Regularized Estimation of High-dimensional Factor-Augmented Vector Autoregressive (FAVAR) Models

George Michailidis

University of Florida
Department of Statistics and the Informatics Institute

February 22, 2018
An afternoon excursion in VAR-land
Two popular frameworks for multivariate temporally dependent data

1. **Vector Autoregressive Models (VAR)**
   - \( X_t = AX_{t-1} + E_t \) captures lead-lag cross-correlations between a set of time series, where \( E_t \) is a noise/shock process
   - Workhorse model in macroeconomics, with many other applications areas such as financial econometrics, and more recently functional genomics and neuroscience
   - Also, standard model for closed loop control in linear dynamical systems \( X_t = AX_{t-1} + BU_{t-1} + E_t \), when coupled with a cost/reward function (e.g. Linear Quadratic Regulator)

2. **Dynamic Factor Models (DFM)**
   - \( X_t = \Lambda F_t + E_t \), where \( F_t \) is a small dimension factor process that follows for example auto-regressive dynamics - \( F_t = \Phi F_{t-1} + W_t \)
   - Popular model in economics and finance, since capable of summarizing contemporaneous correlations

Strong focus on forecasting and especially in economics in conjunction with policy interventions (impulse response functions).
A Brief Overview of VAR modeling - 1

The VAR model:

- $p$-dimensional, discrete time, stationary process $X^t = \{X_1^t, \ldots, X_p^t\}$

$$X^t = A_1 X^{t-1} + \ldots + A_d X^{t-d} + \epsilon^t, \quad \epsilon^t \sim_{i.i.d} N(0, \Sigma_\epsilon). \quad (1)$$

- $A_1, \ldots, A_d$ : $p \times p$ transition matrices (solid, directed edges).
- $\Sigma_{\epsilon}^{-1}$: contemporaneous dependence (dotted, undirected edges).
- stability: Eigenvalues of $A(z) := I_p - \sum_{t=1}^d A_t z^t$ outside $\{z \in \mathbb{C}, |z| \leq 1\}$.

Key challenge: parameter space grows as $O(dp^2)$. 
A Brief Overview of VAR modeling - II

1. **Low-dimensional regime -** $p$ fixed, $T \to \infty$
   - Frequentist approach: map the problem to a regression problem and use least squares or maximum likelihood methods to estimate the transition matrix $A_j$ and covariance $\Sigma_\epsilon$ parameters; also standard $\sqrt{T}$ asymptotics (Lutkepohl, 2005).
   - Bayesian approaches: various prior distributions on the $A_j$’s, including the famous Minnesota prior (Litterman, 1977; Doan, Litterman, Sims, 1984), and its variants (Korobilis, 2013), together with associated computational algorithms.

2. **High-dimensional regime -** $(p, T) \to \infty$
   - Frequentist approach: leverages assumption of (structured) sparsity in both $A_j$ and $\Sigma_\epsilon$ and associated regularized (lasso and its variants) estimation procedures.
   - Theoretical work on sparse estimates.
   - Bayesian approach: use of normal-inverse Wishart prior distributions (Banbura, Giannone, Reichlin, 2010) that gives posterior means akin to ridge estimates.
   - On the theoretical front, for both normal and hierarchical normal-mixture priors (includes Bayesian group lasso t-distributions, etc.) posterior consistency of the $A_j$ estimates was recently established in the $p = o(T)$ regime (Ghosh, Khare, Michailidis, 2018).
Consistency of sparse VAR estimates

Under certain regularity conditions

- VAR process stable;
- restricted eigenvalue (strong convexity) condition that regulates the behavior of the minimum eigenvalue of $X'X/T$ over an appropriately defined cone for the elements of $A_j$'s along the directions of their sparse support;
- deviation condition that regulates the behavior of $\|X' E\|_\infty$;

it is established (see Basu and Michailidis, 2015) that

$$\sum_{h=1}^{d} \left\| \hat{A}_h - A_h \right\| \leq \phi(A_t, \Sigma_\epsilon) \left( \sqrt{k \left( \log dp^2 \right)/T} \right).$$

Further, for Gaussian stable VAR models, the restricted eigenvalue and deviation conditions hold with high probability.
Comments on the Consistency Results

- Estimation error has two components:
  1. \( \phi(A^t, \Sigma_e) \) large \( \Leftrightarrow \) the max eigenvalue \( \mathcal{M}(f_X) \) of the spectral density of \( X_t \) is large, the min eigenvalue \( \mathcal{m}(f_X) \) of the spectral density of \( X_t \) is small
  2. \( \sqrt{k \log dp^2 / T} \): Estimation error for independent data

- Estimation error same as i.i.d. data, modulo a price for temporal dependence
Extensions to the basic sparse VAR framework

- Other penalties, such as group, sparse group, etc. (Basu, Shojaie, Michailidis, 2015; Melnyk, Banerjee, 2016)
- Lag selection through hierarchical VAR models (Nicholson, Matteson, Bien, 2017)
- VAR models with local dependence constraints (Schweinberger, Babkin, Ensor, 2017)
- VAR-X models with exogenous high-dimensional $Z$ variables - $X_t = AX_{t-1} + BZ_t + E_t$ (Lin, Michailidis, 2017)
- Change point problems for VAR models (Safikhani, Shojaie, 2017)
- Joint estimation of related VAR models (Skripnikov, Michailidis, 2017)
- VAR models for count data (Hall, Raskutti, Willett, 2016)
- Finite sample bounds for non-stable VAR models with “heavy”-tailed errors (Faradonbeh, Tewari, Michailidis, 2017)
An application to macroeconomics and forecasting

VAR modeling use in macroeconomic modeling was pioneered by Sims (1972, 1980) and is widely used by policy makers.

A key issue in economics is that if the VAR model size is “small”, it can lead to estimates whose signs violate basic economic relationships (e.g. the prize puzzle by Christiano, Eichenbaum, Evans (1999) that postulates that unexpected tightening monetary policy leads to increases in the economy’s price level).

Hence, economists (for a review see Stock, Watson, 2016) have advocated for large VAR models, but until recently their estimation was not feasible.

We consider the large VAR model (Banbura, Giannone, Recihlin, 2010) of 134 key US macroeconomic indicators for the period January 1973 - December 2014. Variables were transformed according to the recommendations in Stock, Watson (2005).
Forecasting performance

Used various priors, as well as the standard Minnesota prior and did forecasting for horizons $h = 1, 6, 12$. The memory of the VAR model was selected based on the Deviance Information Criterion to $d = 6$.

Bayesian group lasso and normal-inverse Wishart prior exhibited a better performance than other hierarchical priors and all of them outperformed the Minnesota prior, but performance deteriorates at longer horizons.

Forecasting performance of key macro indicators using a normal-inverse Wishart prior:

<table>
<thead>
<tr>
<th></th>
<th>$h = 1$</th>
<th>$h = 6$</th>
<th>$h = 12$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Price Index</td>
<td>0.44</td>
<td>0.38</td>
<td>0.42</td>
</tr>
<tr>
<td>Total Nonfarm</td>
<td>0.96</td>
<td>0.48</td>
<td>0.73</td>
</tr>
<tr>
<td>Employment</td>
<td>0.68</td>
<td>0.68</td>
<td>0.72</td>
</tr>
<tr>
<td>Federal Funds Rate</td>
<td>0.68</td>
<td>0.68</td>
<td>0.72</td>
</tr>
</tbody>
</table>
An Application of VAR-X to a key macroeconomic question

What is the relationship between the stock market and the employment index (or other key macro indicators)?

Note that until the great recession of 2008, standard macroeconomic Dynamic General Stochastic Equilibrium models (both Real Business Cycle and New-Keynesian flavors) did not incorporate the stock market as a driver of real economic activity and the business cycle.

Early work by Phelps (1999) and Fitoussi et al (2000) and more recently by Farmer (2015) advocated the impact of the stock market through a wealth effect mechanism.

Farmer (2015) tested the hypothesis using a bivariate VAR model of the S&P100 index and the employment index, while controlling for other macro indicators. The test for Granger causality of the S&P100 index on the employment index was highly significant.
An Application of VAR-X to a key macroeconomic question

We re-examined this question in Lin, Michailidis (2017) by considering a high-dimensional VAR-X model of the form

\[ Z_t = BX_{t-1} + CZ_{t-1} + E_t, \]

where \( E_t \sim N(0, \Sigma_e) \)

\( Z_t \) contains the log-returns of the S&P 100, and \( X_t \) key macro indicators used in Stock, Watson, 2005
Estimation of a high-dimensional VAR-X model

\[
(\hat{B}, \hat{C}, \hat{\Omega}_\epsilon) = \arg\min_{B, C, \Omega_\epsilon} \left\{ \text{tr} \left[ \Omega_\epsilon (Z_T - XB' - ZC')' (Z_T - XB' - ZC') / T \right] - \log \det \Omega_\epsilon \right.
\]
\[
+ \lambda_B \mathcal{R}(B) + \lambda_C \|C\|_1 + \rho_v \|\Omega_\epsilon\|_{1, \text{off}} \bigg\},
\]

where the regularizer \( \mathcal{R}(B) = \|B\|_1 \) or \( \|B\|_* \) (sparse or low rank).

Solve the optimization:

1. Initialize with \( \hat{\Omega}_\epsilon^{(0)} = I_{p_2} \), then

\[
(\hat{B}^{(0)}, \hat{C}^{(0)}) = \arg\min_{B, C} \left\{ \frac{1}{T} \|ZT - XB' - ZC'\|_F^2 + \lambda_B \mathcal{R}(B) + \lambda_C \|C\|_1 \right\}.
\]

2. Update \( \hat{\Omega}_\epsilon^{(k)} \) by glasso on the residuals with the plug-in \( (\hat{B}^{(k-1)}, \hat{C}^{(k-1)}) \).

3. For fixed \( \hat{\Omega}_\epsilon^{(k)} \), update \( (\hat{B}, \hat{C}) \) alternately until convergence (via Lasso or singular value thresholding).

4. Iterate between 2 and 3 until convergence.
Technical Issues

The objective function is not jointly convex in its parameters, but biconvex. Same issue happens for the simple VAR model with a sparse precision matrix.

Convergence to a stationary point is guaranteed, as long as estimates from all iterations lie within a ball around the true value of the parameters, with the radius of the ball upper bounded by a universal constant that only depends on model dimensions and sample size (Lin, Basu, Banerjee, Michailidis, 2016, Theorem 4.1).

The latter condition is satisfied upon establishing consistency properties of the estimates (technical details in Lin, Michailidis, 2017)
Testing for “Group” Granger Causality

\[ Z_t = BX_{t-1} + CZ_{t-1} + \nu_t. \]

- A testing problem: \( H_0: B = 0 \) vs \( H_1: B \neq 0 \).
- Under two possible structural assumptions: \( B \) is low rank, or \( B \) is sparse.
  - For a low rank \( B \), test \( H_0: B = 0 \) is a special case of testing \( H_0: \text{rank}(B) \leq r \), where \( r < \min(p_1, p_2) \) – can be solved leveraging the idea of canonical correlation (Anderson, 2002).
  - For a sparse \( B \), test \( H_0: B = 0 \) against its sparse alternative, i.e., \( H_1: B \neq 0 \) and is (strongly) sparse – can be solved leveraging the idea of high criticism (Arias-Castro, 2011)
Testing for “Group” Granger Causality

Table 1: Summary for estimated $B$ and $C$ within different periods.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$p$-value for testing $H_0 : B = 0$</td>
<td>0.075</td>
<td>&lt; 0.0001</td>
<td>0.044</td>
</tr>
<tr>
<td>Sparsity level of $\hat{C}$</td>
<td>0.06</td>
<td>0.25</td>
<td>0.06</td>
</tr>
<tr>
<td>Spectral radius of $\hat{C}$</td>
<td>0.35</td>
<td>0.76</td>
<td>0.40</td>
</tr>
</tbody>
</table>
Issues with high-dimensional VAR models

The key challenge is that the parameter space grows as $O(dp^2)$ while the stability constraint needs to be respected.

Sparsity or other regularization approaches help, nevertheless they exhibit natural limitations for modeling very large sets of time series.

An alternative is to use Dynamic Factor Models, but then one focuses on contemporaneous correlations, as opposed to lead-lag cross-correlations.

What if one wants to stay within the VAR world?
Factor-augmented autoregressive models

- Motivation: model the dynamics of a panel of macroeconomic indicators $Y_t$ with VAR, then measure the effects of structural shocks.
  - Some of the underlying economic drivers are not directly observable/measurable hence can not be included in the VAR analysis (e.g. the output gap, the Wicksellian natural rate of interest) – model mis-specification.
  - The set of variables included in the VAR system is often small – underutilized information set.

- FAVAR: use a small set of estimated factors to summarize the large amount of information about the economy, originally introduced in Bernanke, Boivin, Giannone (2005):

  core variables of interest $\leftarrow X_t$

  large information set $\leftarrow Y_t = \Lambda F_t + \Gamma X_t + e_t$

  augmented system $\leftarrow \begin{bmatrix} F_t \\ X_t \end{bmatrix} \sim \text{VAR}(d)$

  $F_t$: economic activity, price pressure, credit condition, etc.
Previous work.

- **Limitation:** $X_t$ is small.
- **Theoretical work:** (Bai, Li, Lu, 2016)
  - established the asymptotic normality of the parameters of interest;
  - estimation procedure and theoretical results rely on the assumption that $X'X$ being invertible;
  - fixed dimension, fixed design.
- **Applied work:**
  - Bernanke, Boivin, Elasz (2005): VAR on core economic variables (IPI, CPI, FFR)
  - Lombardi, Osbat, Schnatz (2010), Forni, Gambetti, Sala (2014), etc.
FAVAR in high-dimensions.

- Extending the low-dimensional FAVAR model to high dimensions, allowing to study cross-dependencies for a large set of $X_t$.
  - size of $X_t$ grows with the available sample size;
  - estimation does not rely on invertibility;
  - model identifiability, its interpretation and link to other high-dim literature;
  - high-probability finite-sample bound;

- Application to commodity price data – uncovering the commodity price linkages, while taking into consideration the influence of the state of the global economy.
Content

A Brief Overview of VAR models.

Model setup and Estimation.

Theoretical Properties.

Real Data Application.
Model setup.

- Let $F_t \in \mathbb{R}^{p_1}$ denote the latent block and $X_t \in \mathbb{R}^{p_2}$ denote the observed block of time series, whose dynamics are described by a VAR($d$) model:

\[
\begin{bmatrix}
F_t \\
X_t
\end{bmatrix} = A^{(1)} \begin{bmatrix} F_{t-1} \\ X_{t-1} \end{bmatrix} + \cdots + A^{(d)} \begin{bmatrix} F_{t-d} \\ X_{t-d} \end{bmatrix} + \begin{bmatrix} w_t^F \\ w_t^X \end{bmatrix}. \tag{2}
\]

- The large observed informational time series $Y_t \in \mathbb{R}^q$ is related to $X_t$ and $F_t$ contemporaneously through a linear model

\[
Y_t = \Lambda F_t + \Gamma X_t + e_t. \tag{3}
\]

- Model assumptions:
  - Let $T$ be the number of time points for $X_t$ and $Y_t$: $\{(x_t, y_t)\}_{t=1}^T$, $p_1 < T$ is fixed and small, $p_2$ and $q$ can grow with $T$ and be potentially larger than $T$.
  - Transition matrices $\{A^{(i)}\}_{i=1}^d$ are assumed sparse, the coefficient matrix $\Gamma$ is sparse, while the factor loadings $\Lambda$ can be dense.
  - Mean-zero Gaussian noise processes with covariance matrices $\Sigma_w^F$, $\Sigma_w^X$ and $\Sigma_e$.
- Goal: recover $\{A^{(i)}\}_{i=1}^d$, and the coefficient matrices $\Gamma$ and $\Lambda$. 

George Michailidis
Model identifiability.

- For matrix $Q_1 \in \mathbb{R}^{p_1 \times p_1}$ (invertible) and $Q_2 \in \mathbb{R}^{p_1 \times p_2}$:
  \[ Y_t = \Lambda F_t + \Gamma X_t + e_t \equiv \Lambda Q_1 (Q_1^{-1} F_t - Q_1^{-1} Q_2 X_t) + (\Gamma + \Lambda Q_2) X_t + e_t. \]

- $(\Lambda, F_t, \Gamma)$ and $(\tilde{\Lambda}, \tilde{F}_t, \tilde{\Gamma})$ yield observationally equivalent models.
- $p_1^2 + p_1 p_2$ restrictions are required for exact identification of $\Lambda$, $\Gamma$ and $F_t$.

- In the low-dim setting:
  - Bernanke, Boivin, Giannone (2005): $\Lambda = [I_{p_1 \times p_1}]$, $\Gamma_{1:p_1,.} = 0$ – too restrictive;
  - Bai, Li, Lu (2016): $\text{Cov}(w_tF, w_tX) = 0$ plus FA-type identification restrictions – can not be easily incorporated in the estimation in the high-dim setting;

- Remedy for the high-dim setting:
  - Identify the factor space $\Lambda F_t$ and $\Gamma X_t$ by solving an optimization problem – a “low rank + sparse”-type decomposition;
  - Recover $F_t$ from the estimated factor space through FA-type identification restrictions:
    \[ \Lambda = [I_{p_1 \times p_1}], \quad F_t \text{ unrestricted} \quad \text{(IR)} \]
    – a total number of $p_1^2$ restrictions; $F_t$ enters the VAR dynamics unrestricted.
Estimation.

- Without loss of generality, focus on the VAR(1) case:
  - **VAR equation**
    \[
    \begin{bmatrix}
    F_t \\
    X_t \\
    \end{bmatrix} =
    \begin{bmatrix}
    A_{11} & A_{12} \\
    A_{21} & A_{22} \\
    \end{bmatrix}
    \begin{bmatrix}
    F_{t-1} \\
    X_{t-1} \\
    \end{bmatrix} +
    \begin{bmatrix}
    w_t^F \\
    w_t^X \\
    \end{bmatrix}
    \]

  - **Calibration equation**
    \[
    Y_t = \Lambda F_t + \Gamma X_t + e_t.
    \]

- **Estimation steps:**
  1. Estimate the factor structure from the calibration equation and subsequently using the identifiability constraints extract the factors;
  2. Use the estimated factors as input time series to estimate the transition matrices of the joint \((\hat{F}_t, X_t)\) VAR system.
Estimation.

- Data: \( X = [x_1', \ldots, x_T']', Y = [y_1', \ldots, y_T']', \) centered and observed. Samples of the factors \( F = [f_1', \ldots, f_T']' \) from process \( \{F_t\} \) are latent.

\[
Y = F\Lambda' + X\Gamma' + E =: \Theta + X\Gamma' + E, \quad \Theta \in \mathbb{R}^{T \times q} \text{ (factor hyperplane)};
\]

\[
\begin{bmatrix}
F_T \\
X_T
\end{bmatrix} = 
\begin{bmatrix}
F_{T-1} \\
X_{T-1}
\end{bmatrix} 
\begin{bmatrix}
A_{11} & A_{12} \\
A_{21} & A_{22}
\end{bmatrix}' + W.
\]

1. Estimate \( \Theta \) and \( \Gamma \), then extract \( F \) based on \( \hat{\Theta} \):

\[
(\hat{\Theta}, \hat{\Gamma}) := \arg\min_{\Theta, \Gamma} \left\{ \frac{1}{2T} \left\| Y - \Theta - X\Gamma' \right\|_F^2 + \lambda \Gamma \| \Gamma \|_1 \right\}, \quad \text{s.t.} \quad \text{rank}(\Theta) \leq r. \quad (v0)
\]

Under IR, \( \hat{F} = \hat{\Theta}_{1:r} \).

2. Estimate \( \hat{A} \) using \( \hat{Z} := [\hat{F}, X] \) based on penalized LS:

\[
\hat{A} = \arg\min_{A} \left\{ \frac{1}{2T} \left\| \hat{Z}_T - \hat{Z}_{T-1} A' \right\|_F^2 + \lambda_A \| A \|_1 \right\}.
\]

- separately obtain rows of \( \hat{A} \) via separate Lasso regressions.
Implementation & tuning.

- Estimation of $\Theta$ and $\Gamma$ based on an alternating minimization algorithm:
  - For fixed $\hat{\Theta}^{(k-1)}$, update $\hat{\Gamma}^{(k)}$ so that each row $i$ solves a Lasso problem with the response being $(Y - \hat{\Theta}^{(k-1)})_i$ and the design matrix being $X$.
  - For fixed $\hat{\Gamma}^{(k)}$, $\hat{\Theta}^{(k)}$ is the best rank-$r$ approximation of $Y - X\hat{\Gamma}^{(k)'}$ – singular value thresholding (SVT).
  - Terminate when the value of the objective function stops descending (w.r.t. some tolerance).

- Estimation of $\hat{A}$: each row separately does Lasso regression with the response being $(\hat{Z}_T)_i$ and the design matrix being $\hat{Z}_{T-1}$ (recall $\hat{Z} := [\hat{F}, X]$).

- Choice of $(r, \lambda_\Gamma)$: panel information criterion (Ando, Bai, 2015).

- Choice of $\lambda_A$: Bayesian information criterion (BIC).
Remarks on the calibration equation estimation problem.

\[ Y_t = \Lambda F_t + \Gamma X_t + e_t. \]

- Theoretical formulation vs. empirical implementation of the optimization problem.
- The “low rank plus sparse”-type decomposition: another view.
- Compactness and identifiability.
Step 1 recovery: theoretical vs empirical.

- \( \mathbf{Y} = \mathbf{F}\Lambda' + \mathbf{X}\Gamma' + \mathbf{E} = \Theta + \mathbf{X}\Gamma' + \mathbf{E}, \) \( \Theta \) is low rank, \( \Gamma \) is sparse.

- The most primitive formulation with rank and \( \ell_1 \)-norm constraints:

\[
\min_{\Theta, \Gamma} \left\{ \frac{1}{2T} \left\| \mathbf{Y} - \Theta - \mathbf{X}\Gamma' \right\|_F^2 \right\}, \quad \text{s.t.} \ \text{rank}(\Theta) \leq r, \ \|\Gamma\|_1 \leq \phi. \quad (v00)
\]

  - Additional compactness-type constraint on \( \Theta \) is required for theoretical guarantees on the global optima.
  - Non-convexity, no guarantee on algorithmic convergence to a global optimum;

- **Compact** constraint formulation; introduce an additional constraint on the nuclear-norm ball \( \mathbb{B}_T(\phi) := \{ \Theta \in \mathbb{R}^{T \times q} | \|\Theta/\sqrt{T}\|_* \leq \phi \} \):

\[
\min_{\Theta, \Gamma} \left\{ \frac{1}{2T} \left\| \mathbf{Y} - \Theta - \mathbf{X}\Gamma' \right\|_F^2 + \lambda_\Gamma \|\Gamma\|_1 \right\}, \quad \text{s.t.} \ \text{rank}(\Theta) \leq r, \ \Theta \in \mathbb{B}_T(\phi). \quad (v1)
\]

- **Compact** constraint and **convexified** formulation:

\[
\min_{\Theta, \Gamma} \left\{ \frac{1}{2T} \left\| \mathbf{Y} - \Theta - \mathbf{X}\Gamma' \right\|_F^2 + \lambda_\Gamma \|\Gamma\|_1 + \lambda_\Theta \|\Theta/\sqrt{T}\|_* \right\}, \quad \text{s.t.} \ \Theta \in \mathbb{B}_T(\phi). \quad (v2)
\]
Step 1 recovery: theoretical vs empirical (ctd)

- With compactness:

\[
\min_{\Theta, \Gamma} \left\{ \frac{1}{2T} \| Y - \Theta - X\Gamma' \|_F^2 + \lambda_\Gamma \| \Gamma \|_1 \right\}, \quad \text{s.t. rank}(\Theta) \leq r, \ \Theta \in B_T(\phi). \quad (v1)
\]

\[
\min_{\Theta, \Gamma} \left\{ \frac{1}{2T} \| Y - \Theta - X\Gamma' \|_F^2 + \lambda_\Gamma \| \Gamma \|_1 + \lambda_\Theta \| \Theta / \sqrt{T} \|_* \right\}, \quad \text{s.t. } \Theta \in B_T(\phi). \quad (v2)
\]

Global optima of (v1) and (v2) have identical finite-sample bounds (rate).

- (v1) is still non-convex, no guarantee on convergence to a global optimum.
- (v2) has both theoretical and algorithmic guarantees.
  - Solved by the alternate minimization between \( \Gamma \) (separate Lasso regressions) and \( \Theta \) (composite gradient descent, Nesterov, 2007);
  - Within each iteration, Lasso steps plus \((2 \times)\) soft SVT steps.

- (v0) is the regularized version of (v00), and its actual implementation is an empirical relaxation of that of (v2), by substituting the soft Singular Value Thresholding (SVT) to a hard SVT.
“Low rank plus sparse”-type decomposition: another view.

- The joint distribution of the three blocks $Y_t$, $F_t$ and $X_t$, parametrized by the concentration matrix:

\[
\begin{bmatrix}
Y_t \\
F_t \\
X_t
\end{bmatrix} \sim \mathcal{N}
\left(0,
\begin{bmatrix}
\Omega_{YY} & \Omega_{YF} & \Omega_{YX} \\
\Omega_{FY} & \Omega_{FF} & \Omega_{FX} \\
\Omega_{XY} & \Omega_{XF} & \Omega_{XX}
\end{bmatrix}^{-1}
\right).
\]

The conditional mean structure links the observed and the latent block:

\[
Y_t = (-\Omega_{YY}^{-1}\Omega_{YF}F_t) + (-\Omega_{YY}^{-1}\Omega_{YX}X_t) + \epsilon_t.
\]

- A similar formulation in Chandrasekaran et al (2012):

\[
(\Sigma_O^{-1}) = K_O - K_{O,H}(K_H)^{-1}K_{H,O}
\]

The conditional covariance structure links the observed and the latent block via Schur complement.
Compactness and identifiability.

- Compactness in essence limits the interaction between the two spaces, respectively spanned by $F_t$ and $X_t$:
  - Stems from the consideration that the two spaces can not be too similar – we require an “incoherence” type of constraint of some form;
  - This usually occurs in the form of some dual norm – from Hölder’s inequality.

- With the compactness constraint, this leads to an approximate identification, but not an exact one.

- In high-dim settings, exact identification often requires excessively stringent conditions:
  - E.g., Chandrasekaran et al (2012): tangent spaces at the low-rank and sparse matrices respectively only intersect at the origin (transversality).
  - E.g., one possibility for this problem: $\text{Cov}(X_t, F_t) = O$. 
Content

A Brief Overview of VAR models.

Model setup and Estimation.

Theoretical Properties.

Real Data Application.
A road map.

- **Deterministic → random; Stage 1 error bound → Stage 2 error bound.**
- **Deterministic analysis:**
  - Stage 1: under certain regularity conditions on the observed samples, the estimates of the parameters and the factor based on the calibration equation have bounded errors;
  - Stage 2: under regularity conditions on the with-error-variables, the transition matrix recovery has bounded error – the regularity conditions depend on both the realizations of the observed process $X_t$, and the magnitude of the error.
- **Random analysis:**
  - Stage 1: for random realizations from the underlying processes, the required regularity conditions are satisfied w.h.p. – high probability error bounds for $\hat{\Theta}$, $\hat{\Gamma}$ and $\hat{F}$.
  - Stage 2: through conditioning arguments, w.h.p the regularity conditions on the with-error-variables are satisfied, provided that the original sequence is “sufficiently regular” – additional regularity on the $\{Z_t\}$ spectrum is required, compared with the case where estimation is conducted directly on the true observed samples.
Calibration equation estimation.

Require:

- Restricted Strong Convexity assumptions on the design matrix $X$
- Bounds on the max/min eigenvalues of $\Sigma_e$ and on the extremes of the spectral density of the $X$ process, i.e., $m(f_X)$ and $M(f_X)$
- Appropriate selection of the penalty parameter in accordance with a deviation bound and other requirements associated with the maximum eigenvalues of $\Sigma_e$ and the extremes of the spectral density of the $X_t$ process

Then, for sample size $T \gtrsim q$, w.h.p.

$$\|\Delta_\Gamma\|_F^2 + \|\Delta_\Theta/\sqrt{T}\|_F^2 \lesssim C(m(f_X), M(f_X), \Lambda_{\text{max}}(\Sigma_e)) \cdot \kappa(s_{\Gamma^*}, p_1, r, \phi^2),$$

where

$C()$ is some function of $m(f_X)$, $M(f_X)$ and $\Lambda_{\text{max}}(\Sigma_e)$,

$\kappa()$ is some linear function of $s_{\Gamma^*}$, $p_1$, $r$ and $\phi^2$. 
VAR system estimation.

- Deterministic realizations of $\hat{Z}_{T-1} := [\hat{F}_{T-1}, \hat{X}_{T-1}]$ satisfy a Restricted Strong Convexity condition.
- Conditions on the max/min eigenvalues of the spectral density of the $Z$ process.
- Appropriate choice of tuning parameter $\lambda_A$.

Then,

$$\|\Delta_A\|^2_F \leq \hat{C}(\mathcal{E}_1, m(f_Z), \mathcal{M}(f_Z)) \cdot \hat{\kappa}(s_{A^*}),$$

where $\mathcal{E}_1$ is the (high-prob) upper bound of the stage 1 estimation error:

$$\mathcal{E}_1 := C(m(f_X), \mathcal{M}(f_X), \Lambda_{\max}(\Sigma_e)) \cdot \kappa(s_{\Gamma^*}, p_1, r, \phi^2).$$
Content

A Brief Overview of VAR models.

Model setup and Estimation.

Theoretical Properties.

Real Data Application.
Application to Commodity Prices Interlinkages.

Some background: commodity prices are driven by several factors including: (I) the state of the global macroeconomy that manifests itself as direct demand for commodities, (II) interest rates that impact the opportunity cost for holding inventories, (III) performance of other assets that impact commodities prices through portfolio compositions (see work by Frankel)

- $X_t$: 16 commodity prices, across Metal, Energy (oil) and Agricultural.
- $Y_t$: 54 macroeconomic indicators from different regions including US, UK, EU, CN and JP, primarily falling into the following categories: Output & Income (industrial production), Labor Market (unemployment), Money & Credit (M2), Interest & Exchange Rate, Prices (CPI), Stock Market (composite indices).
- Data span January 2001 to December 2016 at a monthly frequency, with transformation of the variables applied (according to Stock, Watson, 2005 recommendations, to ensure stationarity of the resulting series.
- Based on other empirical work (Stock, Watson, 2017) we consider three sub-periods:
Pre-crisis: factor composition.

- **F1**: Equity Return, Bond Return and Economic Output.
- **F2**: Price Index and Trade.
- **F3**: Effective Exchange Rate and Bond Return.
- **F4**: Bond Return.
Pre-crisis: commodity vector autoregression.
Crisis: factor composition.

- F1: Equity Return, Bond Return, Unemployment and Money.
- F2: Bond Return, Equity Return and Trade.
- F3: Price Index and Economic Output.
Crisis: commodity vector autoregression.

Transition matrix: lag 1

Transition matrix: lag 2
Post-crisis: factor composition.

- **F1**: Equity Return, Fed Fund Rate, Effective Exchange Rate, Trade.
- **F2**: Money, Economic Output.
- **F3**: Bond Return, Equity Return and Price Index.
Post-crisis: commodity vector autoregression.
Concluding Remarks

VAR models and their extensions offer a rich modeling framework for studying inter-relationships of time series

Recent advances enable us to work with larger models, thus addressing criticisms from the empirical macro/finance community

References:

Thanks to:

Former and current students:
Ali Shojaie (U of Washington), Sumanta Basu (Cornell University), Jiahe Lin (U of Michigan), Andrey Skripnikov, Satyajit Ghosh, Kazem Faradonbeh (U of Florida)

NSF, NIH, NSA for financial support