Dynamic Econometric Methods for Big Data in Real Time

Francis X. Diebold
University of Pennsylvania

November 22, 2016

A Bit of Introduction

"Big Data" or "Vast Data" in Econometrics
"Tall", "Wide", and "Dense"

Consider a $(T \times K)$ regression "X matrix" for T "days" (or whatever) of data each of K variables. Now imagine sampling intra-day data as well, m times per day. Then X is $(mT \times K)$. Big data correspond to huge-X situations arising because one or more of T, K, and m is huge.

– As $K \to \infty$ we speak of "wide data" (in reference to the wide X matrix due to the large number of regressors)

– As $T \to \infty$ we speak of "tall data" (in reference to the tall X matrix, due to the large number of time periods, i.e., the long calendar span of data)

– As $m \to \infty$ we speak of "dense data" (in reference to the high-frequency intra-day sampling, regardless of whether the data are tall.)

Examples

– Consider 2500 days of 1-minute returns (360 per six-hour trading day) for each of 5000 stocks. $K = 5000$, $T = 2500$, $m = 360$, so X is $(1,800,000 \times 5000)$. Data are tall, wide and dense.

– Consider 10 days of 1-minute returns (360 per six-hour trading day) for each of 20 stocks. $K = 20$, $T = 10$, $m = 360$, so X is (3600×20). Data are dense, but neither tall nor wide.

– Consider 2500 days of daily returns for each of 5000 stocks. $K = 5000$, $T = 2500$, $m = 1$, so X is (2500×5000). Data are tall and wide, but not dense.

What's New With Tall, Wide, and Dense Data

- Tall: Nothing new (must wait a million years...)
  - Well, actually, climatological studies...
- Wide: It's here now!
  - High-dimensional VAR's in many contexts
  - Web data like Billion Prices Project
  - Novel apps like fixed effects without panels, forecast combining weights, etc.
- Dense: It's here now!
  - Volatility estimation from ultra-high frequency financial market data
  - Long-memory estimation from ultra-high frequency financial market data?

Long Memory and Dense Data

Dense data may let us estimate long memory with accurately.

- Fractionally-integrated processes are self-similar. "scaling laws"
- $I(d)$ at any observational frequency is $I(d)$ at any other.
- So all we need is very fine sampling (i.e., dense data)? Here dense data deliver information on a low-frequency phenomenon, even with a very short calendar span...

Mixed Frequencies and Big Data

Here the action is in both wide data and dense data.

Mixed frequencies $\Leftrightarrow$ Big Data

- $\Rightarrow$ Mixed frequencies naturally lead to Big (dense) Data. The state-space system must be written at the highest observed frequency, so if the highest frquency is dense, so too will be the entire system.

- $\Leftarrow$ Big (wide) Data naturally lead to mixed frequencies. The wider the dataset, the more likely it is to contain mixed frequencies.

Unbalanced Panels and Big Data

Big Data panels are likely unbalanced.

- This is obvious if mixed frequencies are operative
- Also there may be entry and exit from the "panel"

Big Data Often Have a Real-Time Vintage-Data Aspect

- Due to revisions
- Due to entry / exit
- Makes recursive analysis impossible
- Finally, a real role for "out-of-sample" model comparisons

Background

The Wold Decomposition

Under regularity conditions,
every covariance-stationary process $\{y_t\}$ can be written as:

$$y_t = \sum_{i=0}^{\infty} b_i \varepsilon_{t-i}$$

where:

$$b_0 = 1$$

$$\sum_{i=0}^{\infty} b_i^2 < \infty$$

$$\varepsilon_t = [y_t - P(y_t | y_{t-1}, y_{t-2}, ...)] \sim WN(0, \sigma^2)$$

The General Linear Process

$$y_t = B(L)\varepsilon_t = \sum_{i=0}^{\infty} b_i \varepsilon_{t-i}$$

$$\varepsilon_t \sim WN(0, \sigma^2)$$

$$b_0 = 1$$

$$\sum_{i=0}^{\infty} b_i^2 < \infty$$

Unconditional Moment Structure of the LRCSSP (Assuming Strong WN Innovations)

$$E(y_t) = E\left(\sum_{i=0}^{\infty} b_i \varepsilon_{t-i}\right) = \sum_{i=0}^{\infty} b_i E\varepsilon_{t-i} = \sum_{i=0}^{\infty} b_i \cdot 0 = 0$$

$$var(y_t) = var\left(\sum_{i=0}^{\infty} b_i \varepsilon_{t-i}\right) = \sum_{i=0}^{\infty} b_i^2 var(\varepsilon_{t-i}) = \sigma^2 \sum_{i=0}^{\infty} b_i^2$$

Conditional Moment Structure (Assuming Strong WN Innovations)

$$E(y_t|\Omega_{t-1}) = E(\varepsilon_t|\Omega_{t-1}) + b_1 E(\varepsilon_{t-1}|\Omega_{t-1}) + b_2 E(\varepsilon_{t-2}|\Omega_{t-1}) + \ldots$$

$$(\Omega_{t-1} = \varepsilon_{t-1}, \varepsilon_{t-2}, \ldots)$$

$$= 0 + b_1\varepsilon_{t-1} + b_2\varepsilon_{t-2} + \ldots = \sum_{i=1}^{\infty} b_i\varepsilon_{t-i}$$

$$var(y_t|\Omega_{t-1}) = E[(y_t - E(y_t|\Omega_{t-1}))^2|\Omega_{t-1}]$$

$$= E(\varepsilon_t^2|\Omega_{t-1}) = E(\varepsilon_t^2) = \sigma^2$$

Autocovariance Structure

$$\gamma(\tau) = E\left[\left(\sum_{i=-\infty}^{\infty} b_i \varepsilon_{t-i}\right)\left(\sum_{h=-\infty}^{\infty} b_h \varepsilon_{t-\tau-h}\right)\right]$$

$$= \sigma^2 \sum_{i=-\infty}^{\infty} b_i b_{i-\tau}$$

(where $b_i \equiv 0$ if $i < 0$)

Approximating the Wold Representation

$MA(q)$ process
(Obvious truncation)

$AR(p)$ process
(Stochastic difference equation)

$ARMA(p, q)$ process
("Rational distributed lag,"
later rational spectrum, links to state space)

Unconditional moment structure
Conditional moment structure
Autocovariance functions
Stationarity and invertibility conditions

Wiener-Kolmogorov Prediction

$$y_t = \varepsilon_t + b_1 \varepsilon_{t-1} + \ldots$$

$$y_{T+h} = \varepsilon_{T+h} + b_1 \varepsilon_{T+h-1} + \ldots + b_h \varepsilon_T + b_{h+1} \varepsilon_{T-1} + \ldots$$

Project on $\Omega_T = \{\varepsilon_T, \varepsilon_{T-1}, \ldots\}$ to get:

$$y_{T+h,T} = b_h \varepsilon_T + b_{h+1} \varepsilon_{T-1} + \ldots$$

Note that the projection is on the *infinite* past

Wiener-Kolmogorov Prediction Error

$$e_{T+h,T} = y_{T+h} - y_{T+h,T} = \sum_{i=0}^{h-1} b_i \varepsilon_{T+h-i}$$

(An $MA(h-1)$ process!)

$$E(e_{T+h,T}) = 0$$

$$var(e_{T+h,T}) = \sigma^2 \sum_{i=0}^{h-1} b_i^2$$

Wold's Chain Rule for Autoregressions

Consider an AR(1) process:

$$y_t = \phi y_{t-1} + \varepsilon_t$$

History:

$$\{y_t\}_{t=1}^{T}$$

Immediately,

$$y_{T+1,T} = \phi y_T$$
$$y_{T+2,T} = \phi y_{T+1,T} = \phi^2 y_T$$
$$\vdots$$
$$y_{T+h,T} = \phi y_{T+h-1,T} = \phi^h y_T$$

Extension to $AR(p)$ and $AR(\infty)$ is immediate.

Multivariate

$(y_{1t},\ y_{2t})'$ is covariance stationary if:

$$E(y_{1t}) = \mu_1 \ \forall\ t$$
$$E(y_{2t}) = \mu_2 \ \forall\ t$$

$$\Gamma_{y_1 y_2}(t,\ \tau) = E \left(\begin{array}{c} y_{1t} - \mu_1 \\ y_{2t} - \mu_2 \end{array} \right) (y_{1,t-\tau} - \mu_1,\ y_{2,t-\tau} - \mu_2)$$

$$= \left(\begin{array}{cc} \gamma_{11}(\tau) & \gamma_{12}(\tau) \\ \gamma_{21}(\tau) & \gamma_{22}(\tau) \end{array} \right)$$

$$\tau = 0,\ 1,\ 2,\ ...$$

Cross Covariances

$$\gamma_{12}(\tau) \neq \gamma_{12}(-\tau)$$

$$\gamma_{12}(\tau) = \gamma_{21}(-\tau)$$

$$\Gamma_{y_1 y_2}(\tau) = \Gamma'_{y_1 y_2}(-\tau), \quad \tau = 0, 1, 2, \ldots$$

The Multivariate General Linear Process

$$\begin{pmatrix} y_{1t} \\ y_{2t} \end{pmatrix} = \begin{pmatrix} B_{11}(L) & B_{12}(L) \\ B_{21}(L) & B_{22}(L) \end{pmatrix} \begin{pmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{pmatrix}$$

$$y_t = B(L)\varepsilon_t = (I + B_1 L + B_2 L^2 + ...)\varepsilon_t$$

$$E(\varepsilon_t \varepsilon_s') = \begin{cases} \Sigma & if \ t = s \\ 0 & otherwise \end{cases}$$

$$\sum_{i=0}^{\infty} \| B_i \|^2 < \infty$$

Autocovariance Structure

$$\Gamma_{y_1 y_2}(\tau) = \sum_{i=-\infty}^{\infty} B_i \, \Sigma \, B'_{i-\tau}$$

(where $B_i \equiv 0$ if $i < 0$)

Wiener-Kolmogorov Prediction

$$y_t = \varepsilon_t + B_1\varepsilon_{t-1} + B_2\varepsilon_{t-2} + ...$$

$$y_{T+h} = \varepsilon_{T+h} + B_1\varepsilon_{T+h-1} + B_2\varepsilon_{T+h-2} + ...$$

Project on $\Omega_t = \{\varepsilon_T, \varepsilon_{T-1}, ...\}$ to get:

$$y_{t+h,T} = B_h\varepsilon_T + B_{h+1}\varepsilon_{T-1} + ...$$

Wiener-Kolmogorov Prediction Error

$$\varepsilon_{T+h,T} = y_{T+h} - y_{T+h,T} = \sum_{i=0}^{h-1} B_i \varepsilon_{T+h-i}$$

$$E[\varepsilon_{T+h,T}] = 0$$

$$E[\varepsilon_{T+h,T}\varepsilon'_{T+h,T}] = \sum_{i=0}^{h-1} B_i \Sigma B_i'$$

Vector Autoregressions (VAR's)

N-variable VAR of order p:

$$\Phi(L)y_t = \varepsilon_t$$

$$\varepsilon_t \sim WN(0, \Sigma)$$

where:

$$\Phi(L) = I - \Phi_1 L - ... - \Phi_p L^p$$

A 2-Variable VAR(1) in "Long Form"

$$\begin{pmatrix} y_{1t} \\ y_{2t} \end{pmatrix} = \begin{pmatrix} \phi_{11} & \phi_{12} \\ \phi_{21} & \phi_{22} \end{pmatrix} \begin{pmatrix} y_{1t-1} \\ y_{2t-1} \end{pmatrix} + \begin{pmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{pmatrix}$$

$$\begin{pmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{pmatrix} \sim WN\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix} \right)$$

– Two sources of cross-variable interaction.
What are they?

Understanding VAR's: Bivariate Granger-Sims Causality (One Kind of Predictive Connectedness)

Is the history of y_j useful for predicting y_i,
over and above the history of y_i?

– Granger non-causality tests: Simple exclusion restrictions

– In the simple 2-Variable VAR(1) example,

$$\begin{pmatrix} y_{1t} \\ y_{2t} \end{pmatrix} = \begin{pmatrix} \phi_{11} & \phi_{12} \\ \phi_{21} & \phi_{22} \end{pmatrix} \begin{pmatrix} y_{1t-1} \\ y_{2t-1} \end{pmatrix} + \begin{pmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{pmatrix},$$

y_2 does not Granger cause y_1 iff $\phi_{12} = 0$

– Natural extensions for $N > 2$
(always testing exclusion restrictions)

Understanding VAR's: MA Representation

$$\Phi(L)y_t = \varepsilon_t$$

$$y_t = \Phi^{-1}(L)\varepsilon_t = \Theta(L)\varepsilon_t$$

where:

$$\Theta(L) = I + \Theta_1 L + \Theta_2 L^2 + ...$$

Long-Form MA Representation of 2-Variable VAR(1)

$$\left(\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - \begin{pmatrix} \phi_{11} & \phi_{12} \\ \phi_{21} & \phi_{22} \end{pmatrix} L \right) \begin{pmatrix} y_{1t} \\ y_{2t} \end{pmatrix} = \begin{pmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{pmatrix}$$

$$\begin{pmatrix} y_{1t} \\ y_{2t} \end{pmatrix} = \begin{pmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{pmatrix} + \begin{pmatrix} \theta_{11}^1 & \theta_{12}^1 \\ \theta_{21}^1 & \theta_{22}^1 \end{pmatrix} \begin{pmatrix} \varepsilon_{1t-1} \\ \varepsilon_{2t-1} \end{pmatrix} + \ldots$$

Understanding VAR's: Impulse Response Functions (IRF's) (Another Kind of Predictive Connectedness)

$$(I - \Phi_1 L - ... - \Phi_p L^p) y_t = \varepsilon_t$$

$$\varepsilon_t \sim WN(0, \Sigma)$$

The impulse-response question:
How is y_{it} dynamically affected by a shock to y_{jt} (alone)?

($N \times N$ matrix of IRF *graphs* (over steps ahead))

Problem:
Σ generally not diagonal, so how to shock j alone?

Understanding VAR's: Variance Decompositions (VD's) (Another Kind of Predictive Connectedness)

$$(I - \Phi_1 L - ... - \Phi_p L^p) y_t = \varepsilon_t$$

$$\varepsilon_t \sim WN(0, \Sigma)$$

The variance decomposition question:
How much of the h-step ahead (optimal) prediction-error variance of y_i is due to shocks to variable j?

($N \times N$ matrix of VD *graphs* (over h) could be done.
Or pick an h and examines the $N \times N$ matrix of VD numbers.)

Problem:
Σ generally not diagonal, which makes things tricky, as the variance of a sum of innovations is therefore not the sum of the variances.

Orthogonalizing VAR's by Cholesky Factorization (The Classic Identification Scheme)

Original:

$$(I - \Phi_1 L - ... - \Phi_p L^p) y_t = \varepsilon_t, \quad \varepsilon_t \sim WN(0, \Sigma)$$

Equivalently:

$$(I - \Phi_1 L - ... - \Phi_p L^p) y_t = P v_t, \quad v_t \sim WN(0, I)$$

where $\Sigma = PP'$, for lower-triangular P
(Cholesky factorization)

Now we can shock j alone (for IRF's)

Now we can proceed to calculate forecast-error variances without worrying about covariance terms (for VD's)
But there's no free lunch. Why?

IRF's and VD's from the Orthogonalized VAR

IRF comes from the
orthogonalized moving-average representation:

$$y_t = (I + \Theta_1 L + \Theta_2 L^2 + ...)\ P\, v_t$$

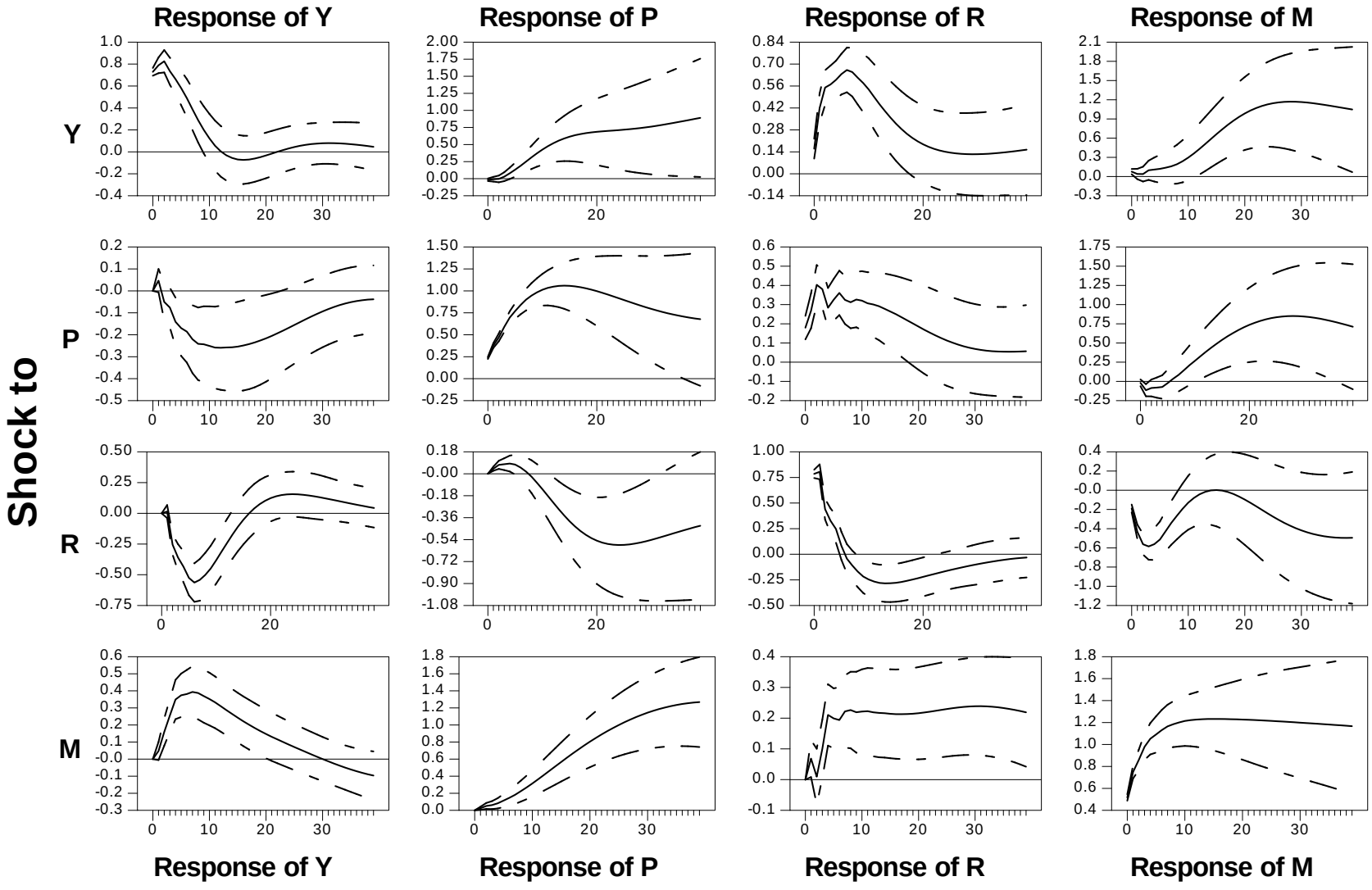
$$= (P + \Theta_1 P\ L + \Theta_2 P\ L^2 + ...)\ v_t$$

$$\text{IRF}_{\text{ij}} \text{ is } \{P_{ij},\ (\Theta_1 P)_{ij},\ (\Theta_2 P)_{ij}, ...\}$$

VD_{ij} comes similarly from the
orthogonalized moving-average representation.

Note how the the contemporaneous IRF and VD for $h = 1$ are
driven by the Cholesky choice of P.
Other choices are possible.

Graphic: IRF Matrix for 4-Variable U.S. Macro VAR



Two-Variable IRF Example (IRF_{12})

$$y_t = Pv_t + \Theta_1 Pv_{t-1} + \Theta_2 Pv_{t-2} + ...$$

$$v_t \sim \ WN(0, I)$$

$$y_t = C_0 v_t + C_1 v_{t-1} + C_2 v_{t-2} + ... \ \ (\mathrm{Q}: \ \mathrm{What\ is}\ C_{12}^0?)$$

$$\begin{pmatrix} y_{1t} \\ y_{2t} \end{pmatrix} = \begin{pmatrix} c_{11}^0 & c_{12}^0 \\ c_{21}^0 & c_{22}^0 \end{pmatrix} \begin{pmatrix} v_{1t} \\ v_{2t} \end{pmatrix} + \begin{pmatrix} c_{11}^1 & c_{12}^1 \\ c_{21}^1 & c_{22}^1 \end{pmatrix} \begin{pmatrix} v_{1t-1} \\ v_{2t-1} \end{pmatrix} + ...$$

$$\mathrm{IRF}_{12} = \mathrm{C}_{12}^0, \mathrm{C}_{12}^1, \mathrm{C}_{12}^2, ...$$

Two-Variable VD Example ($\mathrm{VD}_{12}(2)$)

$$\epsilon_{t+2,t} = C_0 v_{t+2} + C_1 v_{t+1}$$

$$v_t \sim WN(0, I)$$

$$\begin{pmatrix} \epsilon^1_{t+2,t} \\ \epsilon^2_{t+2,t} \end{pmatrix} = \begin{pmatrix} c^0_{11} & c^0_{12} \\ c^0_{21} & c^0_{22} \end{pmatrix} \begin{pmatrix} v_{1t+2} \\ v_{2t+2} \end{pmatrix} + \begin{pmatrix} c^1_{11} & c^1_{12} \\ c^1_{21} & c^1_{22} \end{pmatrix} \begin{pmatrix} v_{1t+1} \\ v_{2t+1} \end{pmatrix}$$

$$\epsilon^1_{t+2,t} = c^0_{11} v_{1t+2} + c^0_{12} v_{2t+2} + c^1_{11} v_{1t+1} + c^1_{12} v_{2t+1}$$

$$var(\epsilon^1_{t+2,t}) = (c^0_{11})^2 + (c^0_{12})^2 + (c^1_{11})^2 + (c^1_{12})^2$$

Part coming from v_2: $(c^0_{12})^2 + (c^1_{12})^2$

$$VD_{12}(2) = \frac{(c^0_{12})^2 + (c^1_{12})^2}{(c^0_{11})^2 + (c^0_{12})^2 + (c^1_{11})^2 + (c^1_{12})^2}$$

Orthogonalizing/Identifying VAR's More Generally "Structural VAR's"

Structure:

$$A_0 y_t = A_1 y_{t-1} + ... + A_p y_{t-p} + v_t, \quad v_t \sim (0, D)$$

where D is diagonal.

Reduced form:

$$\begin{aligned} y_t &= A_0^{-1} A_1 y_{t-1} + ... + A_0^{-1} A_p y_{t-p} + A_0^{-1} v_t \\ &= \Phi_1 y_{t-1} + ... + \Phi_p y_{t-p} + e_t, \end{aligned}$$

where $e_t = A_0^{-1} v_t$.

The structure can be identified from the reduced form if $\frac{N^2-N}{2}$ restrictions are imposed on A_0. One possibility if to impose that A_0 be lower triangular ("recursive structure").

IRF:

$$\begin{aligned} y_t &= (I + \Theta_1 L + \Theta_2 L^2 + ...) e_t \\ &= (I + \Theta_1 L + \Theta_2 L^2 + ...) A_0^{-1} v_t \\ &= (A_0^{-1} + \Theta_1 A_0^{-1} L + \Theta_2 A_0^{-1} L^2 + ...) v_t \end{aligned}$$

Hence recursive SEM's are linked to Cholesky-identified VAR's.

Problem: What to do When N is Huge? (Econometrics Traditionally has $N << T$)

– Estimation?

– Identification?

– Understanding?

Markovian Structure, State Space, and the Kalman Filter

Part I: Markov Processes

Discrete-State, Discrete-Time Stochastic Process

$$\{X_t\}, t = 0, 1, 2, \ldots$$

Possible values ("states") of X_t: $1, 2, 3, \ldots$

First-order homogeneous Markov process:

$$Prob(X_{t+1} = j | X_t = i, X_{t-1} = i_{t-1}, \ldots, X_0 = i_0)$$

$$= Prob(X_{t+1} = j | X_t = i) = p_{ij}$$

Transition Probability Matrix P

1-step transition probabilities:

$$P \equiv \begin{array}{c} \\ [time\ t] \end{array} \begin{array}{c} [time\,(t+1)] \\ \begin{pmatrix} p_{11} & p_{12} & \cdots \\ p_{21} & p_{22} & \cdots \\ \cdot & \cdot & \cdots \\ \cdot & \cdot & \cdots \\ & & \cdot \end{pmatrix} \end{array}$$

$$p_{ij} \geq 0, \quad \sum_{j=1}^{\infty} p_{ij} = 1$$

Chapman-Kolmogorov

m-step transition probabilities:

$$p_{ij}^{(m)} = Prob(X_{t+m} = j \mid \ X_t = i)$$

$$\text{Let } P^{(m)} \equiv \left(p_{ij}^{(m)}\right).$$

Chapman-Kolmogorov theorem:

$$P^{(m+n)} = P^{(m)} P^{(n)}$$

Corollary: $P^{(m)} = P^m$

Illustrations/Variations/Extensions: Network Connectedness

– 1-step network adjacency matrix A similar to Markov transition-probability matrix

– k-step network adjacency matrix A^k in precise parallel to Chapman-Kolmogorov

Illustrations/Variations/Extensions: Regime-Switching Models

$$P = \begin{pmatrix} p_{11} & 1 - p_{11} \\ 1 - p_{22} & p_{22} \end{pmatrix}$$

$$s_t \sim P$$

$$y_t = c_{s_t} + \phi_{s_t} y_{t-1} + \varepsilon_t$$

$$\varepsilon_t \sim iid\ N(0, \sigma^2_{s_t})$$

"Markov switching," or "hidden Markov," model

Illustrations/Variations/Extensions: Constructing Markov Processes with Useful Stationary Distributions

- Markov Chain Monte Carlo (e.g., Gibbs sampling)
  – Construct a Markov process from whose steady-state distribution we want to sample.

- Global Optimization (e.g., simulated annealing)
  – Construct a Markov process the support of whose steady-state distribution is the set of global optima of a function we want to maximize.

Illustrations/Variations/Extensions: Continuous-State Markov Processes

Linear Gaussian state space system:

$$\alpha_t = T\alpha_{t-1} + R\eta_t$$

$$y_t = Z\alpha_t + \varepsilon_t$$

$$\eta_t \sim N,\ \varepsilon_t \sim N$$

Part II: State Space

State Space

$$\underset{mx1}{\alpha_t} = \underset{mxm}{T}\ \underset{mx1}{\alpha_{t-1}} + \underset{mxg}{R}\ \underset{gx1}{\eta_t}$$

$$\underset{Nx1}{y_t} = \underset{Nxm}{Z}\ \underset{mx1}{\alpha_t} + \underset{Nx1}{\varepsilon_t}$$

$$\begin{pmatrix} \eta_t \\ \varepsilon_t \end{pmatrix} \sim WN\left(\underline{0}, \text{diag}(\underbrace{Q}_{g\times g}, \underbrace{H}_{N\times N})\right)$$

$$E(\alpha_0\ \eta_t\ ') = 0_{mxg} \qquad E(\alpha_0\ \varepsilon_t\ ') = 0_{mxN}$$

State-Space in Density Form (Assuming Normality)

$$\alpha_t | \alpha_{t-1} \sim N(T\alpha_{t-1},\ RQR')$$

$$y_t | \alpha_t \sim N(Z\alpha_t,\ H)$$

Tradeoff Between Generality and Tedium

– Could allow time-varying system matrices

– Could allow exogenous variables in measurement equation

– Could allow correlated measurement and transition disturbances

– Could allow for arbitrary non-linear non-Gaussian structure

AR(1) in State Space Form

$$y_t = \phi\, y_{t-1} + \eta_t$$

$$\eta_t \sim WN(0, \sigma_\eta^2)$$

Already in state space form!

$$\alpha_t = \phi\, \alpha_{t-1} + \eta_t$$

$$y_t = \alpha_t$$

$$(T = \phi,\ R = 1,\ Z = 1,\ Q = \sigma_\eta^2,\ h = 0)$$

AR(p) in State Space Form

$$y_t = \phi_1 y_{t-1} + \ldots + \phi_p y_{t-p} + \eta_t$$

$$\eta_t \sim WN(0, \sigma_\eta^2)$$

$$\alpha_t = \begin{pmatrix} \alpha_{1t} \\ \alpha_{2t} \\ \vdots \\ \alpha_{pt} \end{pmatrix} = \begin{pmatrix} \phi_1 & \\ \phi_2 & I_{p-1} \\ \vdots & \\ \phi_p & 0' \end{pmatrix} \begin{pmatrix} \alpha_{1,t-1} \\ \alpha_{2,t-1} \\ \vdots \\ \alpha_{p,t-1} \end{pmatrix} + \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \eta_t$$

$$y_t = (1, 0, \ldots, 0)\, \alpha_t = \alpha_{1t}$$

N-Variable VAR(p) in State Space Form

$$\underset{Npx1}{\begin{pmatrix} \alpha_{1t} \\ \alpha_{2t} \\ \vdots \\ \alpha_{pt} \end{pmatrix}} = \underset{NpxNp}{\begin{pmatrix} \Phi_1 & \\ \Phi_2 & I_{N(p-1)} \\ \vdots & \\ \Phi_p & 0' \end{pmatrix}} \underset{Npx1}{\begin{pmatrix} \alpha_{1,t-1} \\ \alpha_{2,t-1} \\ \vdots \\ \alpha_{p,t-1} \end{pmatrix}} + \underset{NPxN}{\begin{pmatrix} I_N \\ 0_{NxN} \\ \vdots \\ 0_{NxN} \end{pmatrix}} \eta_t$$

$$\underset{Nx1}{y_t} = \underset{NxNp}{(I_N, \quad 0_N \quad , \ ..., \ 0_N)} \underset{Npx1}{\alpha_t}$$

Single-Factor Exact Dynamic Factor Model

(White noise idiosyncratic factors uncorrelated with each other and uncorrelated with AR(1) factor at all leads and lags...)

$$\begin{pmatrix} y_{1t} \\ \vdots \\ y_{Nt} \end{pmatrix} = \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_N \end{pmatrix} f_t + \begin{pmatrix} \varepsilon_{1t} \\ \vdots \\ \varepsilon_{Nt} \end{pmatrix}$$

$$f_t = \phi f_{t-1} + \eta_t$$

Already in state-space form!

Exact Dynamic Factor Model More Generally

$$y_t = \Lambda f_t + \varepsilon_t$$

$$\varepsilon_t \sim WN(0, \Sigma)$$

$$\Phi(L) f_t = v_t$$

$$v_t \perp \varepsilon_{t-\tau}, \ \forall \tau$$

$$\implies \Sigma_y = \Lambda \Sigma_f \Lambda' + \Sigma_\varepsilon$$

– Covariance matrix in static case
– Spectral density matrix in dynamic case
– Many variations and extensions

Part III: The Kalman Filter

State Space Representation

$$\underset{mx1}{\alpha_t} = \underset{mxm}{T}\ \underset{mx1}{\alpha_{t-1}} + \underset{mxg}{R}\ \underset{gx1}{\eta_t}$$

$$\underset{Nx1}{y_t} = \underset{Nxm}{Z}\ \underset{mx1}{\alpha_t} + \underset{Nx1}{\varepsilon_t}$$

$$\begin{pmatrix} \eta_t \\ \varepsilon_t \end{pmatrix} \sim WN\left(0, \text{diag}(\underbrace{Q}_{g\times g}, \underbrace{H}_{N\times N})\right)$$

$$E(\alpha_0\ \eta_t') = 0_{mxg}$$

$$E(\alpha_0\ \varepsilon_t') = 0_{mxN}$$

Statement of the Kalman Filter (After Initializing)

II. Prediction Recursions

$$a_{t/t-1} = T\, a_{t-1}$$

$$P_{t/t-1} = T\, P_{t-1}\, T' + R\, Q\, R'$$

III. Updating Recursions

$$a_t = a_{t/t-1} + P_{t/t-1}\, Z'\, F_t^{-1}\, (y_t - Z a_{t/t-1})$$

$$\text{(where } F_t = Z\, P_{t/t-1}\, Z' + H\text{)}$$

$$P_t = P_{t/t-1} - P_{t/t-1}\, Z'\, F_t^{-1}\, Z\, P_{t/t-1}$$

$$\text{t} = 1, \ \ldots, \ \text{T}$$

Kalman Filter in Density Form (Assuming Normality)

Initialize at a_0, P_0

State prediction:

$$\alpha_t|\tilde{y}_{t-1} \sim N(a_{t/t-1},\ P_{t/t-1})$$
$$a_{t/t-1} = Ta_{t-1}$$
$$P_{t/t-1} = TP_{t-1}T' + RQR'$$

Update:

$$\alpha_t|\tilde{y}_t \sim N(a_t,\ P_t)$$
$$a_t = a_{t/t-1} + K_t(y_t - Za_{t/t-1})$$
$$P_t = P_{t/t-1} - K_tZP_{t/t-1}$$

where $\tilde{y}_t = \{y_1, ..., y_t\}$

Kalman Smoother

1. (Kalman) filter forward through the sample, $t = 1, \ ..., \ T$

2. Smooth backward, $t = T, \ (T-1), \ (T-2), \ ..., \ 1$

Initialize: $a_{T,T} = a_T, \ P_{T,T} = P_T$

Then:

$$a_{t,T} = a_t + J_t(a_{t+1,T} - a_{t+1,t})$$

$$P_{t,T} = P_t + J_t(P_{t+1,T} - P_{t+1,t})J_t'$$

where

$$J_t = P_t T' P_{t+1,t}^{-1}$$

Point Prediction of y_t

Prediction:

$$y_{t/t-1} = Za_{t/t-1}$$

Prediction error:

$$v_t = y_t - Za_{t/t-1}$$

Density Prediction of y_t

$$y_t|\Omega_{t-1} \sim N(Za_{t/t-1},\ F_t)$$

or equivalently

$$v_t \,|\, \Omega_{t-1} \sim N\,(0,\ F_t)$$

Normality follows from linearity of all transformations.

Conditional mean already derived.

Proof that the conditional covariance matrix is F_t:

$$\begin{aligned} E_{t-1}v_t v_{t'} &= E_{t-1}[Z\,(\alpha_t - a_{t/t-1}) + \varepsilon_t]\,[Z\,(\alpha_t - a_{t/t-1}) + \varepsilon_t]' \\ &= ZP_{t/t-1}Z' + H \\ &= F_t \end{aligned}$$

Likelihood Evaluation, Optimization, and Inference

Likelihood I: Brute Force

$$y \sim N(\mu, \Sigma(\theta))$$

Example: AR(1)

$$(y_t - \mu) = \phi(y_{t-1} - \mu) + \varepsilon_t$$

$$\Sigma_{ij}(\phi, \sigma^2) = \frac{\sigma^2}{1 - \phi^2} \phi^{|i-j|}$$

Likelihood I:
Brute Force, Cont'd

$$L(y;\theta) = (2\pi)^{T/2}|\Sigma(\theta)|^{-1/2} exp\left(-\frac{1}{2}(y-\mu)'\Sigma^{-1}(\theta)(y-\mu)\right)$$

$$lnL(y;\theta) = const - \frac{1}{2}ln|\Sigma(\theta)| - \frac{1}{2}(y-\mu)'\Sigma^{-1}(\theta)\ (y-\mu)$$

TxT matrix $\Sigma(\theta)$ can be *very* hard to calculate and invert

Likelihood II:
The Schweppe Decomposition and Kalman Filter

Schweppe's likelihood decomposition is:

$$L(y_1, \ldots, y_T; \theta) = \prod_{t=1}^{T} L_t(y_t | y_{t-1}, \ldots, y_1; \theta)$$

or:

$$\ln L(y_1, \ldots, y_T; \theta) = \sum_{t=1}^{T} \ln L_t(y_t | y_{t-1}, \ldots, y_1; \theta)$$

"Prediction-error decomposition"

Likelihood II:
The Schweppe Decomposition and Kalman Filter, Cont'd

In the univariate Gaussian case, the Schweppe decomposition is

$$\ln L = -\frac{T}{2}\ln 2\pi - \frac{1}{2}\sum_{t=1}^{T}\ln \sigma_t^2 - \frac{1}{2}\sum_{t=1}^{T}\frac{(y_t-\mu_t)^2}{\sigma_t^2}$$

$$= -\frac{T}{2}\ln 2\pi - \frac{1}{2}\sum_{t=1}^{T}\ln F_t - \frac{1}{2}\sum_{t=1}^{T}\frac{v_t^2}{F_t}$$

Kalman filter delivers v_t and F_t!

No matrix inversion!

No need for tedious analytic likelihood derivations!

Likelihood II:
The Schewppe Decomposition and Kalman Filter, Cont'd

In the N-variate Gaussian case, the Schweppe decomposition is

$$\ln L = -\frac{NT}{2}\ln 2\pi - \frac{1}{2}\sum_{t=1}^{T}\ln|\Sigma_t| - \frac{1}{2}\sum_{t=1}^{T}(y_t-\mu_t)'\Sigma_t^{-1}(y_t-\mu_t)$$

$$= -\frac{NT}{2}\ln 2\pi - \frac{1}{2}\sum_{t=1}^{T}\ln|F_t| - \frac{1}{2}\sum_{t=1}^{T}v_t'F_t^{-1}v_t$$

Kalman filter again delivers v_t and F_t.

Only the small matrix F_t $(N \times N)$ need be inverted.

Big Data and Filtering I

Kalman filter requires F^{-1}. F is $(N \times N)$. How to proceed in high dimensions?

– F at least has some significant structure (symmetric, psd) that might be exploited.

– Replace F with a sparse matrix that has (approximately) the same inverse.

– View F as an object to be *estimated* (after all, F is the covariance matrix of the 1-step-ahead data prediction errors from the prediction step) and use an estimator that imposes sparsity (e.g., Fan et al.).

– Impose diagonality? Block diagonality? Equicorrelation?

– See Jungbacker and Koopman.

Big Data and Filtering II

The recursive KF structure is lost with vintage data.

– But Big Data and real time often go together

– And real time leads to vintage data (revisions, entry and exit, ...)

– But apart from benchmark revisions, economic vintage data generally involves revisions going back only four quarters (say). So not all *recursivity* is lost. How to modify the standard filter?

Numerical Maximization of the Gaussian Likelihood

- The key is to be able to evaluate *lnL* for a given parameter configuration
- Then we can climb uphill to maximize *lnL* to get the MLE
- EM especially useful in high dimensions. Approaches optimum quickly.
- e.g., high-dimensional DFM's can still be estimated with SS/EM – Perhaps no need for two-step procedures using first-step PC's, even in high dimensions.
- Plus high dimensions may be unnecessary and tricky (Doz. et al.)

VAR's in High Dimensions

DGP: N-Variable $VAR(p)$, $t = 1, ..., T$

$$\Phi(L)x_t = \varepsilon_t$$

$$\varepsilon_t \sim iid(0, \Sigma)$$

If you understand the VAR, you understand *everything*.

Traditionally, e.g., 4-Variable $VAR(3)$

(1) Estimate the VAR

(2) Identify the estimated VAR

(3) Understand the identified estimated VAR

– Examine variance decompositions, etc.

DGP: N-Variable $VAR(p)$, $t = 1, ..., T$

$$\Phi(L)x_t = \varepsilon_t$$

$$\varepsilon_t \sim iid(0, \Sigma)$$

If you understand the VAR, you understand *everything*.
Traditionally, e.g., 4-Variable $VAR(3)$

Now, perhaps a 5000-Variable $VAR(50)$
(e.g., a high-dim set of asset return volatilities with long memory)
"High dimensionality"
"Big Data"

(1) Estimate the VAR

(2) Identify the estimated VAR

(3) Understand the identified estimated VAR
– Examine variance decompositions, etc.

Variance Decomposition Matrix

D^H

	x_1	x_2	...	x_N
x_1	d^H_{11}	d^H_{12}	$\cdots$	d^H_{1N}
x_2	d^H_{21}	d^H_{22}	$\cdots$	d^H_{2N}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
x_N	d^H_{N1}	d^H_{N2}	$\cdots$	d^H_{NN}

Connectedness involves the **non-diagonal** elements of D^H

(1) Estimate the *VAR*

Key theme:
One way or another, we need to recover d.f.

Selection and Shrinkage

The Parsimony and KISS Principles

– Other things equal, smaller is better

– But be sophisticated

Constraints can be *good*.

Hard constraints: Selection

Soft constraints: Shrinkage

Selection ("Hard Constraints")

All-Subsets Selection I: Information Criteria

What not to do...

$$MSE = \frac{\sum_{t=1}^{T} e_t^2}{T}$$

$$R^2 = 1 - \frac{\sum_{t=1}^{T} e_t^2}{\sum_{t=1}^{T}(y_t - \bar{y})^2}$$

$$= 1 - \frac{MSE}{\frac{1}{T}\sum_{t=1}^{T}(y_t - \bar{y})^2}$$

Still bad:

$$s^2 = \frac{\sum_{t=1}^{T} e_t^2}{T-k}$$

$$s^2 = \left(\frac{T}{T-k}\right)\left(\frac{\sum_{t=1}^{T} e_t^2}{T}\right)$$

$$\bar{R}^2 = 1 - \frac{\sum_{t=1}^{T} e_t^2 \ / \ T-k}{\sum_{t=1}^{T}(y_t - \bar{y}_t)^2 \ / \ T-1}$$

$$= 1 - \frac{s^2}{\sum_{t=1}^{T}(y_t - \bar{y}_t)^2 \ / \ T-1}$$

Good:

$$SIC = T^{\left(\frac{k}{T}\right)} \left(\frac{\sum_{t=1}^{T} e_t^2}{T} \right)$$

More generally,

$$SIC = \frac{-2lnL}{T} + \frac{KlnT}{T}$$

"Consistency" ("oracle property")

Also AIC ("efficiency")

All-Subsets Selection II: Cross Validation

– "Leave one out" ("$T - fold$" CV)
(Split the data into T pieces and predict each)
– "$M - fold$" CV
(Split the data into M pieces ($M < T$) and predict each)
– As M falls, M-fold CV eventually becomes consistent
– $M = 10$ often works well in practice.
– SIC achieves consistency by penalizing in-sample residual MSE to obtain an approximately-unbiased estimate of out-of-sample MSE
– CV achieves consistency by directly obtaining an unbiased estimate of out-of-sample MSE
– CV is more general than information criteria insofar as it can be used even when the model degrees of freedom is unclear
– Non-quadratic loss can be introduced easily
– Generalizations to time-series contexts are available

Partial-Subsets Selection

All-subsets selection, of whatever type, quickly gets hard as there are 2^K subsets of K regressors. Other procedures, like the stepwise selection procedures that we now introduce, don't explore every possible subset. They are more ad hoc but very useful.

Partial-Subsets Selection I: Forward Stepwise Regression

– Begin regressing only on an intercept
– Move to a one-regressor model by including that variable with the smallest t-stat p-value
– Move to a two-regressor model by including that variable with the smallest p-value. Etc.

"Greedy algorithm," producing an increasing sequence of candidate models.
– Often people use information criteria or CV to select from the stepwise sequence of models.
– No guaranteed optimality properties of the selected model. But it often "works"

Partial-Subsets Selection II: Backward Stepwise Regression

– Start with a regression that includes all K variables
– Move to a $K-1$ variable model by dropping the variable with the largest t-stat p-value
– Move to a $K-2$ variable model by dropping the variable with the largest p-value

"Greedy algorithm," producing a decreasing sequence of candidate models.
– Often people use information criteria or CV to select from the stepwise sequence of models.
– No guaranteed optimality properties of the selected model. But it often "works"

Partial-Subsets Selection III: $AR(p)$ Selection Only Over p

– Standard practice for many decades.

– Not clear why.

Shrinkage ("Soft Constraints")

Bayes

Shrinkage is a generic feature of Bayesian estimation. The Bayes rule under quadratic loss is the posterior mean, which is a weighted average of the MLE and the prior mean,

$$\hat{\beta}_{bayes} = \omega_1 \hat{\beta}_{MLE} + \omega_2 \beta_0.$$

Hence the Bayes rule pulls, or "shrinks," the MLE toward the prior mean.

– The weights depend on MLE precision relative to prior precision.

Leading Example: Ridge

$$\hat{\beta}_{ridge} = (X'X + \lambda I)^{-1} X'y.$$

– $\lambda \to 0$ produces OLS

– $\lambda \to \infty$ shrinks completely to 0

– λ can be chosen by CV

– Notice that λ can *not* be chosen by information criteria, as K regressors are included regardless of λ.

– The ridge estimator can be shown to be the posterior mean for a certain prior and likelihood.

– Also "regularizes" and so can handle situations with $K > T$

Shrinkage for High-Dimensional VAR's

M. Banbura, D. Giannone, and L. Reichlin, Large Bayesian Vector Auto-Regressions, Journal of Applied Econometrics, 2010.

G. Koop, Forecasting with Medium and Large Bayesian VARs, Journal of Applied Econometrics, 2013.

A. Carriero, T.E. Clark and M. Marcellino, Large Vector Autoregressions with Asymmetric Priors, w.p., 2015.

Compression

Pettenuzzo, Koop, Korobolis: "Bayesian Compressed VAR's"

– Bayesian shrinkage. NCP (normal-Wishart), so no MCMC

– Selection via imposition of sparsity

– Sparsity enforced on both $\Phi(L)$ and on Σ

– potential rank reduction, with the decision made by Bayesian model averaging.

– Can handle time-varying parameters

Selection *and* Shrinkage

Penalized Estimation for Shrinkage and/or Selection

$$\hat{\beta}_q = \text{argmin}_\beta \sum_{t=1}^{T} \left(y_t - \sum_{i=1}^{K} \beta_i x_{it} \right)^2 \quad s.t. \quad \sum_{i=1}^{K} |\beta_i|^q \leq c$$

$$\hat{\beta}_q = \text{argmin}_\beta \left(\sum_{t=1}^{T} \left(y_t - \sum_{i=1}^{K} \beta_i x_{it} \right)^2 + \lambda \sum_{i=1}^{K} |\beta_i|^q \right)$$

Concave penalty functions non-differentiable at the origin
produce selection to zero (e.g., $q = 1/2$)

Smooth convex penalties produce shrinkage toward 0 (e.g., $q = 2$)

$q = 1$ is both concave and convex,
so it selects to 0 *and* shrinks to 0
("Lasso")

Ridge

$$\hat{\beta}_2 = \arg\min_{\beta} \left(\sum_{t=1}^{T} \left(y_t - \sum_{i=1}^{K} \beta_i x_{it} \right)^2 + \lambda \sum_{i=1}^{K} |\beta_i|^2 \right)$$

($q = 2$, shrinks β_i toward 0 $\forall i$)

– Doesn't select

– Shrinks toward zero
No shrinkage: ($\lambda \to 0$): OLS
Full shrinkage ($\lambda \to \infty$): Zero weights

– Also "regularizes" and so can handle situations with $K > T$

Lasso

$$\hat{\beta}_1 = \arg\min_{\beta} \left(\sum_{t=1}^{T} \left(y_t - \sum_{i=1}^{K} \beta_i x_{it} \right)^2 + \lambda \sum_{i=1}^{K} |\beta_i| \right)$$

$(q = 1, \text{ shrinks } \beta_i \text{ toward } 0 \ \forall i)$

– Selects to zero

– Shrinks toward zero

No shrinkage ($\lambda \to 0$): OLS

Full shrinkage ($\lambda \to \infty$): Zero weights

– Also "regularizes" and so can handle situations with $K > T$

Properties of Lasso

– Selection *and* Shrinkage

– Like ridge and other Bayesian procedures, lasso requires only *one* (convex) estimation.

– Convenient d.f. result. The effective number of parameters is precisely the number of variables selected (number of non-zero β's). This means that we can use info criteria to select among "lasso models" for various λ.

Varieties of Lasso

$$\hat{\beta}_{Lasso} = argmin_{\beta} \left(\sum_{t=1}^{T} \left(y_t - \sum_{i=1}^{K} \beta_i x_{it} \right)^2 + \lambda \sum_{i=1}^{K} |\beta_i| \right)$$

$$\hat{\beta}_{ALasso} = argmin_{\beta} \left(\sum_{t=1}^{T} \left(y_t - \sum_{i=1}^{K} \beta_i x_{it} \right)^2 + \lambda \sum_{i=1}^{K} w_i |\beta_i| \right)$$

$$\hat{\beta}_{Enet} = argmin_{\beta} \left(\sum_{t=1}^{T} \left(y_t - \sum_{i=1}^{K} \beta_i x_{it} \right)^2 + \lambda \sum_{i=1}^{K} \left(\alpha |\beta_i| + (1-\alpha)\beta_i^2 \right) \right)$$

$$\hat{\beta}_{AEnet} = argmin_{\beta} \left(\sum_{t=1}^{T} \left(y_t - \sum_{i=1}^{K} \beta_i x_{it} \right)^2 + \lambda \sum_{i=1}^{K} w_i \left(\alpha |\beta_i| + (1-\alpha)\beta_i^2 \right) \right)$$

where $w_i = 1/|\hat{\beta}_i|^{\nu}$, $\hat{\beta}_i$ is OLS or ridge, $\nu > 0$, and $\alpha \in [0, 1]$.

Adaptive Elastic Net has Emerged as Standard

$$\hat{\beta}_{AEnet} = argmin_{\beta} \left(\sum_{t=1}^{T} \left(y_t - \sum_i \beta_i x_{it} \right)^2 + \lambda \sum_{i=1}^{K} w_i \left(\alpha |\beta_i| + (1-\alpha)\beta_i^2 \right) \right)$$

where $w_i = 1/|\hat{\beta}_i|^{\nu}$, $\hat{\beta}_i$ is OLS or ridge, and $\nu > 0$

– Has the oracle property

Lasso is $\alpha = 1$, $w_i = 1 \forall i$
Adaptive lasso is $\alpha = 1$
Elastic net is $w_i = 1 \forall i$

LASSO Opportunities in VAR's

– Group LASSO

– System LASSO vs. equation-by-equation
(single system λ)

– Shrink and select on Σ, not just β.
This may help ensure p.s.d.

– Shrinking/selecting *functions* of coefficients.
e.g., LASSO directly on variance decompositions?

Pruning

– Thus far we have used lasso for selecting VAR coefficients to 0.

– We can also do a second round pruning on the resulting D

– Can prune "small" elements

– Can prune "insignificant" estimates
(Assess significance using methods of Kilian and Lutkepohl (2016).
They emphasize IRF's but parallel results exist for VD's.)

Alternative Shrinkage/Selection Directions: Generalized Penalized Estimation

– Standard methods shrink/select toward zero.
Not always appropriate.

– e.g., "equal weights prior" for forecast combination.

– e.g., "equicorrelation prior" for conditional covariance.

– e.g., "Minnesota Prior" for conditional mean parameters

$$\hat{\beta} = \textit{argmin}_{\beta} \left(\sum_{t=1}^{T} \left(y_t - \sum_{i=1}^{K} \beta_i x_{it} \right)^2 + \lambda \sum_{i=1}^{K} |\beta_i - \beta_i^0|^q \right)$$

Selects/shrinks to β^0

Generalized ridge, generalized LASSO, etc.

Ridge and Lasso for DF Structure

– We've already seen that DF structure can be *imposed* (selected) using standard SS/ML/Bayes methods.

– What about ridge (pure shrinkage) for DF structure? Need to change the centering to reflect reduced rank. Tricky. (Build on Ledoit.)

– What about LASSO (selection and shrinkage) for DF structure? Need to change the centering to reflect reduced rank. Tricky. (Build on Ledoit.)

Switching From LS to ML

- Instead of min SSR + penalty, could max Gaussian lnL - penalty
  - More generally, for any model with a lnL, max lnL - penalty
  - This opens up LASSO and variants to any model with a lnL

Allowing for Parameter Variation

– Rolling

– Weighted rolling (e.g., EWMA)

– Random-Walk (West, Koop et al.)

– Factor structure in parameter variation (Stevanovic)

Shrinkage for High-Dimensional VAR's with TVP's

G. Koop and D. Korobilis. Large Time-Varying Parameter VARs. Journal of Econometrics, 177:185-198, 2013.

T. Park and G. Casella. The Bayesian Lasso. Journal of American Statistical Association, 103:681-685, 2008.
– Show that lasso is posterior mode for certain prior/likelihood choice. Paves the way for Bayesian MCMC estimation in ultra-high dimensions. Also paves the way for allowance for TVP's in Belmonte et al (2014).

H. Wang. Bayesian Graphical Lasso Models and Efficient Posterior Computation. Bayesian Analysis, 7:867-886, 2012.
– High-dim cov matrices can be estimated using Bayesian MCMC

M. Belmonte, G. Koop, and D. Korobilis. Hierarchical Shrinkage in Time-Varying Parameter Models. Journal of Forecasting, 33:80-94, 2014.
– Bayesian graphical lasso approach using hierarchical priors.

Allowing for Mixed Frequencies (MIDAS)

1. High-dimensional and mixed-frequency data go together in time series.
2. So high-dimensional MIDAS vector autoregression (VAR) may be important.
3. MIDAS VAR is appearing (Ghysels), but it's still low-dimensional.

   Next steps:

   3.1 Move to high dimensions by using regularization methods (e.g. LASSO variants)
   3.2 Allow for many observational frequencies (five or six, say)
   3.3 Allow for the "rough edges" that will invariably arise at the sample beginning and end

(2) Identify the Estimated *VAR*

Key themes:
Mechanical identifications are needed

If You Understand the *VAR*, You Understand *Everything*

But it's hard to understand the *VAR*.

– Staring at coefficient matrices is inadequate

– Staring at coefficient matrices
and innovation covariance matrices
is adequate but unproductive

– Staring at variance decompositions (VD's)
is adequate and maybe productive
But how will you identify them?
And how will you stare at them?

VD's Require Identification

Intricate theory identification (DSGE)
– Generally unavailable in high dimensions and arguably undesirable

Less intricate theory identification (SVAR)
– Generally unavailable in high dimensions and arguably undesirable

Cholesky factor identification
– Desirably mechanical but requires recursive ordering

Koop-Pesaran-Shin generalized identification
– Desirably mechanical and doesn't require recursive ordering
(but of course makes other assumptions)

– Other matrix square roots?

Perhaps DAG Modeling can Help

Cholesky Orthogonalization:

$$(I - \Phi_1 L - ... - \Phi_p L^p) y_t = P v_t$$

$$v_t \sim WN(0, I),$$

where $\Sigma = PP'$ (Cholesky factorization)

Moving-average representation:

$$y_t = (I + \Theta_1 L + \Theta_2 L^2 + ...) P v_t$$

$$= P v_t + \Theta_1 P v_{t-1} + ...$$

Cholesky Corresponds to a Recursive Structural System

Structural Simultaneous-Equations Model:

$$Ay_t = \Phi_1 y_{t-1} + ... + \Phi_p y_{t-p} + \varepsilon_t$$

$$\varepsilon_t \sim (0, \Sigma)$$

Recursive SEM: A triangular and Σ diagonal

$$Ty_t = \Phi_1 y_{t-1} + ... + \Phi_p y_{t-p} + \varepsilon_t$$

$$\varepsilon_t \sim (0, D)$$

The MA representation of the reduced form is:

$$y_t = T^{-1}\varepsilon_t + ...$$

Directed Acyclical Graphs (DAG's)

Start with:

Heckman, J. and Pinto, R. (2015), "Causal Analysis After Haavelmo," *Econometric Theory*, `http://www.nber.org/papers/w19453`

Then back up and read or re-read:

Lauritzen, S. (1996). *Graphical Models*, Clarendon Press.

Pearl, J. (2009), *Causality: Models, Reasoning, and Inference*, Cambridge University Press (second edition).

End with:

Hjsgaard, S. "Graphical Models and Bayesian Networks with R" `www.people.math.aau.dk/~sorenh/misc/2014-useR-GMBN/`

Directed Acyclical Graphs (DAG's) are Also Recursive

– Causal relationships are represented by a graph G,
where nodes correspond to variables.

– Nodes are connected by arrows that
represent causal influences between variables.

– The set of descendants of a variable V consists of all variables
connected to V by arrows of the same direction arising from V.

– Graph G is called a DAG if no variable is a descendant of itself.

– Not fully simultaneous. Instead, *recursive*!

Key DAG Insights

– DAG environment is recursive system environment

– Local Markov condition: In a DAG, a variable is independent of its non-descendants conditional on its ancestors.

(Recall conditional independence:
y is independent of x conditional on z if and only if
$\Pr(y \cap x \mid z) = \Pr(y \mid z) \Pr(x \mid z)$. Coincides with zero partial correlation in the Gaussian case (Baba et al. 2004).)

– So tests for conditional independence might help to determine recursive ordering.

(3) Understand the Identified Estimated *VAR*

Key theme: Staring at a massive variance decomposition matrix (D) is just as hopeless
as staring at massive coefficient matrices

Why?

Can't stare productively at coefficient and covariance matrices.

When $N = 5$ all is well.
We can stare productively at D.

When $N = 5000$ we're in trouble.
We can no longer stare productively at D!

The key tool for "digesting" VAR info
(i.e., examination of D)
is itself now indigestible!

But Graph-Theoretic (Network) Tools Come to the Rescue

Network Theory:
The key to connectedness-based summaries of variance decompositions

Network Visualization:
The key to deep understanding of variance decompositions

Part I: Graph-Theoretic D Summarization

Interpret D as a Network Adjacency Matrix
Summarize Using the Degree Distribution

What does that mean?

A Natural Financial/Economic Connectedness Question:

What fraction of the H-step-ahead prediction-error variance of x_i is due to shocks in x_j, $j \neq i$?

Non-own elements of the variance decomposition: d_{ij}^H, $j \neq i$

Variance Decomposition Matrix

D^H

	x_1	x_2	...	x_N
x_1	d^H_{11}	d^H_{12}	$\cdots$	d^H_{1N}
x_2	d^H_{21}	d^H_{22}	$\cdots$	d^H_{2N}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
x_N	d^H_{N1}	d^H_{N2}	$\cdots$	d^H_{NN}

Connectedness involves the **non-diagonal** elements of D^H

Connectedness Table

Connectedness Table

	x_1	x_2	...	x_N	From Others to i
x_1	d_{11}^H	d_{12}^H	$\cdots$	d_{1N}^H	$\sum_{j=1}^N d_{1j}^H, j \neq 1$
x_2	d_{21}^H	d_{22}^H	$\cdots$	d_{2N}^H	$\sum_{j=1}^N d_{2j}^H, j \neq 2$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
x_N	d_{N1}^H	d_{N2}^H	$\cdots$	d_{NN}^H	$\sum_{j=1}^N d_{Nj}^H, j \neq N$
To Others From j	$\sum_{i=1}^N d_{i1}^H$ $i \neq 1$	$\sum_{i=1}^N d_{i2}^H$ $i \neq 2$	$\cdots$	$\sum_{i=1}^N d_{iN}^H$ $i \neq N$	$\sum_{i,j=1}^N d_{ij}^H$ $i \neq j$

Just the variance decomposition matrix, D^H

Connectedness Graph Table (All H)

– Connectedness table is now a table of *graphs*.

– Just as macroeconomists routinely do with IRF's

Connectedness Measures, $C(x, H, \Phi(L), \Sigma)$

- Pairwise Directional: $C^H_{i \leftarrow j} = d^H_{ij}$ ("i's imports from j")
- Net: $C^H_{ij} = C^H_{j \leftarrow i} - C^H_{i \leftarrow j}$ ("ij bilateral trade balance")

- Total Directional:
  - From others to i: $C^H_{i \leftarrow \bullet} = \sum_{\substack{j=1 \\ j \neq i}}^{N} d^H_{ij}$ ("i's total imports")
  - To others from j: $C^H_{\bullet \leftarrow j} = \sum_{\substack{i=1 \\ i \neq j}}^{N} d^H_{ij}$ ("j's total exports")
- Net: $C^H_i = C^H_{\bullet \leftarrow i} - C^H_{i \leftarrow \bullet}$ ("i's multilateral trade balance")

- System-wide: $C^H = \frac{1}{N} \sum_{\substack{i,j=1 \\ i \neq j}}^{N} d^H_{ij}$ ("total world exports")

Reading and Web Materials

Recent papers:

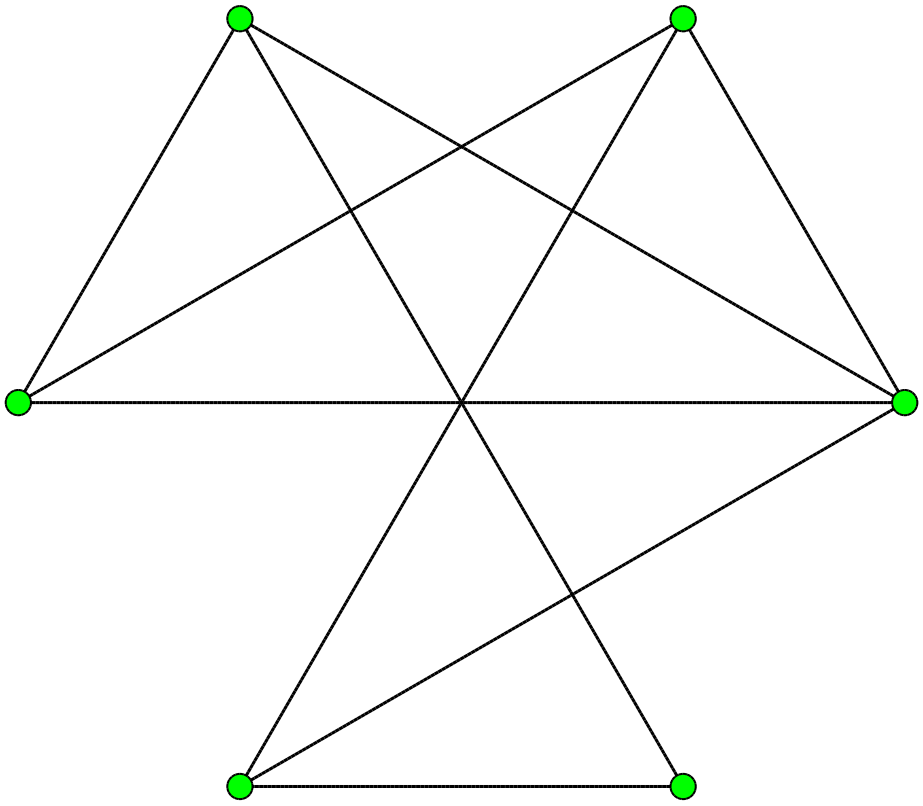
Diebold, F.X. and Yilmaz, K. (2014), "On the Network Topology of Variance Decompositions: Measuring the Connectedness of Financial Firms," *Journal of Econometrics*, 182, 119-134.

Demirer, M., Diebold, F.X., Liu, L. and Yilmaz, K. (2015), "Estimating Global Bank Network Connectedness," Manuscript, MIT, Penn and Koc.

Recent book:

Diebold, F.X. and Yilmaz, K. (2015), *Financial and Macroeconomic Connectedness: A Network Approach to Measurement and Monitoring*, Oxford University Press. With K. Yilmaz.

Network Representation: Graph and Matrix



$$A = \begin{pmatrix} 0 & 1 & 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 \end{pmatrix}$$

Symmetric adjacency matrix A

$A_{ij} = 1$ if nodes i, j linked

$A_{ij} = 0$ otherwise

Network Connectedness: The Degree Distribution

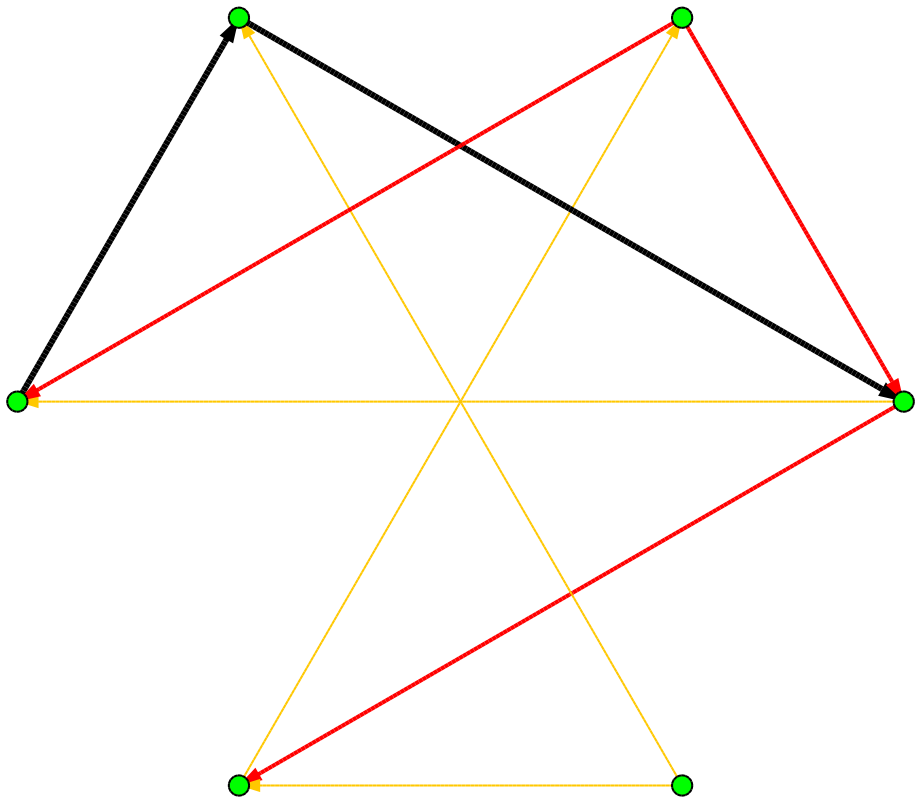
Degree of node i, d_i:

$$d_i = \sum_{j=1}^{N} A_{ij}$$

Discrete *degree distribution* on $0, \ldots, N-1$

Mean degree, $E(d)$, is the key connectedness measure

Network Representation II (Weighted, Directed)



$$A = \begin{pmatrix} 0 & .5 & .7 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & .3 & 0 \\ 0 & 0 & 0 & .7 & 0 & .3 \\ .3 & .5 & 0 & 0 & 0 & 0 \\ .5 & 0 & 0 & 0 & 0 & .3 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

"to i, from j"

Network Connectedness II: The Degree Distribution(s)

$A_{ij} \in [0, 1]$ depending on connection strength

Two degrees:

$$d_i^{from} = \sum_{j=1}^{N} A_{ij}$$

$$d_j^{to} = \sum_{i=1}^{N} A_{ij}$$

"from-degree" and "to-degree" distributions on $[0, N-1]$

Mean degree remains the key connectedness measure

Variance Decompositions as Weighted, Directed Networks

Variance Decomposition / Connectedness Table

	x_1	x_2	...	x_N	From Others
x_1	d_{11}^H	d_{12}^H	$\cdots$	d_{1N}^H	$\sum_{j\neq 1} d_{1j}^H$
x_2	d_{21}^H	d_{22}^H	$\cdots$	d_{2N}^H	$\sum_{j\neq 2} d_{2j}^H$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
x_N	d_{N1}^H	d_{N2}^H	$\cdots$	d_{NN}^H	$\sum_{j\neq N} d_{Nj}^H$
To Others	$\sum_{i\neq 1} d_{i1}^H$	$\sum_{i\neq 2} d_{i2}^H$	$\cdots$	$\sum_{i\neq N} d_{iN}^H$	$\sum_{i\neq j} d_{ij}^H$

Total directional connect. "from," $C_{i\leftarrow\bullet}^H = \sum_{\substack{j=1\\ j\neq i}}^N d_{ij}^H$: "from-degrees"

Total directional connect. "to," $C_{\bullet\leftarrow j}^H = \sum_{\substack{i=1\\ i\neq j}}^N d_{ij}^H$: "to-degrees"

Systemwide connect., $C^H = \frac{1}{N}\sum_{\substack{i,j=1\\ i\neq j}}^N d_{ij}^H$: mean degree

Opportunities with Graph-Theoretic D Summarization

– Examine aspects of the degree distribution *across* H

– Other Connectedness Measures

(a) Multi-step connectedness
The degrees of A track 1-step connectedness.
The degrees of A^k track k-step connectedness,

(b) ∞-step connectedness ("eigenvalue centrality")
Examine second smallest eigenvalue λ_2 of $L = M - A$
(M is a diagonal matrix containing the node degrees)
(A is the adjacency matrix.)

(c) But is any of this necessary/desirable for us?
Could we not simply vary H?

Other Connectedness Measures: Connectedness via Pairwise Granger Causality

– Simple OLS estimation

– Just pairwise so it's like a simple correlation

– Ignores connectedness arising from innovation correlations

– Requires statistical inference (hypothesis testing)

Other Connectedness Measures: Connectedness via Multivariate Granger Causality

– Likely requires regularized estimation
– Not just pairwise so it's like a partial as opposed to simple correlation
– Ignores connectedness arising from innovation correlations
– Requires statistical inference (hypothesis testing)

Other Connectedness Measures: Bonaldi-Hortacsu-Kastl Connectedness

- Treat VAR(1) coefficient matrix as network adjacency matrix
  - Presumably generalize to state-space T matrix
- Ignores connectedness arising from innovation correlations

Other Connectedness Measures: *MES* (S-Risk) Connectedness

$$MES^{j|mkt} = E(r_j \mid \mathbb{C}(r_{mkt}))$$

- Sensitivity of firm j's return to extreme market event $\mathbb{C}$
- Market-based "stress test" of firm j's fragility

"Total directional connectedness *from*" (from-degrees)

"From others to j"

Other Connectedness Measures: *CoVaR* Connectedness

$$VaR^p: \; p = P\left(r < -VaR^p\right)$$

$$CoVaR^{p,j|i}: \; p = P\left(r_j < -CoVaR^{p,j|i} \mid \mathbb{C}\left(r_i\right)\right)$$

$$CoVaR^{p,mkt|i}: \; p = P\left(r_{mkt} < -CoVaR^{p,mkt|i} \mid \mathbb{C}\left(r_i\right)\right)$$

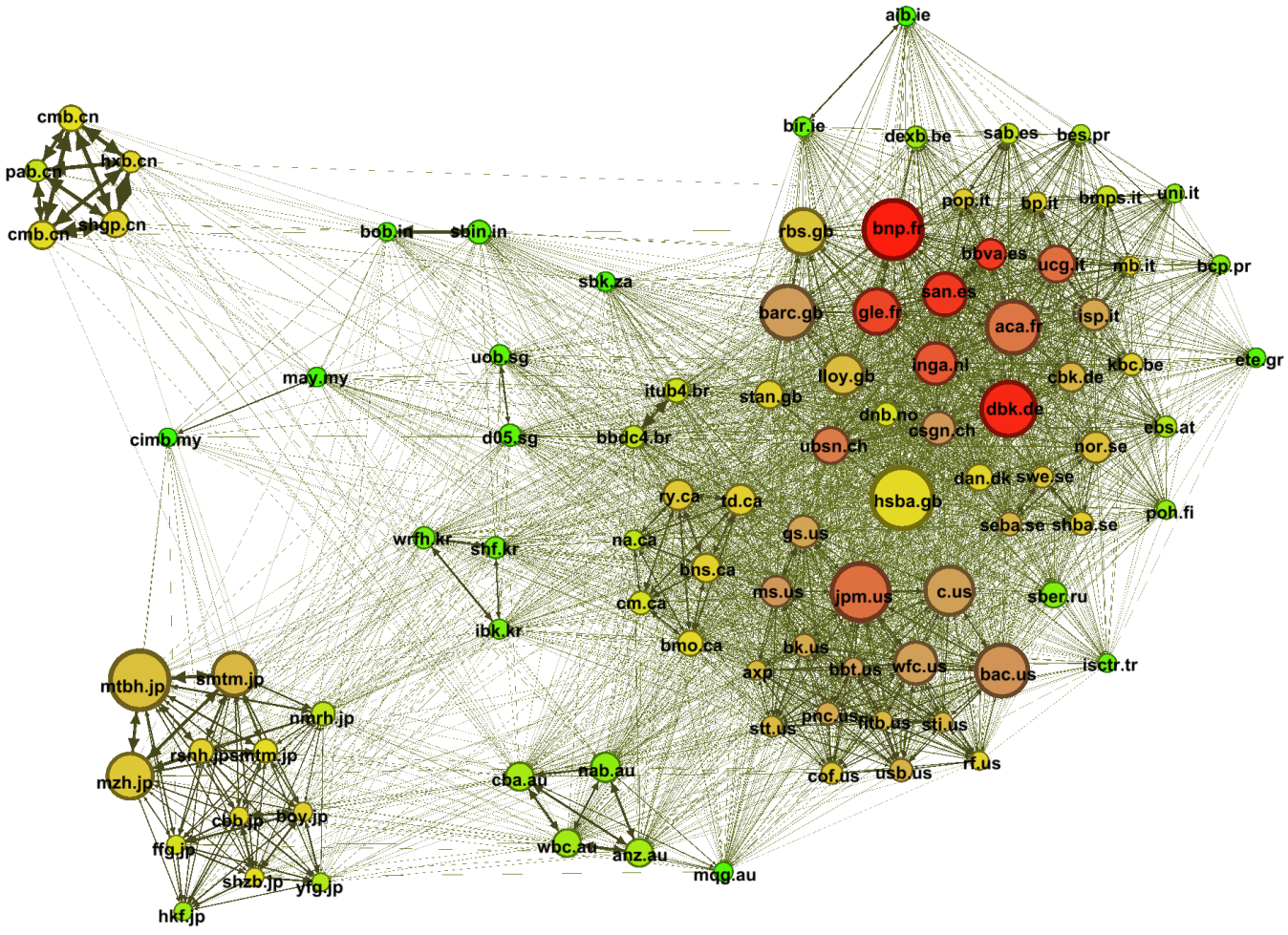
- Measures tail-event linkages
- Leading choice of $\mathbb{C}\left(r_i\right)$ is a VaR breach

"Total directional connectedness *to*" (to-degrees)

"From i to others"

Part II: Graph-Theoretic *D Visualization*

Spring Graph



Spring Graph Detail

- Node size: Asset size (firms), GDP (countries), etc.
- Node color: Total directional connectedness "to others"

- Node location: Average pairwise directional connectedness (Equilibrium of repelling and attracting forces, where (1) nodes repel each other, but (2) edges attract the nodes they connect according to average pairwise directional connectedness "to" and "from.")
- Edge thickness: Average pairwise directional connectedness
- Edge arrow sizes: Pairwise directional connectedness "to" and "from"

Opportunities With D Graphs

– Do we really need different node sizes?

– There does not exist a natural gradation
across colors from "cool" to "hot"
(So use layering rather than colors)

– Animate over H
(But spring graphs have a flip-flop issue that needs solving.)

– Time-varying coefficients and dynamic network graphs
(Rolling estimation, explicitly time-varying coefficients, etc.)

– Animate over time for fixed H,
animate over H for fixed time.
(Again, spring graphs have a flip-flop issue that needs solving.)

Curley at Columbia

http://curleylab.psych.columbia.edu/netviz/netviz1.html
http://curleylab.psych.columbia.edu/netviz/netviz2.html
http://curleylab.psych.columbia.edu/netviz/netviz3.html
http://curleylab.psych.columbia.edu/netviz/netviz4.html
http://curleylab.psych.columbia.edu/netviz/netviz5.html

Novel Econometric Network Visualizations

– Multivariate forecast error covariance matrix

– Multivariate state-space model "A matrix"

– Separate visualizations of "A matrix" and transition shock covariance matrix.

Concluding Perspective

Old view: *VAR*'s unworkable in high dimensions
(Actually no one even *thought* about high dimensions)

New view: *VAR*'s *are* workable in high dimensions

(1) Regularized estimation

(3) Network theory for numerical summarization,
and network graphs for visual understanding

Still *VAR*'s, but:

Important new tools for estimation and analysis in high dimensions
are opening important new research areas

Penn
University of Pennsylvania