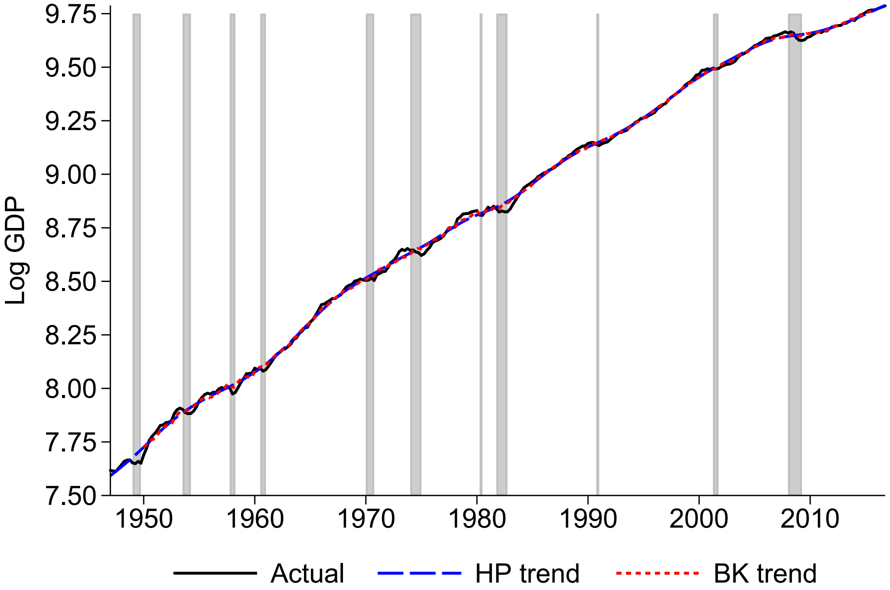
- Exact time domain weights require infinite observations because ideal weights do not decay. Intuition from the frequency domain is that with finite data we cannot obtain the exact spectrum from estimated autocovariances.

- Baxter and King suggest setting q to equivalent of 3 years.

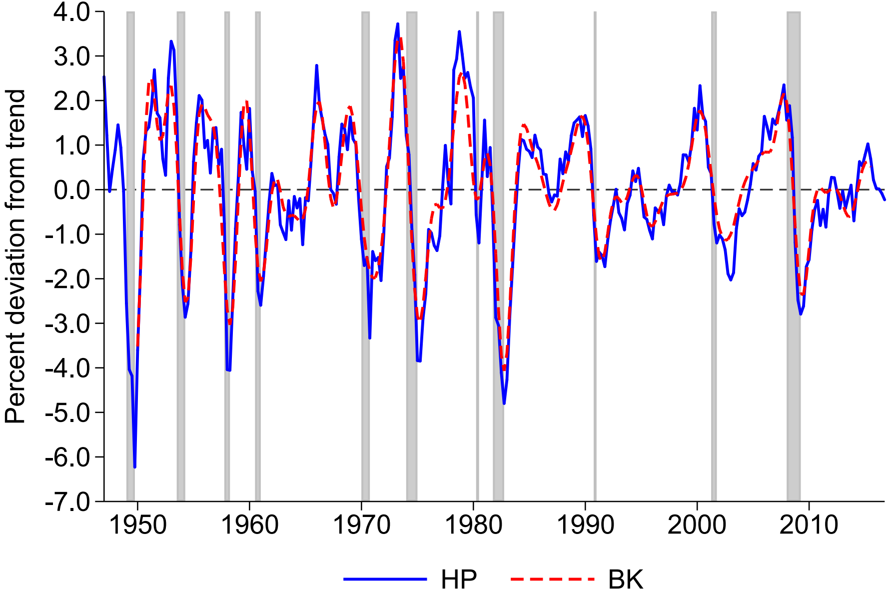
- Higher $q \Rightarrow$ weights closer to ideal, but at loss of data at sample start and end.

- Burns and Mitchell define business cycle as 1.5-8 years. Set ω_L, ω_H accordingly.

Log GDP and its Trend



HP versus BK



Percent deviation from trend
4.0
3.0
2.0
1.0
0.0
-1.0
-2.0
-3.0
-4.0
-5.0
-6.0
-7.0
1950
1960
1970
1980
1990
2000
2010
HP
BK

Hamilton (RESTAT 2018) critique

- HP and BK are both two-sided filters. Use caution when applying to expectational errors, Granger causality, etc.

- Recovered trends are functions of full time series. This can induce a correlation structure where none existed. (But so is seasonally-adjusted data.)

- Two-sided filters may perform poorly at sample start and end.

- If comparing data to model simulated data, filter both.

Hamilton (RESTAT 2018) filter

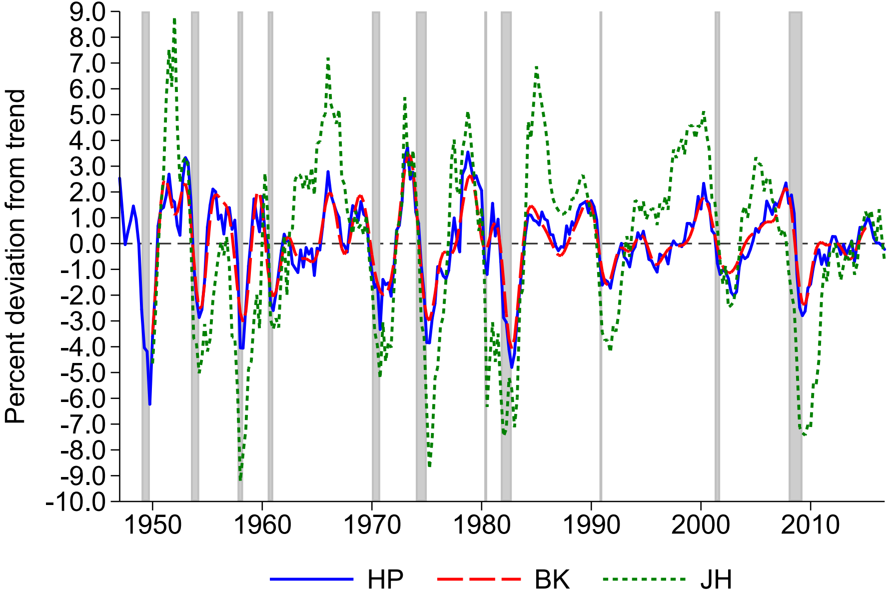
- Regress y_{t+h} on $y_t, y_{t-1}, \ldots, y_{t-p}$.

- Intuitively, cyclical component is unforecastable part of series.

- Hamilton proves induces stationarity in wide class of settings.

- Suggests setting $h = 2$ years (i.e. $h = 8$ if quarterly data), $p - 4$.

HP versus BK versus JH



Outline

OUTLINE

Overview

- Dominant methodology from mid 1980s to mid 2000s.
- Joint system of equations, usually estimated by OLS.
- Identify shocks from restrictions on moments of the residuals.
- Produce easily interpretable impulse response functions which can guide and discipline theory and policy.
- Usually associated with transparent identification assumptions.
- Identification assumptions not always plausible.
- Source of variation in treatment not always transparent.

Sims' EMCA 1980 critique

...what "economic theory" tells us about them is mainly that any variable which appears on the right-hand-side of one of these equations belongs in principle on the right-hand-side of all of them. To the extent that models end up with very different sets of variables on the right-handsides of these equations, they do so not by invoking economic theory, but (in the case of demand equations) by invoking an intuitive, econometrician's version of psychological and sociological theory, since constraining utility functions is what is involved here. Furthermore, unless these sets of equations are considered as a system in the process of specification, the behavioral implications of the restrictions on all equations taken together may be much less reasonable than the restrictions on any one equation taken by itself.

Reduced form VAR

- Consider a system with k equations, $T+p$ observations, and p lags.
- The j,t equation of the **reduced form VAR** take the form:

$$y_{j,t} = \beta_{j1,1} L^1 y_{1,t} + \ldots + \beta_{jk,1} L^1 y_{k,t} + \beta_{j1,2} L^2 y_{1,t} + \ldots \beta_{jk,p} L^p y_{k,t} + e_{j,t}.$$

- The k period t equations take the form:

$$\underbrace{\begin{pmatrix} y_{1,t} \\ \vdots \\ y_{k,t} \end{pmatrix}}_{Y_t : kx1} = \underbrace{\begin{pmatrix} \beta_{11,1} & \beta_{12,1} & \cdots & \beta_{1k,1} \\ \beta_{21,1} & \beta_{22,1} & \cdots & \beta_{2k,1} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{k1,1} & \beta_{k2,1} & \cdots & \beta_{kk,1} \end{pmatrix}}_{B_1 : kxk} L \begin{pmatrix} y_{1,t} \\ \vdots \\ y_{k,t} \end{pmatrix} + \ldots + \underbrace{\begin{pmatrix} \beta_{11,p} & \beta_{12,p} & \cdots & \beta_{1k,p} \\ \beta_{21,p} & \beta_{22,p} & \cdots & \beta_{2k,p} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{k1,p} & \beta_{k2,p} & \cdots & \beta_{kk,p} \end{pmatrix}}_{B_p : kxk} L^p \begin{pmatrix} y_{1,t} \\ \vdots \\ y_{k,t} \end{pmatrix} + \underbrace{\begin{pmatrix} e_{1,t} \\ \vdots \\ e_{k,t} \end{pmatrix}}_{e_t : kx1}$$

- Let B_l be a matrix containing $\{\beta_{ij,l}\}$, $Y_t = (y_{1,t} \ldots y_{k,t})'$:

$$Y_t = B_1 L Y_t + \ldots B_p L^p Y_t + e_t.$$

$$B(L) Y_t = e_t, \; E\left[e_t e_t'\right] = \Omega_t.$$

MA representation

$$Y_t = B_1 L Y_t + \ldots B_p L^p Y_t + e_t.$$

- VAR(p) is covariance stationary if all values of z satisfying

$$|I_k - B_1 z - \ldots - B_p z^p|$$

lie outside the unit circle.

- A covariance stationary VAR has the vector MA(∞) representation

$$Y_t = \tilde{\Psi}(L) e_t = e_t + \tilde{\Psi}_1 e_{t-1} + \tilde{\Psi}_2 e_{t-2} + \ldots$$

.

- $\tilde{\Psi}_j$ gives impulse response at horizon j.

VAR as SUR

- Write the jth equation as:

$$\underbrace{y_j}_{(T-p)x1} = \underbrace{\begin{pmatrix} Ly_1 & \ldots & L^p y_1 & \ldots & Ly_j \ldots & Ł^p y_j & \ldots L^p y_k \end{pmatrix}}_{X_j:(T-p)x(pk)} \underbrace{\beta_j(L)}_{(pk)x1} + \underbrace{e_j}_{(T-p)x1}.$$

  - Note: $X_1 = X_2 = \ldots = X_k$.

- Stacking equations:

$$\begin{pmatrix} y_1 \\ \vdots \\ y_k \end{pmatrix} = \begin{pmatrix} X_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & X_k \end{pmatrix} \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} e_1 \\ \vdots \\ e_k \end{pmatrix}.$$

- More compactly:

$$Y = X\beta + e.$$

Comments

- Reduced form VARs have proved popular for forecasting.
- Kruskal theorem: if the right hand side variables are the same for every equation, OLS coincides with GLS. In practice, use equation-by-equation OLS to estimate $\hat{B}(L)$ and $\hat{e}_t$.
- Lag selection:

$$\text{AIC} = \log\left|\Omega\right| + 2n/T.$$
$$\text{BIC} = \log\left|\Omega\right| + n\log(T)/T.$$

  - $n = k^2 p$.
  - Log likelihood with normal errors is $-\frac{T}{2}\left[\ln 2\pi + \log\left|\Omega\right| + k\right]$. IC based on log likelihood after dropping constants.
- Alternative: likelihood ratio test.

Outline

Structural VAR

- The **structural VAR** has the form:

$$A_0 Y_t = A_1 L Y_t + \dots A_p L^p Y_t + v_t,$$

where

$$E\left[v_t v_t'\right] = \Sigma_t,$$

and Σ_t is diagonal matrix with positive entries on the main diagonal.

- Σ diagonal means that innovations are uncorrelated. In economic terms, can think of a disturbance to only one part of the system. (WLOG?)
- Define: $A(L) \equiv [A_0 - A_1 L - \dots - A_p L^p]$:

$$A(L) Y_t = v_t.$$

Comparison

- Reduced form:

$$Y_t = B_1 L Y_t + \dots B_p L^p Y_t + e_t,\ E\left[e_t e_t'\right] = \Omega_t.$$

- Structural:

$$A_0 Y_t = A_1 L Y_t + \dots A_p L^p Y_t + v_t,\ E\left[v_t v_t'\right] = \Sigma_t.$$

- Comparing:

$$\begin{aligned} A_i &= A_0 B_i, \\ A(L) &= A_0 B(L), \\ v_t &= A_0 e_t, \\ Var\left[e_t\right] &= \Omega_t = A_0^{-1} \Sigma_t A_0^{-1\prime}. \end{aligned}$$

Rank/invertibility condition

- Invertibility: $v_t = \text{Proj}[v_t | Y_t, Y_{t-1}, \ldots]$.
- In words: structural shocks v_t linearly determined by current and lagged observables.
- Invertibility $\Rightarrow e_t = A_0^{-1} v_t$.
- In words: observable innovations are linear combinations of the structural shocks and span the same space.
- With invertibility, structural shocks are redundant for forecasting: $\text{Proj}[Y_t | Y_{t-1}, Y_{t-2}, \ldots, v_{t-1}, v_{t-2}, \ldots] = \text{Proj}[Y_t | Y_{t-1}, Y_{t-2}, \ldots]$.
- Interpretation: if structural shocks add forecasting power to observables, then VAR has omitted variables.
- Structural VARs *always* assume invertibility, even if it is not explicitly stated as an identification assumption.

Order condition for identification

- Assume $\Omega_t = \Omega$, $\Sigma_t = \Sigma$, and invertibility holds.

- Estimate $\hat{\Omega}$ from matrix of reduced form residuals, $\hat{\Omega} = \frac{1}{T}\sum_{t=1}^{T} \hat{e}_t \hat{e}_t{}' \to^p \Omega$.

- Symmetry of $\hat{\Omega}$ implies $k(k+1)/2$ free parameters.

- The decomposition $A_0^{-1}\Sigma_t A_0^{-1\prime}$ has $k^2 + k$ free parameters.

- Identification requires $\left[k^2 + k\right] - \frac{k(k+1)}{2} = \frac{k(k+1)}{2}$ additional restrictions.

- k restrictions come from normalizing either $\Sigma = I$ or $diag\,(A_0) = I$, leaving $\frac{k(k-1)}{2}$ behavioral restrictions to be imposed (interpret).

- This is the order condition for identification.

CHOLESKY

- Leading example of additional restrictions: **Cholesky ordering**.
- Linear algebra fact: any symmetric, positive definite matrix can be written as $\Omega = RR'$, where R is lower triangular and unique.
  - Constructive proof using LU factorization.
- Covariance matrices are symmetric positive definite.
- Compare: $\Omega = A_0^{-1}\Sigma A_0^{-1\prime} = RR' \Longrightarrow A_0^{-1}\Sigma^{\frac{1}{2}} = R$.
- R lower triangular imposes recursive structure on contemporaneous response to structural shocks.
- R unique given $\hat{\Omega}$, but $\hat{\Omega}$ depends on ordering of variables in VAR.

Cholesky example

- Bivariate VAR in government purchases and output, government purchases ordered first:

$$Y_t = \{\text{Government spending}_t, \text{Output}_t\}'.$$

$$\begin{pmatrix} g_t \\ y_t \end{pmatrix} = \sum_{i=1}^{p} B_i \begin{pmatrix} g_{t-i} \\ y_{t-i} \end{pmatrix} + \begin{pmatrix} 1 & 0 \\ r_{yg} & 1 \end{pmatrix} \begin{pmatrix} v_{g,t} \\ v_{y,t} \end{pmatrix}.$$

- Σ un-normalized.
- Identification assumption: government spending does not respond to contemporaneous shocks to output, but output responds to contemporaneous shocks to government spending.
- Recursive intuition: regress g_t on lags of g_t and lags of y_t, recovering the structural residuals $v_{g,t}$. Then regress y_t on lags of y_t, lags of g_t, and $v_{g,t}$ to recover the structural residuals $v_{y,t}$.
- Invertibility: are there other shocks that also affect g and y?

Outline

Impulse response functions

- Moving average representation of Y_t:

$$Y_t = [\Psi_0 + \Psi_1 L + \ldots +] \, v_t = \Psi(L) \, v_t.$$

- The matrix Ψ_j gives the **impulse response** of Y at horizon j.

- Following the earlier example, the impulse response provides the effect on output j quarters after the government increases spending. It is directly of interest, and also can help to discipline models.

Computing impulse response functions

- $Y_t = \Psi(L) v_t = \Psi(L) A(L) Y_t$. Since this equality must hold for all Y_t:

$$
\begin{aligned}
I &= \Psi(L) A(L) \\
&= [\Psi_0 + \Psi_1 L + \ldots +] [A_0 - A_1 L - \ldots - A_p L^p] \\
&= \Psi_0 A_0 + [\Psi_1 A_0 - \Psi_0 A_1] L + [\Psi_2 A_0 - \Psi_1 A_1 - \Psi_0 A_2] L^2 \\
&+ \ldots + \left[\Psi_j A_0 - \sum_{i=1}^{j} \Psi_{j-i} A_i \right] L^j + \ldots + .
\end{aligned}
$$

- Taking the roots of this polynomial,

$$
\begin{aligned}
I &= \Psi_0 A_0, \\
\Psi_j A_0 &= \sum_{i=1}^{j} \Psi_{j-i} A_i \ \forall j > 0.
\end{aligned}
$$

Computing impulse response functions

-

$$I = \Psi_0 A_0,$$

$$\Psi_j A_0 = \sum_{i=1}^{j} \Psi_{j-i} A_i \ \forall j > 0.$$

- Recursive computation:

$$\Psi_0 = A_0^{-1},$$

$$\Psi_j = \left[\sum_{i=1}^{j} \Psi_{j-i} A_i \right] A_0^{-1} \text{ for } j > 0.$$

COMPUTING IMPULSE RESPONSE FUNCTIONS

- Reduced form impulse response found from above setting $B_0 = I$:

$$\tilde{\Psi}_0 = I,$$
$$\tilde{\Psi}_j = \left[\sum_{i=1}^{j} \tilde{\Psi}_{j-i} B_i\right] \text{ for } j > 0.$$

- Relationship:

Initial step: $\Psi_0 = A_0^{-1} = \tilde{\Psi}_0 A_0^{-1}$.

Conjecture: $\Psi_{j-1} = \tilde{\Psi}_{j-1} A_0^{-1}$.

Induction step:
$$\Psi_j = \left[\sum_{i=1}^{j} \Psi_{j-i} A_i\right] A_0^{-1} = \left[\sum_{i=1}^{j} \overbrace{\tilde{\Psi}_{j-i} A_0^{-1}}^{\Psi_{j-i}} \overbrace{A_0 B_i}^{A_i}\right] A_0^{-1}$$
$$= \left[\sum_{i=1}^{j} \tilde{\Psi}_{j-i} B_i\right] A_0^{-1} = \tilde{\Psi}_j A_0^{-1}.$$

- Often easiest to obtain $\tilde{\Psi}_j$ using B and then convert to Ψ_j using A_0^{-1}.

Impulse response intuition

- Let $dY_{t+j} = \Psi_j v = \tilde{\Psi}_j e$ be the average change in Y_{t+j} given $v_t = v, e_t = e$:

$$dY_{t+j} = E[Y_{t+j}|e_t = e, v_t = v, Y_{t-1}, \ldots] - E[Y_{t+j}|e_t = v_t = 0, Y_{t-1}, \ldots].$$

- Then:

$$\begin{array}{rlll}
 & \text{Reduced form} & & \text{Structural} \\
dY_t = & \tilde{\Psi}_0 e_t = e & & = \Psi_0 v = A_0^{-1} v = e \\
dY_{t+1} = & \tilde{\Psi}_1 e = B_1 e & & = \Psi_1 v = \Psi_0 A_1 A_0^{-1} v = B_1 A_0^{-1} v \\
dY_{t+2} = & \tilde{\Psi}_2 e = \left(\tilde{\Psi}_1 B_1 + B_2\right) e & & = \Psi_2 v = \left(\Psi_1 A_1 + \Psi_0 A_2\right) A_0^{-1} v \\
 & & & = \left(\tilde{\Psi}_1 B_1 + B_2\right) A_0^{-1} v \\
\vdots & & &
\end{array}$$

Forecast variance decomposition

- Example: what fraction of output variance 4 quarters ahead caused by monetary policy shocks?
- Let $e_{t+h|t} \equiv Y_{t+h} - E_t Y_{t+h}$ denote the h step ahead forecast error:

$$e_{t+h|t} = \Psi_0 v_{t+h} + \Psi_1 v_{t+h-1} + \ldots + \Psi_{h-1} v_{t+1} = \sum_{j=0}^{h-1} \Psi_j v_{t+h-j}.$$

- Forecast variance, where $\Psi_{j,rc}$ denotes the (r,c) entry of Ψ_j:

$$Var\left(e_{t+h|t}\right) = Var\left(\sum_{j=0}^{h-1} \Psi_j v_{t+h-j}\right) = \sum_{j=0}^{h-1} Var\left(\Psi_j v_{t+h-j}\right) = \sum_{j=0}^{h-1} \Psi_j \Sigma \Psi_j'$$

$$= \sum_{j=0}^{h-1} \begin{pmatrix} \sum_{m=1}^{k} \Psi_{j,1m}^2 \sigma_m^2 & \cdots & \sum_{m=1}^{k} \Psi_{j,1m} \Psi_{j,km} \sigma_m^2 \\ \vdots & \ddots & \vdots \\ \sum_{m=1}^{k} \Psi_{j,km} \Psi_{j,1m} \sigma_m^2 & \cdots & \sum_{m=1}^{k} \Psi_{j,km}^2 \sigma_m^2 \end{pmatrix}.$$

- Contribution of variable c to horizon h variance of variable r:

$$\frac{\sum_{j=0}^{h-1} \Psi_{j,rc}^2 \sigma_c^2}{\sum_{m=1}^{k} \sum_{j=0}^{h-1} \Psi_{j,rm} \sigma_m^2}.$$

Outline

1. ARMA processes
2. Filters
3. VARs
   - VAR overview
   - Structural VAR and Cholesky example
   - Impulse response and variance decomposition
   - Other structural point identification strategies
   - Identification concerns
   - Sign restrictions and set identification
4. Local projection
5. Principal Components Analysis

Block recursive (Keating, JEDC 1996)

- Suppose $R = A_0^{-1}$ has the structure:[1]

$$\begin{pmatrix} R_{11} & 0 \\ R_{21} & R_{22} \end{pmatrix} \begin{pmatrix} v_{1,t} \\ v_{2,t} \end{pmatrix} = \begin{pmatrix} e_{1,t} \\ e_{2,t} \end{pmatrix},$$

where R_{11} and R_{21} are unrestricted matrices, R_{22} is lower triangular.

- Then:

$$\begin{aligned} e_{1,t} &= R_{11} v_{1,t} \\ e_{2,t} &= R_{21} v_{1,t} + R_{22} v_{2,t} = R_{21} R_{11}^{-1} e_{1,t} + R_{22} v_{2,t}, \\ E\left[e_{2,t} | e_{1,t}\right] &= R_{21} R_{11}^{-1} e_{1,t}. \end{aligned}$$

[1] Note that A_0 is also block recursive with: $R = \begin{pmatrix} A_{0,11}^{-1} & 0 \\ -A_{0,22}^{-1} A_{0,21} A_{0,11}^{-1} & A_{0,22}^{-1} \end{pmatrix}$.

Block recursive (Keating, JEDC 1996)

- Let $C = chol(\Omega)$ and $\mu_t = C^{-1}e_t$:

$$\begin{pmatrix} C_{11} & 0 \\ C_{21} & C_{22} \end{pmatrix} \begin{pmatrix} \mu_{1,t} \\ \mu_{2,t} \end{pmatrix} = \begin{pmatrix} e_{1,t} \\ e_{2,t} \end{pmatrix}.$$

- Following the same steps as before: $E\left[e_{2,t}|e_{1,t}\right] = C_{21}C_{11}^{-1}e_{1,t}$. Then:

$$\begin{aligned} R_{21}R_{11}^{-1} &= C_{21}C_{11}^{-1}, \\ e_{2,t} &= C_{21}C_{11}^{-1}e_{1,t} + R_{22}v_{2,t}, \\ R_{22}v_{2,t} &= C_{22}\mu_{2,t}. \end{aligned}$$

- Then $R_{22}R_{22}' = E[R_{22}v_{2,t}v_{2,t}'R_{22}'] = E[C_{22}\mu_{2,t}\mu_{2,t}'C_{22}'] = C_{22}C_{22}'$. Since Cholesky factors are unique, $R_{22} = C_{22}$.
- Then $v_{2,t} = \mu_{2,t}$, i.e. $v_{2,t}$ is identified from Cholesky factor.

Non-recursive system direct estimation: GMM

- Suppose A_0 is not lower triangular but the order condition is satisfied, and let $R = A_0^{-1}$. Write:

$$R = \begin{pmatrix} r'_{1.} \\ \vdots \\ r'_{k.} \end{pmatrix}, \qquad \Omega = RR' = \begin{pmatrix} r'_{1.}r_{1.} & r'_{1.}r_{2.} & \dots & r'_{1.}r_{k.} \\ r'_{2.}r_{1.} & r'_{2.}r_{2.} & \dots & r'_{2.}r_{k.} \\ \vdots & \vdots & \ddots & \dots \\ r'_{k.}r_{1.} & r'_{k.}r_{2.} & \dots & r'_{k.}r_{k.} \end{pmatrix}.$$

- This gives $\frac{k(k+1)}{2}$ moment conditions $E\left[vech\left(\hat{\Omega}\right) - vech\left(RR'\right)\right] = 0$, where $vech(X)$ is the column vector that stacks the entries on or below the main diagonal of matrix X.

- Additional restrictions from imposing zeros or other known coefficients.

Non-recursive system direct estimation: MLE

If $v_t \sim N(0, I)$, then LLH is:[2]

$$
\begin{aligned}
\mathscr{L}(\Omega|\hat{e}_1, \ldots, \hat{e}_T) &= -\sum_{t=1}^{T}\left[\frac{k}{2}\ln 2\pi + \frac{1}{2}\ln|\Omega| + \frac{1}{2}e_t'\Omega^{-1}e_t\right] \\
&= -\frac{Tk}{2}\ln 2\pi - \frac{T}{2}\ln|\Omega| - \frac{1}{2}\sum_{t=1}^{T}e_t'\Omega^{-1}e_t \\
&= -\frac{Tk}{2}\ln 2\pi - \frac{T}{2}\ln\left|A_0^{-1}A_0^{-1\prime}\right| - \frac{T}{2}\left[tr\left(A_0'A_0\hat{\Omega}\right)\right] \\
&= -\frac{Tk}{2}\ln 2\pi + T\ln(|A_0|) - \frac{T}{2}\left[tr\left(A_0'A_0\hat{\Omega}\right)\right].
\end{aligned}
$$

Maximize $\ln(|A_0|) - \frac{1}{2}\left[tr\left(A_0'A_0\hat{\Omega}\right)\right]$ over the free parameters in A_0.

---

[2] Third line: $\sum_{t=1}^{T} e_t'\Omega^{-1}e_t = \sum_{t=1}^{T} tr\left(e_t'\Omega^{-1}e_t\right) = \sum_{t=1}^{T} tr\left(\Omega^{-1}e_te_t'\right) = tr\left(\Omega^{-1}\sum_{t=1}^{T}(e_te_t')\right) = T\left[tr\left(\Omega^{-1}\hat{\Omega}\right)\right] = T\left[tr\left(A_0'A_0\hat{\Omega}\right)\right]$.
Fourth line: $\det(AB) = \det(A)\det(B)$, $\det(A') = \det(A)\det\left(A^{-1}\right) = \det(A)^{-1} \Rightarrow$
$\ln\left|A_0^{-1}A_0^{-1\prime}\right| = \ln\left|A_0^{-1}A_0^{-1}\right| = \ln\left(\left|A_0^{-1}\right|^2\right) = \ln\left(|A_0|^{-2}\right) = -2\ln(|A_0|)$.

Blanchard-Quah (1989) long-run restrictions

- Output growth Δx_t, unemployment u_t, demand shock v_t^d, supply shock v_t^s:

$$\text{Structural:} \quad \begin{pmatrix} \Delta x_t \\ u_t \end{pmatrix} = \begin{pmatrix} \Psi_{11}(L) & \Psi_{12}(L) \\ \Psi_{21}(L) & \Psi_{22}(L) \end{pmatrix} \begin{pmatrix} v_t^d \\ v_t^s \end{pmatrix},$$

$$\text{Reduced form:} \quad \begin{pmatrix} \Delta x_t \\ u_t \end{pmatrix} = \begin{pmatrix} \tilde{\Psi}_{11}(L) & \tilde{\Psi}_{12}(L) \\ \tilde{\Psi}_{21}(L) & \tilde{\Psi}_{22}(L) \end{pmatrix} \begin{pmatrix} e_{1t} \\ e_{2t} \end{pmatrix}.$$

- Relationship of residuals:

$$e_t = \begin{pmatrix} e_{1t} \\ e_{2t} \end{pmatrix} = \begin{pmatrix} \Delta x_t \\ u_t \end{pmatrix} - E_{t-1} \begin{pmatrix} \Delta x_t \\ u_t \end{pmatrix} = \begin{pmatrix} \Psi_{11}(0) & \Psi_{12}(0) \\ \Psi_{21}(0) & \Psi_{22}(0) \end{pmatrix} \begin{pmatrix} v_t^d \\ v_t^s \end{pmatrix} = R v_t.$$

- Normalization: $\Sigma = Var(v_t) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$.
- Observable moments:

$$Var(e_t) = RR' = \begin{pmatrix} r_{11}^2 + r_{12}^2 & r_{11} r_{21} + r_{12} r_{22} \\ r_{11} r_{21} + r_{12} r_{22} & r_{21}^2 + r_{22}^2 \end{pmatrix}.$$

- Three moments and four unknowns in R. So far just the order restriction.
- Assumption: only supply shocks v_t^s can have long run effect on level of GDP.

Blanchard-Quah (1989) long-run restrictions

- Step 1: solve for reduced-form MA representation of $\begin{pmatrix}\Delta x_t & u_t\end{pmatrix}'$:

$$
\begin{aligned}
\begin{pmatrix}\Delta x_t \\ u_t\end{pmatrix} &= \sum_{j=1}^{p} B_j \begin{pmatrix}\Delta x_{t-j} \\ u_{t-j}\end{pmatrix} + e_t = \left(I - \sum_{j=1}^{p} B_j L^j\right)^{-1} e_t \\
&= \begin{pmatrix} 1 - \sum_{j=1}^{p} B_{j,11} L^j & -\sum_{j=1}^{p} B_{j,12} L^j \\ -\sum_{j=1}^{p} B_{j,21} L^j & 1 - \sum_{j=1}^{p} B_{j,22} L^j \end{pmatrix}^{-1} e_t \\
&= \frac{1}{\det\left(I - \sum_{j=1}^{p} B_j L^j\right)} \begin{pmatrix} 1 - \sum_{j=1}^{p} B_{j,22} L^j & \sum_{j=1}^{p} B_{j,12} L^j \\ \sum_{j=1}^{p} B_{j,21} L^j & 1 - \sum_{j=1}^{p} B_{j,11} L^j \end{pmatrix} e_t .
\end{aligned}
$$

Blanchard-Quah (1989) long-run restrictions

- Step 1: solve for reduced-form MA representation of $\begin{pmatrix}\Delta x_t & u_t\end{pmatrix}'$:

$$\begin{pmatrix}\Delta x_t \\ u_t\end{pmatrix} = \sum_{j=1}^{p} B_j \begin{pmatrix}\Delta x_{t-j} \\ u_{t-j}\end{pmatrix} + e_t = \left(I - \sum_{j=1}^{p} B_j L^j\right)^{-1} e_t$$

$$= \begin{pmatrix} 1-\sum_{j=1}^{p} B_{j,11} L^j & -\sum_{j=1}^{p} B_{j,12} L^j \\ -\sum_{j=1}^{p} B_{j,21} L^j & 1-\sum_{j=1}^{p} B_{j,22} L^j \end{pmatrix}^{-1} e_t$$

$$= \frac{1}{\det\left(I-\sum_{j=1}^{p} B_j L^j\right)} \begin{pmatrix} 1-\sum_{j=1}^{p} B_{j,22} L^j & \sum_{j=1}^{p} B_{j,12} L^j \\ \sum_{j=1}^{p} B_{j,21} L^j & 1-\sum_{j=1}^{p} B_{j,11} L^j \end{pmatrix} e_t.$$

- Step 2: Substitute structural shocks:

$$\Delta x_t = \frac{\left(1-\sum_{j=1}^{p} B_{j,22} L^j\right) e_{1t} + \sum_{j=1}^{p} B_{j,12} L^j e_{2t}}{\det\left(I-\sum_{j=1}^{p} B_j L^j\right)}$$

$$= \frac{\left(1-\sum_{j=1}^{p} B_{j,22} L^j\right)\left(r_{11} v_t^d + r_{12} v_t^s\right) + \sum_{j=1}^{p} B_{j,12} L^j \left(r_{21} v_t^d + r_{22} v_t^s\right)}{\det\left(I-\sum_{j=1}^{p} B_j L^j\right)}.$$

Blanchard-Quah (1989) long-run restrictions

- Step 3: Impose restriction as $\sum_{h=0}^{\infty} \Psi_{11}(h) = 0$:

$$
\begin{aligned}
0 &= \sum_{h=0}^{\infty} \Psi_{11}(h) \\
&= \left(1 - \sum_{j=1}^{p} B_{j,22} L^j\right) r_{11} + \sum_{j=1}^{p} B_{j,12} L^j r_{21} \\
&= \left(1 - \sum_{j=1}^{p} B_{j,22}\right) r_{11} + \sum_{j=1}^{p} B_{j,12} r_{21}.
\end{aligned}
$$

Blanchard-Quah (1989) long-run restrictions

- Step 3: Impose restriction as $\sum_{h=0}^{\infty} \Psi_{11}(h) = 0$:

$$
\begin{aligned}
0 &= \sum_{h=0}^{\infty} \Psi_{11}(h) \\
&= \left(1 - \sum_{j=1}^{p} B_{j,22} L^{j}\right) r_{11} + \sum_{j=1}^{p} B_{j,12} L^{j} r_{21} \\
&= \left(1 - \sum_{j=1}^{p} B_{j,22}\right) r_{11} + \sum_{j=1}^{p} B_{j,12} r_{21}.
\end{aligned}
$$

- Step 4: use four moment conditions to solve for four elements of matrix R.

Blanchard-Quah (1989) long-run restrictions

- Step 3: Impose restriction as $\sum_{h=0}^{\infty} \Psi_{11}(h) = 0$:

$$
\begin{aligned}
0 &= \sum_{h=0}^{\infty} \Psi_{11}(h) \\
&= \left(1 - \sum_{j=1}^{p} B_{j,22} L^j\right) r_{11} + \sum_{j=1}^{p} B_{j,12} L^j r_{21} \\
&= \left(1 - \sum_{j=1}^{p} B_{j,22}\right) r_{11} + \sum_{j=1}^{p} B_{j,12} r_{21}.
\end{aligned}
$$

- Step 4: use four moment conditions to solve for four elements of matrix R.

- In practice, can be sensitive to number of lags and some demand shocks may be very persistent.

Blanchard-Quah and estimates of potential output

- Many government forecasting agencies provide estimates of potential output and output gap (Federa; Reserve, CBO, IMF, etc.)
- Guide whether policy should be loose or tight.
- Used by academics e.g. to estimate Phillips curve.
- Following slides from Coibon, Gorodnichenko, Ulate (BPEA 2018).

Figure 3. Real-Time Estimates of U.S. Potential Output Growth Rate and Trends in Actual Output Growth Rate[a]

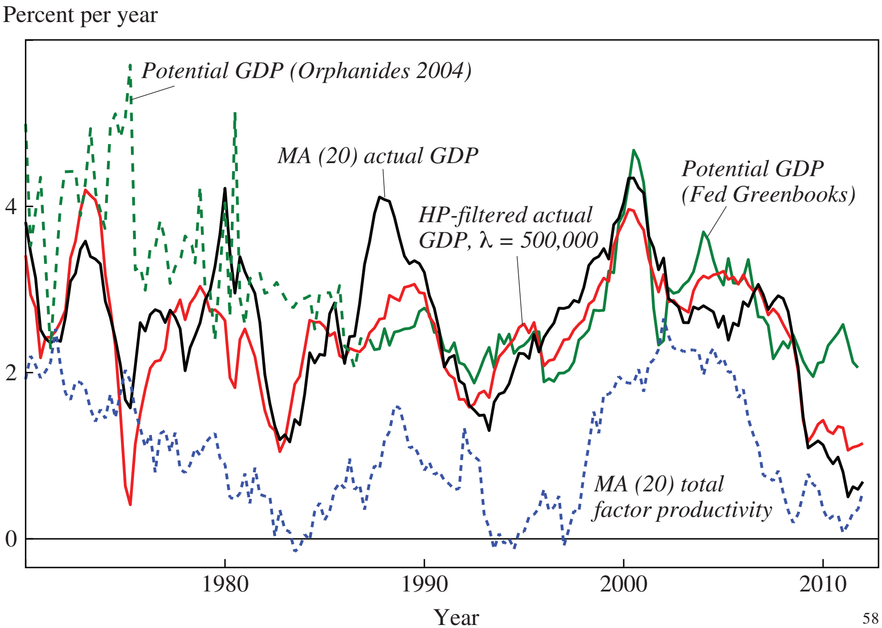
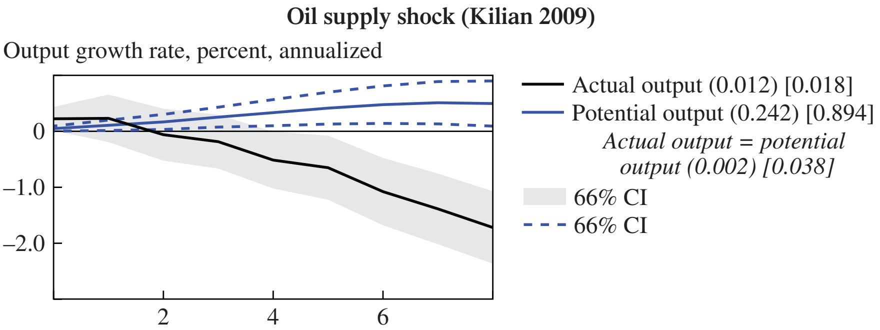


Figure 4. Responses of U.S. Output and Greenbook Estimates of Potential



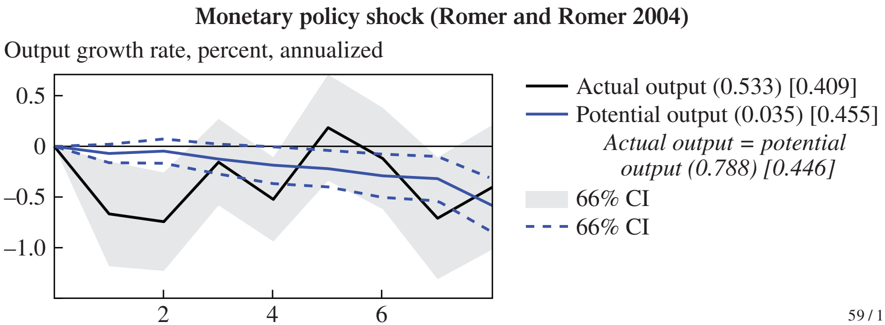
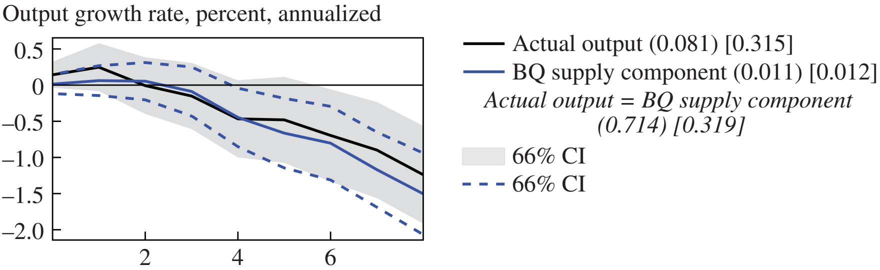
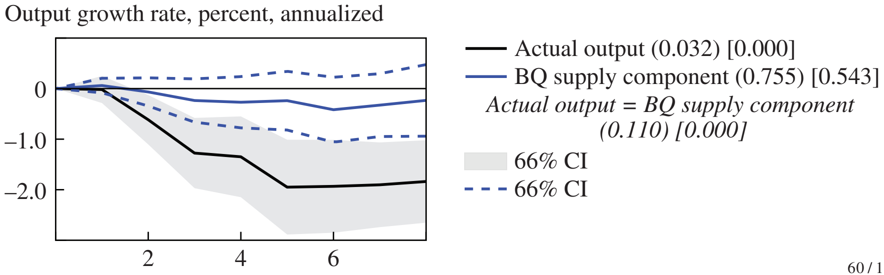


Figure 12. Response of the U.S. Growth Rate for Actual Output and the SVAR Identified

Oil supply shock (Kilian 2009)



Monetary policy shock (Romer and Romer 2004)



Heteroskedastic identification (Lutkepohl (2012))

- Suppose structural variance matrix is I in first t_1 periods and Σ_{t_2} in next t_2 periods with all non-unity diagonal elements. Then:

$$\Omega_{t_1} = A_0^{-1} A_0^{-1\prime},$$
$$\Omega_{t_2} = A_0^{-1} \Sigma_{t^2} A_0^{-1\prime}.$$

- $\Omega_{t_1} \neq \Omega_{t_2}$ provides $\frac{k(k+1)}{2} + \frac{k(k+1)}{2}$ restrictions, exactly enough to identify k^2 parameters in A_0^{-1} and k variances in Σ_{t_2}.
- Advantage is less structure, disadvantage is opacity of identification and interpretation of structural shocks.
- Lewis (RESTUD, forthcoming) generalizes to arbitrary time-varying volatility.

External instruments

- Measure shock or proxy for shock external to VAR.
- Contrast to "internal instruments" that are restrictions on residuals estimated in the VAR.
- Example: part of tax changes taken for reasons unrelated to business cycle trajectory.
- More next time.

Outline

Granger causality

$$y_{1,t} = c_1 + \Phi_{1,1}(L)y_{1,t-1} + \Phi_{1,2}(L)y_{2,t-1} + e_{1,t}$$
$$y_{2,t} = c_2 + \Phi_{2,1}(L)y_{1,t-1} + \Phi_{2,2}(L)y_{2,t-1} + e_{2,t}.$$

- $y_{1,t}, y_{2,t}$ are univariate processes.
- y_2 does not **Granger cause** y_1 if $\Phi_{1,2} = 0$.
- Interpret: y_2 Granger causes y_1 if it helps to forecast y_1 beyond what past values of y_1 would predict.

Granger causality and VARs

- VAR restrictions generate time series of structural shocks.
- If Cholesky, by construction shocks not Granger caused by variables in the VAR because they are linear combinations of OLS residuals.
- Can test whether variables excluded from the VAR Granger cause shock series.
  - If yes, suggests agents in the world have more information than does the VAR. If agents' response to this information affects variables in the VAR, there is an omitted variables problem.
  - Formally, invertibility fails: VAR variables do not span space of shocks.
- Absence of Granger causality of structural shocks not sufficient to prove proper identification.

Example: Ramey (QJE 2011)

- Structural residuals from VAR in government spending and output a la Blanchard and Perotti (more in next lecture).
- Simple test: do forecasts of government spending or war dates Granger cause VAR residuals?

TABLE I
Granger Causality Tests

Hypothesis tests	p-value in parenthesis
Do war dates Granger-cause VAR shocks? 1948:1–2008:4	Yes (0.012)
Do one-quarter ahead professional forecasts Granger-cause VAR shocks? 1981:3–2008:4	Yes (0.032)
Do four-quarter ahead professional forecasts Granger-cause VAR shocks? 1981:3–2008:4	Yes (0.016)
Do VAR shocks Granger-cause war dates? 1948:1–2008:4	No (0.115)

Notes. VAR shocks were estimated by regressing the log of real per capita government spending on 4 lags of itself, the Barro–Redlick tax rate, log real per capita GDP, log rea per capita nondurable plus services consumption, log real per capita private fixed investment, log real per capita total hours worked, and log compensation in private business divide by the deflator for private business. Except for the professional forecasts, 4 lags were also used in the Granger-causality tests. For the professional forecaster test, the VAR shock i period t is regressed on either the forecast made in period t-1 of the growth rate of real federal spending from t-1 to t for the forecast made in period t-4 of the growth from t-4 to

- What other variables might one want to test?

Solving omitted variable problem

- Suppose you find an omitted variable. What then?
- Can add to VAR with appropriate ordering (this is what Ramey does).
- If original VAR system included p lags and k variables, adds pk coefficients to be estimated.
- This is VAR "curse of dimensionality."
- FAVAR as one possible solution: next lecture.

Outline

Sign restrictions overview (my notation)

- Recall short run restrictions:

$$
\begin{aligned}
A_0 Y_t &= A_1 Y_{t-1} + \ldots A_p Y_{t-p} + v_t,\ \ Var(v_t) = I, \\
Y_t &= B_1 Y_{t-1} + \ldots B_p Y_{t-p} + e_t,\ \ Var(e_t) = \Omega, \\
e_t &= R v_t.
\end{aligned}
$$

- Conditional on the estimated reduced form coefficient matrices, $R = A_0^{-1}$ uniquely defines the impulse responses to structural shocks.
- Researchers will sometimes justify restrictions by appealing to "reasonableness" of the impulse response functions.
- Sign restrictions formalize this and take it to the logical limit: a priori define what a resonable impulse response function looks like, then identify set of R matrices that give rise to such impulse responses.

Sign restrictions details

- Specify only enough restrictions to recover the shocks of interest.
- Brute force.
  - Estimate $B(L)$ and Ω with OLS. For every possible R, compute IRFs and verify whether restrictions are satisfied. Set identifies IRFs.
- Penalty function.
  - Assign score to each admissable draw.
  - A_0s that generate strong comovement viewed as more likely to be truth. Justification is combination of other shocks could "accidentally" result in weak comovement of required sign, but actual shock should generate stronger comovement.
  - Generates point estimate and standard error bands.
- Bayesian estimation.
  - Uhlig argues conceptually cleaner: draws from posterior treated as candidate truths, and either satisfy sign restrictions or not.

Bayesian approach

- Uhlig (JME 2005), Rubio-Ramirez, Waggoner, Zha (RESTUD 2010).
- If $e_t = Rv_t$ (R is instantaneous IRF), and C is cholesky factor of $\Omega = Var(e_t)$, then $R = CQ$, where Q is orthonormal matrix:

$$\Omega = CC' = RR' \Rightarrow \quad I = C^{-1}RR'C'^{-1} \Rightarrow \quad C^{-1}R = Q \Rightarrow \quad R = CQ.$$

- Specify joint prior on $(B(L), \Omega, Q)$ with $(B(L), \Omega)$ independent of Q.
- Take joint draw from posterior on $(B(L), \Omega)$ and uniform (HAAR) distribution for Q. Construct IRF and keep if satisfy sign restrictions.
- Note: for uninformative prior on $B(L), \Omega$, posterior draw is random draw from asymptotic distribution of OLS estimates.
- Note: data are informative about $(B(L), \Omega)$ but not about Q. In large sample, *should be* similar to brute force approach.

Typical algorithm

1. Estimate reduced form VAR $B(L)Y_t = e_t$ and form $\hat{\Omega}_t = T^{-1}\sum_{t=1}^{T} \hat{e}_t \hat{e}_t'$.
2. Draw M random values $\Omega^{(m)}, B^{(m)}(L)$ from asymptotic distribution of OLS estimates $\hat{\Omega}_t, \hat{B}(L)$. Define $C^{(m)}$ as Cholesky factor of $\Omega^{(m)}$.
3. Draw orthonormal matrix $Q^{(m)}$ from HAAR prior (more soon) and define $R^{(m)} = C^{(m)}Q^{(m)}$.
4. Check sign restrictions for structural impulse responses given R^m.
5. Plot distribution (median, 16% upper and lower values) of retained draws.

Mountford and Uhlig (JOE 2009)

- VAR in GDP, private consumption, total government expenditure, total government revenue, real wages, private non-residential investment, interest rate, adjusted reserves, the producer price index for crude materials and the GDP deflator.
- Quarterly frequency, 1955-2000, six lags.
- Combines sign restrictions with recursive identification:
  - First identify general business cycle shock to satisfy comovement.
  - Monetary shock has required signs and is orthogonal to business cycle shock.
  - Goverment spending and revenue shocks orthogonal to monetary and business cycle shocks.

Table I. Identifying sign restrictions

	Gov. revenue	Gov. spending	GDP, cons, non-res.inv.	Interest rate	Adjusted reserves	Prices
Non-fiscal shocks						
Business cycle	$+$		$+$			
Monetary policy				$+$	$-$	$-$
Basic fiscal policy shocks						
Government revenue	$+$					
Government spending		$+$				

This table shows the sign restrictions on the impulse responses for each identified shock. 'Cons' stands for private consumption and 'Non-res. inv.' stands for non-residential investment. A '$+$' means that the impulse response of the variable in question is restricted to be positive for four quarters following the shock, including the quarter of impact. Likewise, a '$-$' indicates a negative response. A blank entry indicates that no restrictions have been imposed.

restriction that government revenues increase with output in the business cycle shock should be emphasized. This is our crucial identifying assumption for fiscal policy shocks: when output and government revenues move in the same direction, we essentially assume that this must be due to some improvement in the business cycle generating the increase in government revenue, not the other way around. We regard this is as a reasonable assumption and consistent with a number of theoretical views. Furthermore, our identifying assumptions are close to minimal: some assumptions are needed to say anything at all. The orthogonality assumption a priori excludes the

Baumeister and Hamilton (ECMA 2015) critique

- Attraction of sign restrictions is agnostic over R.
- But uniform prior on one parameter in R does not imply uniform prior over other parameters.
- 2 variable example: $e_{1t} = r_{11}v_{1t} + r_{12}v_{2t}$. Require $r_{11} > 0$.
  - $Var(e_{1t}) = \omega_1^2 = r_{11}^2 + r_{12}^2 \Rightarrow r_{11} \leq \omega_1,\ r_{12} = \sqrt{\omega_1^2 - r_{11}^2}$.
  - Uniform prior: $r_{11}|\Omega \sim U(0, \omega_1)$. What is distribution ofr_{12}?
  - $0.1 = Pr\left(r_{11} < 0.1\omega_1\right) = Pr\left(r_{12} > \sqrt{\omega_1^2 - (0.1)^2\omega_1^2}\right) = Pr\left(r_{12} > \sqrt{0.99}\omega_1\right)$. Density of r_{12} sharply nonuniform.
  - Can show other parameters (e.g. r_{21}) also nonuniform.
- "Researchers using the traditional methodology can end up performing hundreds of thousands of calculations, ostensibly analyzing the data, but in the end are doing nothing more than generating draws from a prior distribution that they never even acknowledged that they had assumed."

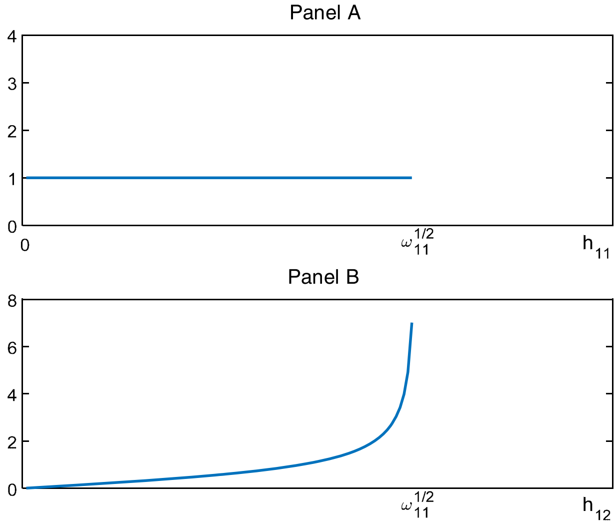


FIGURE 3.—Prior densities when a uniform prior is used for the effect of shock 1 and the number of variables is 2. Panel A: Prior density for response of variable 1 to structural shock 1. Panel B: Prior density for response of variable 1 to structural shock 2.

the results would be to calculate the effect on output if the Fed raised interest rates by 25 basis points, namely h_{21}/h_{11}. A uniform prior on h_{11} and h_{21} implies

Sign restrictions summary

- Within the class of admissable IRFs, data are not informative about which is correct $\Rightarrow$ plot only bounds!.
- Can do inference on identified set, such as minimum and maximum IRFs (Gafarov,Meier,Olea, JOE 2018).
  - Unfortunately, just sign restrictions appear not to restrict the identified set much (Giacomini,Kitagawa, ECMA forthcoming).
  - "Masquerading shocks": even if restrictions uniquely identify a single shock, combinations of other shocks may satisfy restrictions. But these other shock combinations may have very different implications for other variables (Wolf, AEJ:Macro, 2020).
- Plagborg-Møller (QE 2019) offers flexible approach to identification from impulse response functions.
  - "Correct" way to impose uninformed prior given sign restriction.
  - Can add other restrictions on smoothness, maximum response, etc.
- Embrace Bayesian justification and treat sign restrictions as Bayesian prior (next time).

Outline

Overview

- VARs combine identification of structural shocks and construction of impulse response functions into a single step.
- Recursive structure of VAR impulse response functions can compound errors at long horizions if VAR is mis-specified.
- Alternative approach: estimate reduced form impulse matrices directly.
- Seminal reference is Jorda (AER 2005), but article contains strange notation and some unfortunate typos.

Frisch-Waugh-Lovell statement

- Model:
$$y = X\beta + u = X_1\beta_1 + X_2\beta_2 + u.$$

- $y \sim (Tx1), X_1 \sim (TxK_1), \beta_1 \sim (K_1x1), X_2 \sim (Tx1), \beta_2 \sim (1x1), X = \begin{pmatrix} X_1 & X_2 \end{pmatrix} \sim (T \times K), u \sim (Tx1)$.

- Claim: $\hat{\beta}_2$ from regressing y on X_1 and X_2, and $\tilde{\beta}_2$ from regressing y on the residuals from a regression of X_2 on X_1, are the same.

Frisch-Waugh-Lovell notation

- Some notation:

$$\begin{aligned}
P_X &= X(X'X)^{-1}X' \sim (K \times K), \\
M_X &= I - P_X \sim (K \times K), \\
P_1 &= X_1(X_1'X_1)^{-1}X_1' \sim (K_1 \times K_1), \\
M_1 &= I - P_1 \sim (K_1 \times K_1).
\end{aligned}$$

- P is a "hat" matrix: $P_X y = X\hat{\beta} = \hat{y}$.
- M is a residual matrix: $M_X y = y - X\hat{\beta} = \hat{u}$.
- P, M are projection matrices, and hence symmetric and idempotent.
- $P_X X = X, M_X X = 0$.

Frisch-Waugh-Lovell proof

The residuals from a regression of X_2 on X_1 are $M_1 X_2$. Thus

$$\begin{aligned}
\tilde{\beta}_2 &= \left((M_1X_2)' M_1 X_2\right)^{-1} (M_1 X_2)' y \\
&= \left(X_2' M_1 X_2\right)^{-1} X_2' M_1 y \\
&= \left(X_2' M_1 X_2\right)^{-1} X_2' M_1 (P_X + I - P_X) y \\
&= \left(X_2' M_1 X_2\right)^{-1} X_2' M_1 (P_X y + M_X y) \\
&= \left(X_2' M_1 X_2\right)^{-1} X_2' M_1 \left(X_1 \hat{\beta}_1 + X_2 \hat{\beta}_2 + M_X y\right) \\
&= \hat{\beta}_2.
\end{aligned}$$

The last line follows from $M_1 X_1 = 0$, and

$$\begin{aligned}
X_2' M_1 M_X &= X_2' (I - P_1) M_X \\
&= X_2' (M_X - P_1 M_X) \\
&= X_2' \left(M_X - (M_X P_1)'\right) \\
&= X_2' M_X \\
&= 0.
\end{aligned}$$

Local projection

- Definition of IR at horizon h to reduced form disturbance e at time t:

$$\tilde{\Psi}_h = E\left[Y_{t+h}|e_t = e; Y_t, \dots, Y_{t-p}\right] - E\left[Y_{t+h}|e_t = 0; Y_t, \dots, Y_{t-p}\right].$$

- Consider the direct regression

$$Y_{t+h} = \alpha_h + C_{h,0} Y_t + C_{h,1} Y_{t-1} + \dots + C_{h,p} Y_{t-p} + u_{t+h}. \qquad (1)$$

- By FWL, can also recovery $C_{h,0}$ from the two step procedure:

$$Y_t = \hat{F}_0 + \hat{F}_1 Y_{t-1} + \dots + \hat{F}_p Y_{t-p} + \hat{e}_t, \qquad (2)$$

$$Y_{t+h} = C_{h,0} \hat{e}_t. \qquad (3)$$

- (2) is a standard VAR.
- From (3), $C_{h,0} = \tilde{\Psi}_h$ can be estimated directly using (1).

Implementation

$$Y_{t+h} = \alpha_h + C_{h,0} Y_t + C_{h,1} Y_{t-1} + \cdots + C_{h,p} Y_{t-p} + u_{t+h}$$

- **Local projection** because IR estimated independently for each horizon.
- Lag length p can vary across horizons.
- To obtain structural IRFs, either
  - Multiply local IR by VAR estimate of A_0^{-1}, or
  - If Cholesky ordering, replace $C_{h,0} Y_t$ with $C_{h,0} \begin{pmatrix} y_{1t} & 0 & \ldots & 0 \end{pmatrix}'$ to construct response to Cholesky shock directly.

Equivalence result to VARs

- VAR uses iterative forecast for IRF while LP uses direct projection.
- LP(∞) and VAR(∞) give exactly the same IRFs (Plagborg-Moller and Wolf, ECMA forthcoming).
- Intuition: VAR(∞) perfectly captures all covariance properties of the data, so VAR iterated forecasts perfectly coincide with local projection direct forecasts.
- LP(p) and VAR(p) give approximately the same IRF up to horizon p in finite samples.
  - Approximate because horizon h LP depends on first $p+h$ autocovariances of data, while VAR(p) captures only first p autocovariances.
  - Exact if LP regressor is a "shock" that is uncorrelated with past data (so can run LP with 0 lags).

Comparison to VARs

- Any structural SVAR identification implementable in LP.
- In practice, sometimes easier to implement LP:
  - Confidence bands straightforward, including HAR standard errors.
  - Easily extend to nonlinear settings.
- LP IRFs can look "weird" since they don't impose smoothness of VAR step-ahead forecasts. Can "tune" optimal smoothness (Plagborg-Moller, 2019).
- More generally, bias-variance trade-off in small samples.
- Other advantages from flexibility of LP: for example, obtain IRFs for variables not in shock information set – avoid curse of dimensionality.
- More flexibility is good, but beware of the Sims critique.

Example: state-dependent fiscal multipliers

- Question: are government spending multipliers larger in recessions than expansions?
- Auerbach and Gorodnichenko (AEJ Policy 2012) trivariate VAR, government spending ordered first:

$$
\begin{aligned}
Y_t &= \{\text{Government spending}_t, \text{Taxes}_t, \text{Output}_t\}', \\
Y_t &= [1 - F(z_{t-1})]\Pi_E(L)Y_{t-1} + F(z_{t-1})\Pi_R(L)Y_{t-1} + u_t, \\
F(z_t) &= \frac{\exp(-\gamma z_t)}{1 + \exp(-\gamma z_t)}.
\end{aligned}
$$

- z_t = deviation of output growth from trend is an index of the business cycle.
- Paper includes controls for forecasts of government spending and output – more in a few weeks.
- Highly nonlinear system: VAR estimated using MCMC methods.

Example: state-dependent fiscal multipliers

- Auerbach and Gorodnichenko (2013) instead use local projection.
- For any response variable x_t, they estimate (essentially):

$$\begin{aligned} x_{t+h} = \alpha_h &+ F(z_t)\Pi_{R,h}(L)y_{t-1} + [1-F(z_{t-1})]\Pi_{E,h}(L)y_{t-1} \\ &+ F(z_t)\Phi_{R,h}(L)g_{t-1} + [1-F(z_{t-1})]\Phi_{E,h}(L)g_{t-1} \\ &+ F(z_t)\Psi_{R,h}g_t + [1-F(z_{t-1})]\Psi_{E,h}g_t. \end{aligned}$$

- Shock is equivalent to a Cholesky shock.
- Easy to examine response of variables such as private consumption, hours, etc. without overwhelming number of variables in the VAR.
- After appropriately interacting variables, this is a linear regression.

Figure 1. Comparison of impulse responses from VAR and direct projection

Panel A: Full sample, 1960-2010.

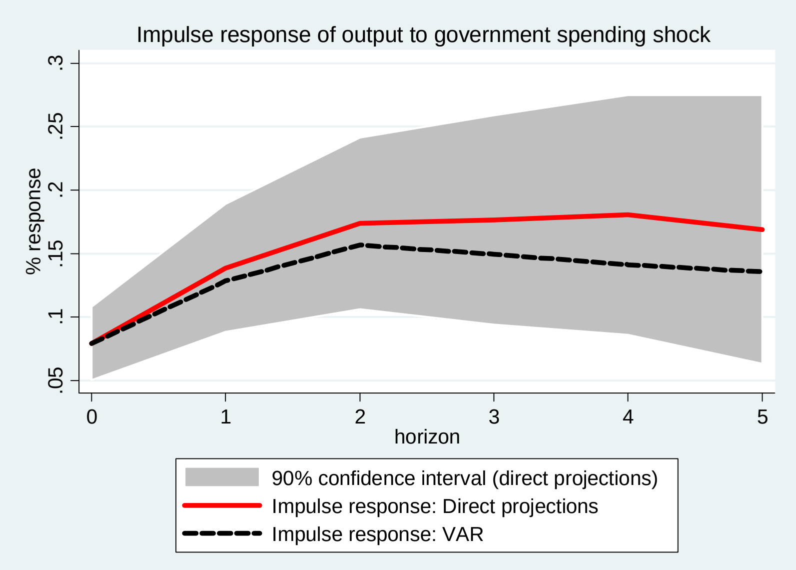
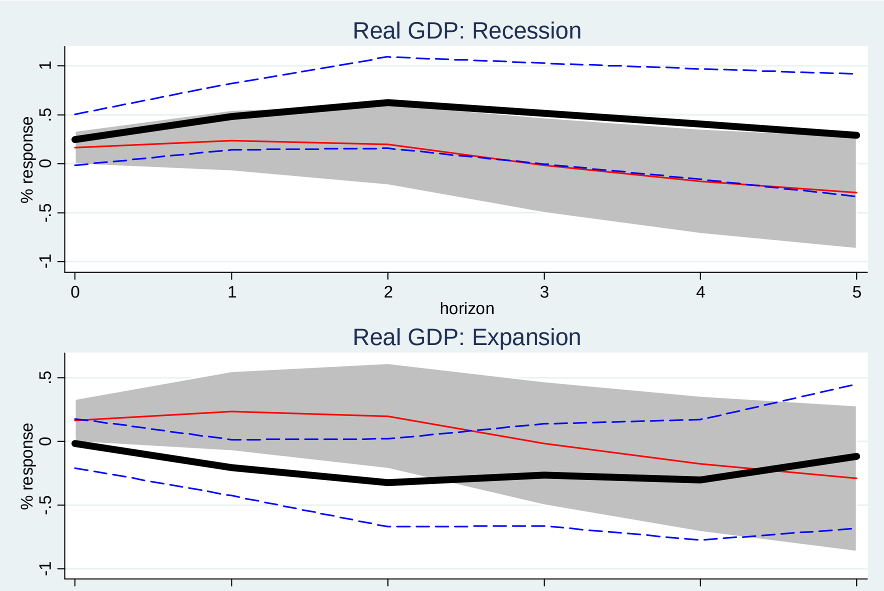
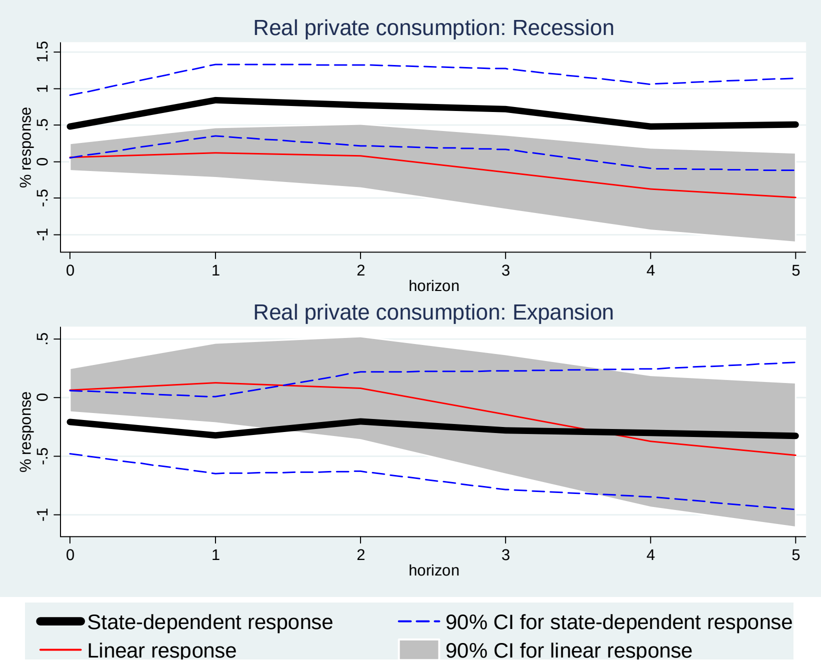


Figure 4. State-dependent vs. Linear responses

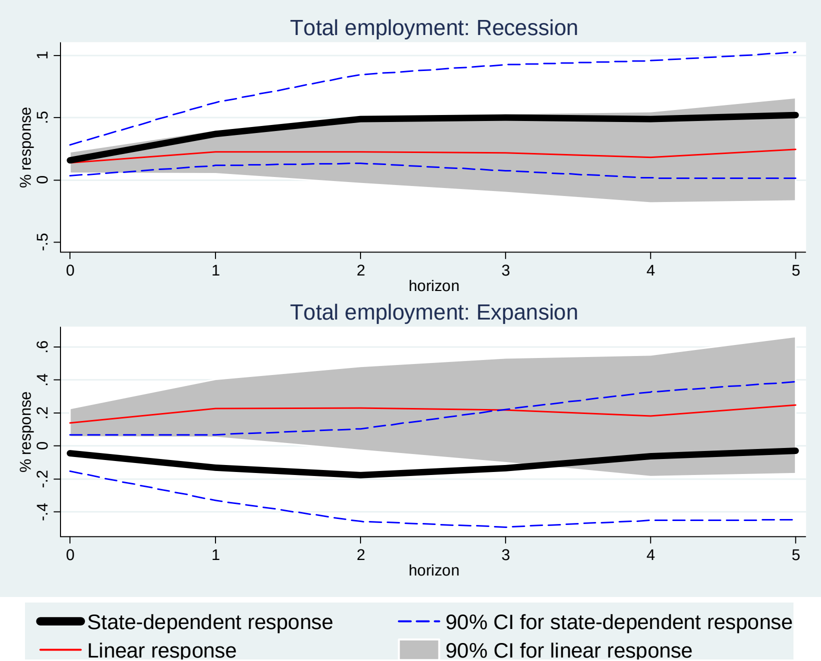
Panel A. Real GDP



Panel B. Private consumption



Panel D. Total employment



LP WITHOUT VARs

- Jorda specification:
  $Y_{t+h} = \alpha_h + C_{h,0} Y_t + C_{h,1} Y_{t-1} + \cdots + C_{h,p} Y_{t-p} + u_{t+h}$.

- Suppose instead: $Y_{t+h} = \alpha_h + C_{h,0} Y_{1,t}$.

- Valid only if $Y_{1,t}$ is a structural shock.

- Close cousin: excluded instrument for $Y_{1,t}$. LP-IV most common use in practice.

- Return to it when we discuss excluded instruments.

Outline

Overview

- Data reduction technique to extract common component of series.
- Many models in finance and macro have "factor-loading" structure.
- PCA does not generate economic interpretation of factors.
  - Compare to observed factors such as Fama-French.
- PCA does efficiently condense information from large number of series into a few factors.
- Sometimes possible to interpret principal component factors.
  - Example: term structure as level, slope, curvature.
- Used in FAVAR (next time).

Setup

- $e_t \equiv \begin{pmatrix} e_{1,t} & e_{2,t} & \dots & e_{S,t} \end{pmatrix}'$: vector of observations at time t on S variables. WLOG, assume the mean of each element of e_t is zero.
- Factor model (observation t):

$$\underbrace{e_t}_{SxK1} = \underbrace{\Lambda}_{SxK} \underbrace{F_t}_{Kx1} + \underbrace{\varepsilon_t}_{Sx1},$$

  - Λ: matrix of factor-loadings.
  - F_t: vector of K factors.
- Λ and F_t are identified only up to a rotation, i.e. $\Lambda F_t = \Lambda Q' Q F_t$ for an orthonormal matrix Q.
  - Impose a normalization such as $Var[F_t] = I$.
- Factor model (full system):

$$\underbrace{e}_{TxS} = \underbrace{F}_{TxK} \Lambda' + \underbrace{\varepsilon}_{TxS}.$$

Observation t

$$\begin{pmatrix} e_{1,t} \\ e_{2,t} \\ \vdots \\ e_{S,t} \end{pmatrix} = \begin{pmatrix} \lambda_{1,1} & \lambda_{1,2} & \dots & \lambda_{1,K} \\ \lambda_{2,1} & \lambda_{2,2} & \vdots & \lambda_{2,K} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_{S,1} & \lambda_{S,2} & \dots & \lambda_{S,K} \end{pmatrix} \begin{pmatrix} f_{1,t} \\ f_{2,t} \\ \vdots \\ f_{K,t} \end{pmatrix} + \begin{pmatrix} \varepsilon_{1t} \\ \varepsilon_{2,t} \\ \vdots \\ \varepsilon_{S,t} \end{pmatrix}.$$

Full system

$$
\begin{pmatrix} e_1' \\ e_2' \\ \vdots \\ e_T' \end{pmatrix} = \begin{pmatrix} e_{1,1} & e_{2,1} & \dots & e_{S,1} \\ e_{1,2} & e_{2,2} & \vdots & e_{S,2} \\ \vdots & \vdots & \ddots & \vdots \\ e_{1,T} & e_{2,T} & \dots & e_{S,T} \end{pmatrix}
$$

$$
= \begin{pmatrix} f_{1,1} & f_{2,1} & \dots & f_{K,1} \\ f_{1,2} & f_{2,2} & \vdots & f_{K,2} \\ \vdots & \vdots & \ddots & \vdots \\ f_{1,T} & f_{2,T} & \dots & f_{K,T} \end{pmatrix} \begin{pmatrix} \lambda_{1,1} & \lambda_{1,2} & \dots & \lambda_{1,K} \\ \lambda_{2,1} & \lambda_{2,2} & \vdots & \lambda_{2,K} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_{S,1} & \lambda_{S,2} & \dots & \lambda_{S,K} \end{pmatrix}'
$$

$$
+ \begin{pmatrix} \varepsilon_{1,1} & \varepsilon_{2,1} & \dots & \varepsilon_{S,1} \\ \varepsilon_{1,2} & \varepsilon_{2,2} & \vdots & \varepsilon_{S,2} \\ \vdots & \vdots & \ddots & \vdots \\ \varepsilon_{1,T} & \varepsilon_{2,T} & \dots & \varepsilon_{S,T} \end{pmatrix}.
$$

First PC

- First PC maximizes explained variance in e_t from a single factor.
- With one factor,
$$e = F_1 \lambda_1' + \varepsilon.$$
- Variance:
$$TVar[e] = e'e = \lambda_1 F_1' F_1 \lambda_1' + \varepsilon' \varepsilon.$$
- Letting $tr\{.\}$ denote the trace function, the minimization problem is
$$\min_{F_1, \lambda_1} \sum_{s=1}^{S} \sum_{t=1}^{T} \varepsilon_{s,t}^2 = \min tr\{\varepsilon' \varepsilon\}.$$

Minimization

$$\min_{F_1,\lambda_1} \sum_{s=1}^{S} \sum_{t=1}^{T} \varepsilon_{s,t}^2 = \min tr\left\{\varepsilon'\varepsilon\right\}.$$

- Some algebra:

$$\begin{aligned} tr\left\{\varepsilon'\varepsilon\right\} &= tr\left\{\left(e - F_1\lambda_1'\right)'\left(e - F_1\lambda_1'\right)\right\} \\ &= tr\left\{e'e - \lambda_1 F_1' e - e' F_1 \lambda_1' + \lambda_1 F_1' F_1 \lambda_1'\right\} \\ &= tr\left\{e'e - 2\lambda_1 F_1' e + \lambda_1' \lambda_1\right\}. \end{aligned}$$

  - Third equality uses the fact that trace passes under a sum, $tr\{AB\} = tr\{BA\}$, and the normalization $F'F = I$.
- FOC w.r.t λ_1':

$$-2e'F_1 + 2\lambda_1 = 0 \Longrightarrow e'F_1 = \lambda_1.$$

  - Use: $\frac{\partial tr\{AZB\}}{\partial Z} = \frac{\partial tr\left\{B'Z'A'\right\}}{\partial Z} = BA$, $\frac{\partial tr\left\{AZBZ'C\right\}}{\partial Z} = BZ'CA + B'Z'A'C'$.

...CONTINUED

- Use FOC $e'F_1 = \lambda_1$ to concentrate the objective function:

$$\begin{aligned} tr\left\{\varepsilon'\varepsilon\right\} &= tr\left\{e'e - 2\lambda_1 F_1' e + \lambda_1'\lambda_1\right\} \\ &= tr\left\{e'e - 2e'F_1F_1'e + F_1'ee'F_1\right\} \\ &= tr\left\{e'e - F_1'ee'F_1\right\}. \end{aligned}$$

- Drop the first term (which does not depend on F_1) and restate problem as:

$$\max_{F_1} tr\left\{F_1'ee'F_1\right\} \; s.t. \; F_1'F_1 = I.$$

- Let d_1 denote the lagrange multiplier on the constraint. The FOC w.r.t F_1' is

$$\begin{aligned} 0 &= ee'F_1 + ee'F_1 - 2d_1F_1 \\ &= \left(ee' - d_1 I\right) F_1. \end{aligned}$$

First PC solution

- FOC:

$$\left(ee' - d_1 I\right) F_1 = 0.$$

- F_1 is an eigenvector of ee' and d_1 is its eigenvalue.
- Linear algebra fact: the largest eigenvalue maximizes the quadratic form $fee'f$ over all unit vectors f.
- Premultiplying FOC by e':

$$0 = \left(e'ee' - d_1 Ie'\right) F_1 = \left(e'e - d_1 I\right) e'F_1 = \left(e'e - d_1 I\right) \lambda_1.$$

- λ_1 is the eigenvector of $e'e$ corresponding to the same eigenvalue as F_1.
- In sum, the vector of the first principal components is the eigenvector corresponding to the largest eigenvalue of of ee', and the eigenvector corresponding to the same eigenvalue of $e'e$ provides the loadings.

Additional PCs

- Suppose j principal components have been identified:

$$e = F_j \Lambda_j' + \varepsilon.$$

  - $\Lambda_j = \begin{pmatrix} \lambda_1 & \lambda_2 & \dots & \lambda_j \end{pmatrix}$; $F_j = \begin{pmatrix} f_1 & f_2 & \dots & f_j \end{pmatrix}$.
- The $j+1$'th PC maximizes the explained variance remaining after the first j PCs, i.e. it chooses $\{f_{j+1}, \lambda_{j+1}\}$ to minimize $\tilde{\varepsilon}'\tilde{\varepsilon}$, where [3]

$$\begin{aligned} \tilde{e} &= e - F_j \Lambda_j' \\ &= f_{j+1}\lambda_{j+1}' + \tilde{\varepsilon}. \end{aligned}$$

- Following the same logic as above, the $j+1$ PC will be the eigenvector associated with the largest eigenvalue of $\tilde{e}\tilde{e}'$, and the $j+1$ factor loading will be the eigenvector of $\tilde{e}'\tilde{e}$ associated with the same eigenvalue.

[3] Note that $\tilde{e}'\tilde{e} = e'e - e'F_j\Lambda_j' - \Lambda_j F_j' e + \Lambda_j F_j' F_j \Lambda_j' = e'e - \Lambda_j F_j' F_j \Lambda_j'$.

Additional PCs, more directly

Claim: The largest eigenvalue of $\tilde{e}'\tilde{e}$ is also the $j+1$'th largest eigenvalue of $e'e$, and the corresponding eigenvector is also the eigenvector corresponding to the same eigenvalue of $e'e$.

Proof: By induction.

1. Claim holds trivially for first eigenvalue and eigenvector.
2. Suppose it holds for the first j eigenvalues and eigenvectors. Let $\left(d_{j+1}, \lambda_{j+1}\right)$ denote the $j+1$'th largest eigenvalue and associated eigenvector of $e'e$ and ee':
$$\left(e'e - d_{j+1}I\right)\lambda_{j+1} = 0.$$
3. λ_{j+1} is orthogonal to each column of Λ_j by the induction assumption (the columns of Λ_j contain eigenvectors corresponding to distinct eigenvalues of $e'e$). In particular,
$$\Lambda_j \Lambda_j' \lambda_{j+1} = 0.$$
4. Subtracting the last equation from the penultimate:
$$\begin{aligned} 0 &= \left(e'e - \Lambda_j \Lambda_j' - d_{j+1}I\right)\lambda_{j+1} \\ &= \left(\tilde{e}'\tilde{e} - d_{j+1}I\right)\lambda_{j+1}. \end{aligned}$$

Singular value decomposition

- Singular value decomposition of e:

$$e = FD^{\frac{1}{2}}\Lambda'.$$

  - F is orthonormal matrix of factors.
  - Λ is orthonormal matrix of loadings.
  - D is diagonal matrix of factor variances.

- Confirm:

$$e'e = \Lambda D^{\frac{1}{2}}F'FD^{\frac{1}{2}}\Lambda' = \Lambda D\Lambda'.$$
$$ee' = FD^{\frac{1}{2}}\Lambda'\Lambda D^{\frac{1}{2}}F' = FDF'.$$

- Intuition: S linearly independent factors span the subspace of e, so with full set of factors, error term disappears.

Selecting number of factors

- Informal criteria using scree plots of eigenvalues or threshold for explained variance.
- Bai and Ng (ECMA 2002) criteria:

$$IC_{p1}(k) = \log V(k, \hat{F}^k) + k\left(\frac{S+T}{ST}\right)\log\left(\frac{ST}{S+T}\right)$$
$$IC_{p2}(k) = \log V(k, \hat{F}^k) + k\left(\frac{S+T}{ST}\right)\log\left(C_{ST}^2\right)$$
$$IC_{p3}(k) = \log V(k, \hat{F}^k) + kC_{ST}^{-2}\log\left(C_{ST}^2\right).$$

  - $V(k, \hat{F}^k) = \min_{\Lambda}(ST)^{-1}\sum_{s=1}^{S}\sum_{t=1}^{T}\left(e_{s,t} - \hat{\lambda}_s(k)'\hat{F}_t(k)\right)^2$ measures the goodness of fit with k factors, and
  - $C_{ST} = \min\left(\sqrt{T}, \sqrt{S}\right)$ captures the speed of convergence.

Comments

- Matrix decomposition very fast to compute.
- Can do PCA on covariance or correlation matrix, depending whether you want to upweight data series with larger variances. If covariance matrix, make sure series have comparable units.
- Always pay attention to what the default normalization is in statistical software.
- If SVD, check ordering of factors and explained shares.
- Often only a few factors will explain more than 80 or 90 percent of the variance.
- When factors or factors and loadings are to be included in subsequent model, one can instead directly estimate in one step. See Bai (ECMA 2009) for recent treatment.