Visualizing VAR’s: Regularization and Network Tools for High-Dimensional Financial Econometrics

Francis X. Diebold
University of Pennsylvania

March 7, 2015
DGP: $N$-Variable $\text{VAR}(p)$, $t = 1, \ldots, T$

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

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

Traditionally, e.g., 4-Variable $\text{VAR}(3)$
If you understand the $\text{VAR}$, you understand everything.
DGP: \( N \)-Variable \( \text{VAR}(p) \), \( t = 1, \ldots, T \)

\[
\Phi(L)x_t = \varepsilon_t
\]

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

Traditionally, e.g., 4-Variable \( \text{VAR}(3) \)
If you understand the \( \text{VAR} \), you understand everything.

(1) Estimate the \( \text{VAR} \)

(2) Identify the estimated \( \text{VAR} \)

(3) Understand the estimated \( \text{VAR} \)
   - Examine variance decompositions, etc.
(1) Background Motivation

Financial/Economic Connectedness Measurement
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 / Connectedness Table

\[ D \]

<table>
<thead>
<tr>
<th></th>
<th>( x_1 )</th>
<th>( x_2 )</th>
<th>( \ldots )</th>
<th>( x_N )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( x_1 )</td>
<td>( d_{11}^H )</td>
<td>( d_{12}^H )</td>
<td>( \ldots )</td>
<td>( d_{1N}^H )</td>
</tr>
<tr>
<td>( x_2 )</td>
<td>( d_{21}^H )</td>
<td>( d_{22}^H )</td>
<td>( \ldots )</td>
<td>( d_{2N}^H )</td>
</tr>
<tr>
<td>( \vdots )</td>
<td>( \vdots )</td>
<td>( \vdots )</td>
<td>( \ddots )</td>
<td>( \vdots )</td>
</tr>
<tr>
<td>( x_N )</td>
<td>( d_{N1}^H )</td>
<td>( d_{N2}^H )</td>
<td>( \ldots )</td>
<td>( d_{NN}^H )</td>
</tr>
</tbody>
</table>


www.FinancialConnectedness.org
DGP: $N$-Variable VAR($p$), $t = 1, \ldots, T$

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

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

Traditionally, e.g., 4-Variable VAR(3)
If you understand the VAR, you understand everything.
DGP: $N$-Variable VAR($p$), $t = 1, \ldots, T$

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

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

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

Now, perhaps 5000-Variable VAR(50)
(e.g., asset returns or return volatilities, long memory)
“High dimensionality”
“Big Data”
DGP: $N$-Variable $VAR(p)$, $t = 1, \ldots, T$

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

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

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

Now, perhaps 5000-Variable $VAR(50)$
(e.g., asset returns or return volatilities, long memory)
  “High dimensionality”
  “Big Data”

(1) Estimate the $VAR$
(2) Identify the estimated $VAR$
(3) Understand the estimated $VAR$
  – Examine variance decompositions, etc.
(1) Estimate the \textit{VAR}

Key theme: One way or another, we need to recover d.f.
Leading example: adaptive elastic net (lasso variant)

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

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

Lasso is \( \alpha = 1 \), \( w_i = 1 \forall i \)
Adaptive lasso is \( \alpha = 1 \)
Elastic net is \( w_i = 1 \forall i \)
Opportunities with Shrinkage, Selection, Combinations

- Lasso equation-by-equation vs. system
  We's like system estimation, and system lasso
  (i.e., single system $\lambda$)
Opportunities with Shrinkage, Selection, Combinations

– Lasso equation-by-equation vs. system estimation
We’s like system estimation, and system lasso (i.e., single system $\lambda$)

– Lasso may shrink in an awkward direction.
“Minnesota prior” via modified lasso penalty function
Opportunities with Shrinkage, Selection, Combinations

– Lasso equation-by-equation vs. system
  We’s like system estimation, and system lasso
  (i.e., single system $\lambda$)

– Lasso may shrink in an awkward direction.
  “Minnesota prior” via modified lasso penalty function

– What to Shrink/Select? Coefficients? $D$?
  First-round coefficient lasso with second-round $D$ thresholding?
Opportunities with Shrinkage, Selection, Combinations

– Lasso equation-by-equation vs. system
  We’s like system estimation, and system lasso
  (i.e., single system \( \lambda \))

– Lasso may shrink in an awkward direction.
  “Minnesota prior” via modified lasso penalty function

– What to Shrink/Select? Coefficients? \( D \)?
  First-round coefficient lasso with second-round \( D \) thresholding?

– Is the group lasso appealing?
Opportunities with Shrinkage, Selection, Combinations

- Lasso equation-by-equation vs. system estimation
  We's like system estimation, and system lasso
  (i.e., single system $\lambda$)

- Lasso may shrink in an awkward direction.
  “Minnesota prior” via modified lasso penalty function

- What to Shrink/Select? Coefficients? $D$?
  First-round coefficient lasso with second-round $D$ thresholding?

  - Is the group lasso appealing?

  - Christian Hansen two-step lasso?
“Regularization”: Dimensionality Reduction

Leading example: dynamic factor model

\[ y_t = \lambda f_t + \varepsilon_t \]
\[ \varepsilon_t \sim WN(0, \Sigma) \]
\[ \Phi(L)f_t = \nu_t \]
\[ \nu_t \perp \varepsilon_{t-\tau}, \forall \tau \]
Opportunities with Dimensionality Reduction

– Precise implications of factor structure for $D$?
Opportunities with Dimensionality Reduction

– Precise implications of factor structure for $D$?

– Exact vs. approximate vs. switching factor structure
An Opportunity with Everything: $N > T$

- Equation-by-equation coefficient estimation
An Opportunity with Everything: $N > T$

- Equation-by-equation coefficient estimation
- System (multivariate) coefficient estimation
An Opportunity with Everything: $N > T$

– Equation-by-equation coefficient estimation

– System (multivariate) coefficient estimation

  – Estimate $\Sigma$ (and its Cholesky factor)
(2) Identify the Estimated VAR

Key theme: SVAR-style identification is hopeless

We need to return to basics, like Cholesky
If You Understand the VAR, You Understand Everything

But it’s hard to understand the VAR.

– Pairwise Granger-causality is inadequate
If You Understand the VAR, You Understand *Everything*

But it’s hard to understand the VAR.

– Pairwise Granger-causality is inadequate

– Staring at coefficient matrices is inadequate
If You Understand the VAR, You Understand Everything

But it’s hard to understand the VAR.

- Pairwise Granger-causality is inadequate
- Staring at coefficient matrices is inadequate
  - Staring at coefficient matrices and innovation covariance matrices is adequate but unproductive
If You Understand the VAR, You Understand Everything

But it’s hard to understand the VAR.

– Pairwise Granger-causality is inadequate

– 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
- Generally unavailable in high dimensions and arguably undesirable
VD’s Require Identification

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

Cholesky factor identification
– Orthogonalizes but requires ordering
VD’s Require Identification

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

Cholesky factor identification
– Orthogonalizes but requires ordering

But perhaps Cholesky ordering doesn’t matter much except for very small $H$’s
Intricate theory identification
– Generally unavailable in high dimensions and arguably undesirable

Cholesky factor identification
– Orthogonalizes but requires ordering

But perhaps Cholesky ordering doesn’t matter much except for very small $H$’s

But perhaps there is a natural Cholesky ordering in many finance applications (e.g., put large-cap institutions first)
And, Perhaps Graphical Modeling can Help

Cholesky Orthogonalization:

$$(I - \Phi_1 L - \ldots - \Phi_p L^p) y_t = P \nu_t$$

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

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

Moving-average representation:

$$y_t = (I + \Theta_1 L + \Theta_2 L^2 + \ldots) P \nu_t$$

$$= P \nu_t + \Theta_1 P \nu_{t-1} + \ldots$$
Recursive Structural System

Structural Simultaneous-Equations Model (SEM):

\[ Ay_t = \Phi_1 y_{t-1} + \ldots + \Phi_p y_{t-p} + \varepsilon_t \]

\[ \varepsilon_t \sim (0, \Sigma) \]
Recursive Structural System

Structural Simultaneous-Equations Model (SEM):

\[ Ay_t = \Phi_1 y_{t-1} + \ldots + \Phi_p y_{t-p} + \varepsilon_t \]
\[ \varepsilon_t \sim (0, \Sigma) \]

Recursive SEM: A triangular and \( \Sigma \) diagonal

\[ Ty_t = \Phi_1 y_{t-1} + \ldots + \Phi_p y_{t-p} + \varepsilon_t \]
\[ \varepsilon_t \sim (0, D) \]
Recursive Structural System

Structural Simultaneous-Equations Model (SEM):

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

\[ \varepsilon_t \sim (0, \Sigma) \]

Recursive SEM: A triangular and \( \Sigma \) 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:

\[ y_t = T^{-1} \varepsilon_t + ... \]
DAG’s, Bayes Nets, and all That...

Start with:

http://www.nber.org/papers/w19453
DAG’s, Bayes Nets, and all That...

Start with:
http://www.nber.org/papers/w19453

Then back up and read or re-read:
DAG’s, Bayes Nets, and all That...

Start with:
http://www.nber.org/papers/w19453

Then back up and read or re-read:

Potentially good primers:
Hjsgaard, S. ”Graphical Models and Bayesian Networks with R”
www.people.math.aau.dk/~sorenh/misc/2014-useR-GMBN/
– Causal relationships are represented by a graph $G$, where nodes correspond to variables.
– Causal relationships are represented by a graph $G$, where nodes correspond to variables.

– Nodes are connected by arrows that represent causal influences between variables.
– 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 $V$. 
– 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.
– 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
Key DAG Insights

- DAG environment is recursive system environment
- DAG causality is about determining recursive system ordering
Key DAG Insights

- DAG environment is recursive system environment
- DAG causality is about determining recursive system ordering
- Conditional independence: $y$ is independent of $x$ conditional on $z$ if and only if $\Pr(y \cap x \mid z) = \Pr(y \mid z) \Pr(x \mid z)$. 
Key DAG Insights

- DAG environment is recursive system environment

- DAG causality is about determining recursive system ordering

- Conditional independence: $y$ is independent of $x$ conditional on $z$ if and only if $\Pr(y \cap x \mid z) = \Pr(y \mid z) \Pr(x \mid z)$.

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

Potentially a very big step:

*Local Markov / conditional independence can be used to help determine recursive orderings. “Likelihood information”*

and

*Bayesian analysis can blend likelihood and prior information.*
(3) Learn From the Identified Estimated VAR

Key theme: Staring at massive $D$ is just as hopeless as staring at massive coefficient matrices

But graph-theoretic tools come to the rescue
How?

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!*
Part I: Graph-Theoretic $D$ Distillation
Interpret $D$ as a Network Adjacency Matrix
Distill Using the Degree Distribution

### Variance Decomposition / Connectedness Table

<table>
<thead>
<tr>
<th></th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>...</th>
<th>$x_N$</th>
<th>From Others</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_1$</td>
<td>$d_{11}^H$</td>
<td>$d_{12}^H$</td>
<td>...</td>
<td>$d_{1N}^H$</td>
<td>$\sum_{j \neq 1} d_{1j}^H$</td>
</tr>
<tr>
<td>$x_2$</td>
<td>$d_{21}^H$</td>
<td>$d_{22}^H$</td>
<td>...</td>
<td>$d_{2N}^H$</td>
<td>$\sum_{j \neq 2} d_{2j}^H$</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>$x_N$</td>
<td>$d_{N1}^H$</td>
<td>$d_{N2}^H$</td>
<td>...</td>
<td>$d_{NN}^H$</td>
<td>$\sum_{j \neq N} d_{Nj}^H$</td>
</tr>
</tbody>
</table>

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 connectedness “from,” $C_{i \leftarrow \bullet}^H = \sum_{j=1}^{N} d_{ij}^H$: “from-degrees”

Total directional connectedness “to,” $C_{\bullet \leftarrow j}^H = \sum_{i=1}^{N} d_{ij}^H$: “to-degrees”

Systemwide connectedness, $C^H = \frac{1}{N} \sum_{i,j=1}^{N} d_{ij}^H$: mean degree
Opportunities with Graph-Theoretic $D$ Distillation

- Examine aspects of the degree distribution _across $H$_
Opportunities with Graph-Theoretic D Distillation

- Examine aspects of the degree distribution across $H$
  - Other Connectedness Measures

- Multi-step connectedness

- $k$-step connectedness,

- $\infty$-step connectedness ("eigenvalue centrality")

Examine second smallest eigenvalue $\lambda_2$ of $L = M - A$ ($M$ is a diagonal matrix containing the node degrees) $A$ (is the adjacency matrix).

- But is any of this necessary/desirable for us?
  - Should we not simply vary $H$?
Opportunities with Graph-Theoretic $D$ Distillation

– 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,
Opportunities with Graph-Theoretic $D$ Distillation

– 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) $\infty$-step connectedness ("eigenvalue centrality")
   Examine second smallest eigenvalue $\lambda_2$ of $L = M - A$
   ($M$ is a diagonal matrix containing the node degrees)
   ($A$ is the adjacency matrix.)
Opportunities with Graph-Theoretic $D$ Distillation

- 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) $\infty$-step connectedness ("eigenvalue centrality")
  Examine second smallest eigenvalue $\lambda_2$ 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?
  Should we not simply vary $H$?
Part II: Graph-Theoretic $D$ Visualization
Spring Graph
Spring Graph Detail

- Node size: Asset size
- 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

- Examine graphs across $H$
  (Multiple comparisons or a single animation across $H$)
Opportunities With $D$ Graphs

– Examine graphs across $H$
  (Multiple comparisons or a single animation across $H$)

– Spring graphs have a flip-flop issue
Opportunities With $D$ Graphs

- Examine graphs across $H$
  (Multiple comparisons or a single animation across $H$)

  - Spring graphs have a flip-flop issue

  - There does not exist a natural gradation across colors from “cool” to “hot”
    (So use layering rather than colors)
Opportunities With $D$ Graphs

– Examine graphs across $H$
  (Multiple comparisons or a single animation across $H$)

  – Spring graphs have a flip-flop issue

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

– Time-varying coefficients and dynamic network graphs
  (Rolling estimation, explicitly time-varying coefficients, etc.)
Opportunities With $D$ Graphs

– Examine graphs across $H$
  (Multiple comparisons or a single animation across $H$)

– Spring graphs have a flip-flop issue

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

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

– Animate over time for fixed $H$,
  for a variety of $H$
Concluding Perspective

Old view: VAR’s unworkable in high dimensions
(Actually no one even thought about high dimensions)
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) Regularization for estimation

(2) Bayes nets for Cholesky identification

(3) Network graphs for visual understanding
Concluding Perspective

Old view: \textit{VAR’s} unworkable in high dimensions (Actually no one even \textit{thought} about high dimensions)

New view: \textit{VAR’s} \textit{are} workable in high dimensions

(1) Regularization for estimation

(2) Bayes nets for Cholesky identification

(3) Network graphs for visual understanding

Still \textit{VAR’s, but:}

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