Estimation of High-dimensional Vector Autoregressive (VAR) models

George Michailidis

Department of Statistics, University of Michigan

www.stat.lsa.umich.edu/∼gmichail

CANSSI-SAMSI Workshop,
Fields Institute, Toronto May 2014

Joint work with Sumanta Basu

Outline

1. Introduction
2. Modeling Framework
3. Theoretical Considerations
4. Implementation
5. Performance Evaluation

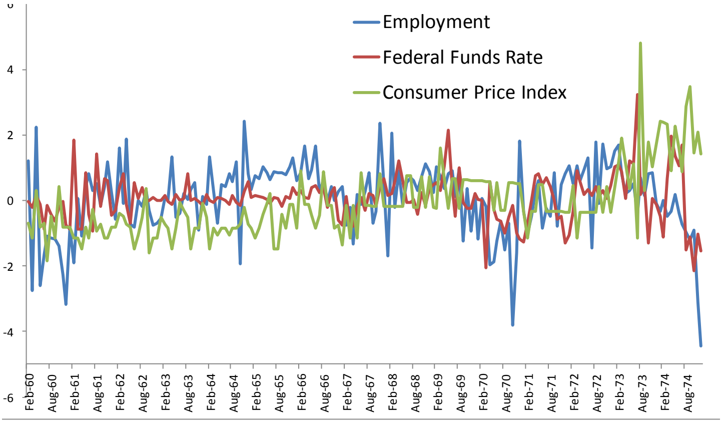
Vector Autoregressive models (VAR)

- widely used for structural analysis and forecasting of time-varying systems
- capture rich dynamics among system components
- popular in diverse application areas
  - control theory: system identification problems
  - economics: estimate macroeconomic relationships (Sims, 1980)
  - genomics: reconstructing gene regulatory network from time course data
  - neuroscience: study functional connectivity among brain regions from fMRI data (Friston, 2009)

VAR models in Economics

- testing relationship between money and income (Sims, 1972)
- understanding stock price-volume relation (Hiemstra et al., 1994)
- dynamic effect of government spending and taxes on output (Blanchard and Jones, 2002)
- identify and measure the effects of monetary policy innovations on macroeconomic variables (Bernanke et al., 2005)

VAR models in Economics

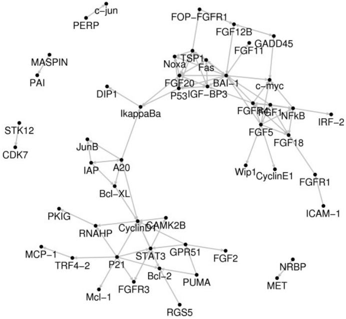


VAR models in Functional Genomics

- technological advances allow collecting huge amount of data
  - DNA microarrays, RNA-sequencing, mass spectrometry
- capture meaningful biological patterns via network modeling
- difficult to infer direction of influence from co-expression
- transition patterns in time course data helps identify regulatory mechanisms

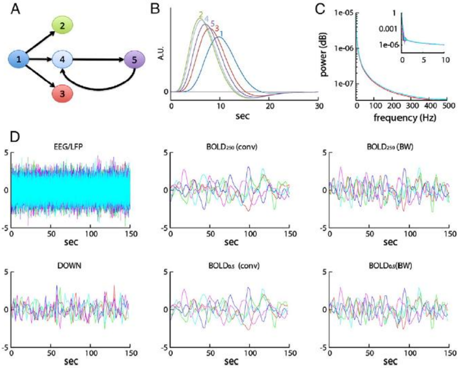
VAR models in Functional Genomics (ctd)

- HeLa gene expression regulatory network [Courtesy: Fujita et al., 2007]



VAR models in Neuroscience

- identify connectivity among brain regions from time course fMRI data
- connectivity of VAR generative model (Seth et al., 2013)

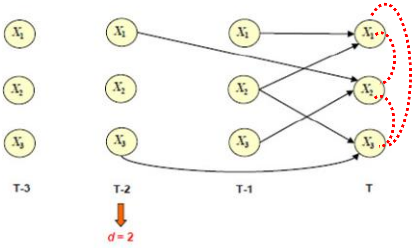


Model

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

$$X^t = A_1 X^{t-1} + \dots + A_d X^{t-d} + \varepsilon^t, \quad \varepsilon^t \overset{i.i.d}{\sim} N(\mathbf{0}, \Sigma_\varepsilon) \tag{1}$$

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



Why high-dimensional VAR?

- The parameter space grows quadratically (p^2 edges for p time series)
- order of the process (d) often unknown
- Economics:
  - Forecasting with many predictors (De Mol et al., 2008)
  - Understanding structural relationship - “price puzzle" (Christiano et al., 1999)
- Functional Genomics:
  - reconstruct networks among hundreds to thousands of genes
  - experiments costly - small to moderate sample size
- Finance:
  - structural changes - local stationarity

Literature on high-dimensional VAR models

- Economics:
  - Bayesian vector autoregression (lasso, ridge penalty; Litterman, Minnesota Prior)
  - Factor model based approach (FAVAR, dynamic factor models)
- Bioinformatics:
  - Discovering gene regulatory mechanisms using pairwise VARs (Fujita et al., 2007 and Mukhopadhyay and Chatterjee, 2007)
  - Penalized VAR with grouping effects over time (Lozano et al., 2009)
  - Truncated lasso and thesholded lasso variants (Shojaie and Michailidis, 2010 and Shojaie, Basu and Michailidis, 2012)
- Statistics:
  - lasso (Han and Liu, 2013) and group lasso penalty (Song and Bickel, 2011)
  - low-rank modeling with nuclear norm penalty (Negahban and Wainwright, 2011)
  - sparse VAR modeling via two-stage procedures (Davis et al., 2012)

Outline

1 Introduction

2 Modeling Framework

3 Theoretical Considerations

4 Implementation

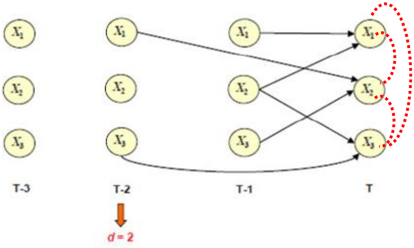
5 Performance Evaluation

Model

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

$$X^t = A_1 X^{t-1} + \ldots + A_d X^{t-d} + \varepsilon^t, \quad \varepsilon^t \overset{i.i.d}{\sim} N(\mathbf{0}, \Sigma_\varepsilon) \qquad (2)$$

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



Detour: VARs and Granger Causality

- Concept introduced by Granger (1969)
- A time series X is said to Granger-cause Y if it can be shown, usually through a series of F-tests on lagged values of X (and with lagged values of Y also known), that those X values provide statistically significant information about future values of Y.
- In the context of a high-dimensional VAR model we have that X_j^{T-t} is Granger-causal for X_i^T if $A_{i,j}^t \neq 0$.
- Granger-causality does not imply true causality; it is built on correlations
- Also, related to estimating a Directed Acyclic Graph (DAG) with $(d+1)\times p$ variables, with a known ordering of the variables

Estimating VARs through regression

- data: $\{X^0, X^1, \ldots, X^T\}$ - one replicate, observed at $T+1$ time points
- construct autoregression

$$\underbrace{\begin{bmatrix}(X^T)' \\ (X^{T-1})' \\ \vdots \\ (X^d)'\end{bmatrix}}_{\mathscr{Y}} = \underbrace{\begin{bmatrix}(X^{T-1})' & (X^{T-2})' & \cdots & (X^{T-d})' \\ (X^{T-2})' & (X^{T-3})' & \cdots & (X^{T-1-d})' \\ \vdots & \ddots & \vdots & \vdots \\ (X^{d-1})' & (X^{d-2})' & \cdots & (X^0)'\end{bmatrix}}_{\mathscr{X}} \underbrace{\begin{bmatrix}A_1' \\ \vdots \\ A_d'\end{bmatrix}}_{B^*} + \underbrace{\begin{bmatrix}(\varepsilon^T)' \\ (\varepsilon^{T-1})' \\ \vdots \\ (\varepsilon^d)'\end{bmatrix}}_{E}$$

$$\begin{aligned} vec(\mathscr{Y}) &= vec(\mathscr{X} B^*) + vec(E) \\ &= (I \otimes \mathscr{X})\, vec(B^*) + vec(E) \\ \underbrace{Y}_{Np\times 1} &= \underbrace{Z}_{Np\times q} \underbrace{\beta^*}_{q\times 1} + \underbrace{vec(E)}_{Np\times 1} \qquad vec(E) \sim N(\mathbf{0}, \Sigma_\varepsilon \otimes I) \end{aligned}$$

$$N = (T-d+1),\ q = dp^2$$

- Assumption : A_t are sparse, $\sum_{t=1}^{d} \|A_t\|_0 \le k$

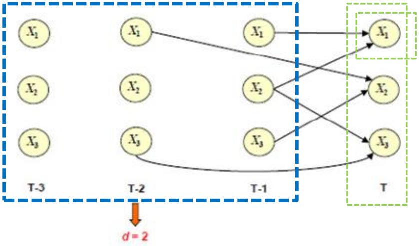
Estimates

- ℓ_1-penalized least squares (ℓ_1-LS)

$$\underset{\beta \in \mathbb{R}^q}{\operatorname{argmin}} \frac{1}{N} \|Y - Z\beta\|^2 + \lambda_N \|\beta\|_1$$

- ℓ_1-penalized log-likelihood (ℓ_1-LL) (Davis et al., 2012)

$$\underset{\beta \in \mathbb{R}^q}{\operatorname{argmin}} \frac{1}{N} (Y - Z\beta)' \left(\Sigma_\varepsilon^{-1} \otimes I\right) (Y - Z\beta) + \lambda_N \|\beta\|_1$$



Outline

1 Introduction

2 Modeling Framework

3 Theoretical Considerations

4 Implementation

5 Performance Evaluation

Detour: Consistency of Lasso Regression

$$\begin{bmatrix} Y \end{bmatrix}_{n\times 1} = \begin{bmatrix} \qquad X \qquad \end{bmatrix}_{n\times p} \begin{bmatrix} \beta^* \end{bmatrix}_{p\times 1} + \begin{bmatrix} \varepsilon \end{bmatrix}_{n\times 1}$$

LASSO : $\hat{\beta} := \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \frac{1}{n} \|Y - X\beta\|^2 + \lambda_n \|\beta\|_1$

- $S = \left\{ j \in \{1, \ldots, p\} | \beta_j^* \neq 0 \right\}$, $card(S) = k$, $k \ll n$, $\varepsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$
- Restricted Eigenvalue (RE): Assume

$$\alpha_{RE} := \min_{v \in \mathbb{R}^p, \|v\| \leq 1, \|v_{S^c}\|_1 \leq 3\|v_S\|_1} \frac{1}{n} \|Xv\|^2 > 0$$

Estimation error: $\|\hat{\beta} - \beta^*\| \leq \mathbb{Q}(X, \sigma) \frac{1}{\alpha_{RE}} \sqrt{\frac{k \log p}{n}}$ with high probability

Verifying Restricted Eigenvalue Condition

- Raskutti et al. (2010): If the rows of $X \overset{i.i.d.}{\sim} N(\mathbf{0}, \Sigma_X)$ and Σ_X satisfies RE, then X satisfies RE with high probability.
- Assumption of independence among rows crucial

- Rudelson and Zhou (2013): If the design matrix X can be factorized as $X = \Psi A$ where A satisfies RE and Ψ acts as (almost) an isometry on the images of sparse vectors under A, then X satisfies RE with high probability.

Back to Vector Autoregression

- Random design matrix $\mathscr{X}$, correlated with error matrix E

$$\underbrace{\begin{bmatrix} (X^T)' \\ (X^{T-1})' \\ \vdots \\ (X^d)' \end{bmatrix}}_{\mathscr{Y}} = \underbrace{\begin{bmatrix} (X^{T-1})' & (X^{T-2})' & \cdots & (X^{T-d})' \\ (X^{T-2})' & (X^{T-3})' & \cdots & (X^{T-1-d})' \\ \vdots & \ddots & \vdots & \vdots \\ (X^{d-1})' & (X^{d-2})' & \cdots & (X^0)' \end{bmatrix}}_{\mathscr{X}} \underbrace{\begin{bmatrix} A_1' \\ \vdots \\ A_d' \end{bmatrix}}_{B^*} + \underbrace{\begin{bmatrix} (\varepsilon^T)' \\ (\varepsilon^{T-1})' \\ \vdots \\ (\varepsilon^d)' \end{bmatrix}}_{E}$$

$$\begin{aligned} vec(\mathscr{Y}) &= vec(\mathscr{X} B^*) + vec(E) \\ &= (I \otimes \mathscr{X}) vec(B^*) + vec(E) \\ \underbrace{Y}_{Np\times 1} &= \underbrace{Z}_{Np\times q} \underbrace{\beta^*}_{q\times 1} + \underbrace{vec(E)}_{Np\times 1} \qquad vec(E) \sim N(\mathbf{0}, \Sigma_\varepsilon \otimes I) \end{aligned}$$

$$N = (T-d+1), \; q = dp^2$$

Vector Autoregression (ctd)

Key Questions:

- How often does RE hold?
- How small is α_{RE}?
- How does the cross-correlation affect convergence rates?

Consistency of VAR estimates

- Restricted Eigenvalue (RE) assumption: $(I \otimes \mathscr{X})_{q \times q} \sim RE(\alpha, \tau(N,q))$ with $\alpha > 0, \tau(N,q) > 0$ if

$$\theta' \left(I \otimes \mathscr{X}' \mathscr{X} / N \right) \theta \geq \alpha \|\theta\|_2^2 - \tau(N,q) \|\theta\|_1^2 \text{ for all } \theta \in \mathbb{R}^q \qquad (3)$$

- Deviation Condition: There exists a function $\mathbb{Q}(\beta^*, \Sigma_\varepsilon)$ such that

$$\|vec\left(\mathscr{X}' E / N\right)\|_{\max} \leq \mathbb{Q}(\beta^*, \Sigma_\varepsilon) \sqrt{\frac{\log d + 2 \log p}{N}} \qquad (4)$$

- Key Result:
  Estimation Consistency: If (3) and (4) hold with $k\tau(N,q) \leq \alpha/32$, then, for any $\lambda_N \geq 4\mathbb{Q}(\beta^*, \Sigma_\varepsilon)\sqrt{(\log d + 2 \log p)/N}$, lasso estimate $\hat{\beta}_{\ell_1}$ satisfies

$$\|\hat{\beta}_{\ell_1} - \beta^*\| \leq 64 \, \frac{\mathbb{Q}(\beta^*, \Sigma_\varepsilon)}{\alpha} \sqrt{\frac{k\left(\log d + 2 \log p\right)}{N}}$$

Verifying RE and Deviation Condition

- Negahban and Wainwright, 2011: for VAR(1) models, assume $\|A_1\| < 1$, where $\|A\| := \sqrt{\Lambda_{\max}(A'A)}$
- For $p = 1$, $d = 1$, $X^t = \rho X^{t-1} + \varepsilon^t$, reduces to $|\rho| < 1$ - equivalent to stability

- Han and Liu, 2013: for VAR(d) models, reformulate as VAR(1): $\tilde{X}^t = \tilde{A}_1 \tilde{X}^{t-1} + \tilde{\varepsilon}^t$, where

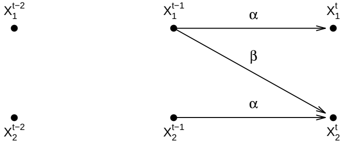
$$\tilde{X}^t = \begin{bmatrix} X^t \\ X^{t-1} \\ \vdots \\ X^{t-d+1} \end{bmatrix}_{dp\times 1} \quad \tilde{A}_1 = \begin{bmatrix} A_1 & A_2 & \cdots & A_{d-1} & A_d \\ I_p & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & I_p & \cdots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & I_p & \mathbf{0} \end{bmatrix}_{dp\times dp} \quad \tilde{\varepsilon}^t = \begin{bmatrix} \varepsilon^t \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{bmatrix}_{dp\times 1}$$

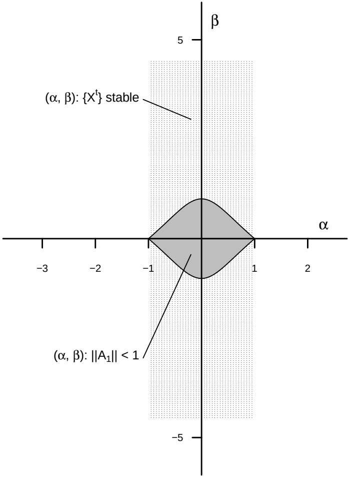
- Assume $\|\tilde{A}_1\| < 1$

VAR(1): Stability and $\|A_1\| < 1$

- $\|A_1\| \not< 1$ for *many* stable VAR(1) models

$$X^t = A_1 X^{t-1} + \varepsilon^t, \qquad A_1 = \begin{bmatrix} \alpha & 0 \\ \beta & \alpha \end{bmatrix}$$

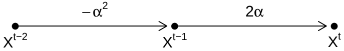


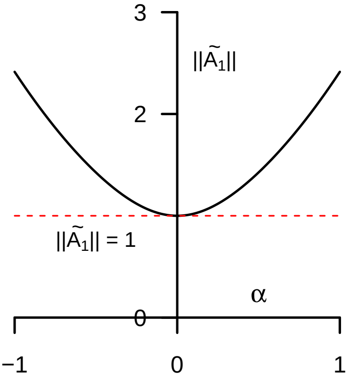


VAR(d): Stability and $\|\tilde{A}_1\| < 1$

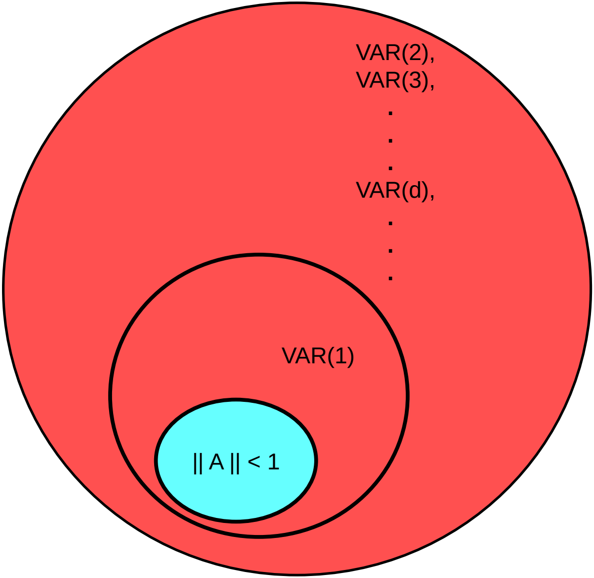
- $\|\tilde{A}_1\| \not< 1$ for *any* stable VAR(d) models, if $d > 1$

$$X^t = 2\alpha X^{t-1} - \alpha^2 X^{t-2} + \varepsilon^t, \quad \begin{bmatrix} X^t \\ X^{t-1} \end{bmatrix} = \begin{bmatrix} 2\alpha & -\alpha^2 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} X^{t-1} \\ X^{t-2} \end{bmatrix} + \begin{bmatrix} \varepsilon^t \\ 0 \end{bmatrix}$$

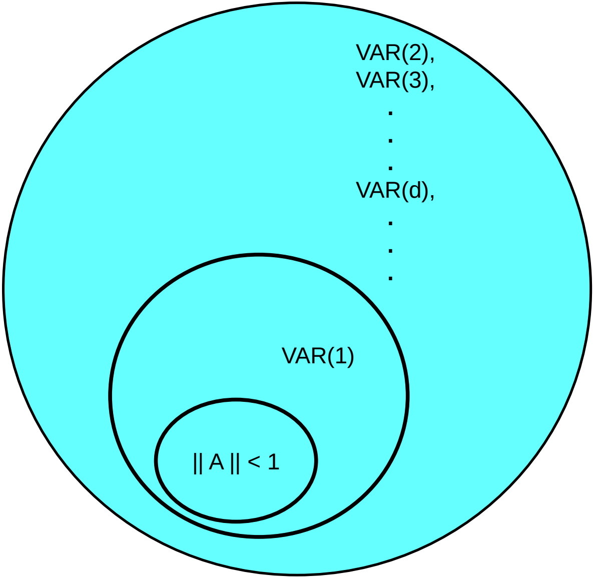




Stable VAR models



Stable VAR models



Quantifying Stability through the Spectral Density

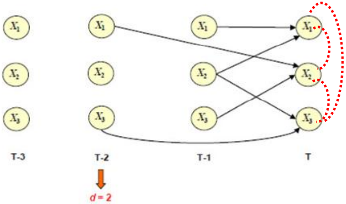
- Spectral density function of a covariance stationary process $\{X^t\}$,

$$f_X(\theta) = \frac{1}{2\pi} \sum_{l=-\infty}^{\infty} \Gamma_X(l) e^{-il\theta}, \qquad \theta \in [-\pi, \pi]$$

- $\Gamma_X(l) = \mathbb{E}\left[X^t (X^{t+l})'\right]$, autocovariance matrix of order l
- If the VAR process is stable, it has a closed form (Priestley, 1981)

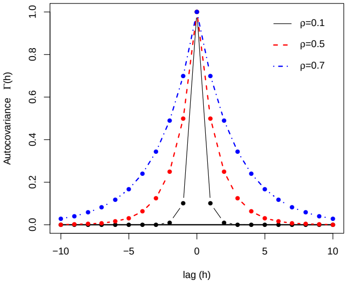
$$f_X(\theta) = \frac{1}{2\pi} \left(\mathscr{A}(e^{-i\theta})\right)^{-1} \Sigma_\varepsilon \left(\mathscr{A}^*(e^{-i\theta})\right)^{-1}$$

- The two sources of dependence factorize in frequency domain

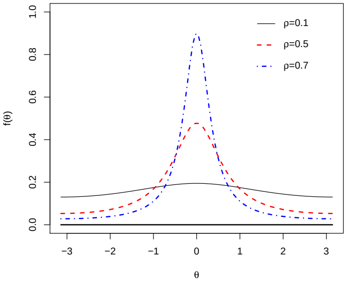


Quantifying Stability by Spectral Density

- For univariate processes, the “peak" of the spectral density measures stability of the process - (sharper peak = less stable)



(f) Autocovariance of $AR(1)$



(g) Spectral Density of $AR(1)$

- For multivariate processes, similar role is played by the maximum eigenvalue of the (matrix-valued) spectral density

Quantifying Stability by Spectral Density

- For a stable VAR(d) process $\{X^t\}$, the maximum eigenvalue of its spectral density captures its stability

$$\mathcal{M}(f_X) = \max_{\theta \in [-\pi,\pi]} \Lambda_{\max}(f_X(\theta))$$

- The minimum eigenvalue of the spectral density captures dependence among its components

$$\mathfrak{m}(f_X) = \min_{\theta \in [-\pi,\pi]} \Lambda_{\min}(f_X(\theta))$$

- For stable VAR(1) processes, $\mathcal{M}(f_X)$ scales with $(1-\rho(A_1))^{-2}$, $\rho(A_1)$ is the spectral radius of A_1
- $\mathfrak{m}(f_X)$ scales with the capacity (maximum incoming + outgoing effect at a node) of the underlying graph

Consistency of *VAR* estimates

Theorem

Consider a random realization $\{X^0, \ldots, X^T\}$ *generated according to a stable VAR(d) process with* $\Lambda_{\min}(\Sigma_\varepsilon) > 0$ *. Then there exist deterministic functions* $\phi_i(A_t, \Sigma_\varepsilon) > 0$ *and constants* $c_i > 0$ *such that for* $N \succsim \phi_0(A_t, \Sigma_\varepsilon)\sqrt{k(\log d + 2\log p)/N}$*, the lasso estimate* ($\ell_1$*-LS) with* $\lambda_N \asymp \sqrt{(2\log p + \log d)/N}$ *satisfies, with probability at least* $1 - c_1 \exp[-c_2(2\log p + \log d)]$,

$$\sum_{h=1}^{d} \|\hat{A}_h - A_h\| \leq \phi_1(A_t, \Sigma_\varepsilon)\left(\sqrt{k(\log d + 2\log p)/N}\right)$$

$$\frac{1}{\sqrt{N}} \sum_{t=d}^{T} \left\| \sum_{h=1}^{d} (\hat{A}_h - A_h) X^h \right\| \leq \phi_2(A_t, \Sigma_\varepsilon)\left(\sqrt{k(\log d + 2\log p)/N}\right)$$

Further, a thresholded version of lasso $\tilde{A} = \left(\hat{A}_{t,ij}\mathbf{1}_{\{|\hat{A}_{t,ij}| > \lambda_N\}}\right)$ *satisfies*

$$\left| supp(\tilde{A}^{1:d}) \backslash supp(A^{1:d}) \right| \leq \phi_3(A_t, \Sigma_\varepsilon) k$$

$\phi_i(A^t, \Sigma_\varepsilon)$ are large when $\mathscr{M}(f_X)$ is large and $\mathfrak{m}(f_X)$ is small.

Some Remarks

Convergence rates governed by:

- dimensionality parameters - dimension of the process (p), order of the process (d), number of parameters (k) in the transition matrices A_i and sample size (N = T -d + 1)
- internal parameters - curvature (α), tolerance (τ) and the deviation bound $Q(\beta^\star, \Sigma_\varepsilon)$

The squared ℓ_2-errors of estimation and prediction scale with the dimensionality parameters as

$$k(2logp + logd)/N,$$

similar to the rates obtained when the observations are independent

The temporal and cross-sectional dependence affect the rates only through the internal parameters.
Typically, the rates are better when α is large and $Q(\beta^\star, \Sigma_\varepsilon), \tau$ are small.
This dependence is captured in the next results.

Verifying RE

Proposition

Consider a random realization $\{X^0, \ldots, X^T\}$ *generated according to a stable VAR(d) process. Then there exist universal positive constants* c_i *such that for all* $N \gtrsim \max\{1, \omega^{-2}\} k \log(dp)$*, with probability at least* $1 - c_1 \exp(-c_2 N \min\{\omega^2, 1\})$*,*

$$I_p \otimes (\mathscr{X}' \mathscr{X} / N) \sim RE(\alpha, \tau),$$

where

$$\omega = \frac{\Lambda_{\min}(\Sigma_\varepsilon)/\Lambda_{\max}(\Sigma_\varepsilon)}{\mu_{\max}(\mathscr{A})/\mu_{\min}(\tilde{\mathscr{A}})}, \quad \alpha = \frac{\Lambda_{\min}(\Sigma_\varepsilon)}{2\mu_{\max}(\mathscr{A})},$$

$$\tau(N, q) = c_3 \frac{\Lambda_{\min}(\Sigma_\varepsilon)}{\mu_{\max}(\mathscr{A})} \max\{\omega^{-2}, 1\} \frac{\log(dp)}{N}.$$

Verifying Deviation Condition

Proposition

If $q \geq 2$, then, for any $A > 0$, $N \gtrsim \log d + 2\log p$, with probability at least $1 - 12q^{-A}$, we have

$$\left\| vec\left(\mathscr{X}'E/N\right)\right\|_{\max} \leq \mathbb{Q}(\beta^*, \Sigma_\varepsilon)\sqrt{\frac{\log d + 2\log p}{N}},$$

where

$$\mathbb{Q}(\beta^*, \Sigma_\varepsilon) = (18 + 6\sqrt{2(A+1)})\left[\Lambda_{\max}(\Sigma_\varepsilon) + \frac{\Lambda_{\max}(\Sigma_\varepsilon)}{\mu_{\min}(\mathscr{A})} + \frac{\Lambda_{\max}(\Sigma_\varepsilon)\mu_{\max}(\mathscr{A})}{\mu_{\min}(\mathscr{A})}\right]$$

Some Comments

RE:

the convergence rates are faster for larger α and smaller τ. From the expressions of ω, α and τ, it is clear that the VAR estimates have lower error bounds when $\Lambda_{\max}(\Sigma_\varepsilon), \mu_{\max}(\mathscr{A})$ are smaller and $\Lambda_{\min}(\Sigma_\varepsilon), \mu_{\min}(\mathscr{A})$ are larger.

Deviation bound:

VAR estimates exhibit lower error bounds when $\Lambda_{\max}(\Sigma_\varepsilon)$, $\mu_{\max}(\mathscr{A})$ are smaller and $\Lambda_{\min}(\Sigma_\varepsilon)$, $\mu_{\min}(\mathscr{A})$ are larger (similar to RE)

Outline

1 Introduction

2 Modeling Framework

3 Theoretical Considerations

4 Implementation

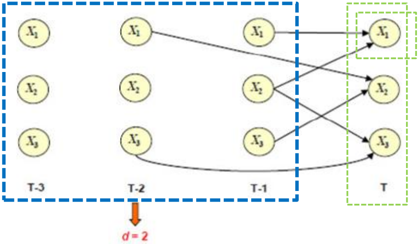
5 Performance Evaluation

ℓ_1-LS:

Denote the i^{th} column of a matrix M by M_i.

$$\begin{aligned} & arg \quad \min_{\beta \in \mathbb{R}^q} \frac{1}{N} \|Y - Z\beta\|^2 + \lambda_N \|\beta\|_1 \\ \equiv \quad & arg \quad \min_{B_1,\ldots,B_p} \frac{1}{N} \sum_{i=1}^{p} \|\mathscr{Y}_i - \mathscr{X} B_i\|^2 + \lambda_N \sum_{i=1}^{p} \|B_i\|_1 \end{aligned}$$

- Amounts to running p separate LASSO programs, each with dp predictors: $\mathscr{Y}_i \sim \mathscr{X}, i = 1, \ldots, p$.



ℓ_1-LL:

Davis et al, 2012, proposed the following algorithm:

$$\begin{aligned} & arg \min_{\beta \in \mathbb{R}^q} \frac{1}{N} (Y - Z\beta)' \left(\Sigma_\varepsilon^{-1} \otimes I\right) (Y - Z\beta) + \lambda_N \|\beta\|_1 \\ \equiv \; & arg \min_{\beta \in \mathbb{R}^q} \frac{1}{N} \left\| \left(\Sigma_\varepsilon^{-1/2} \otimes I\right) Y - \left(\Sigma_\varepsilon^{-1/2} \otimes \mathscr{X}\right) \beta \right\|^2 + \lambda_N \|\beta\|_1 \end{aligned}$$

- Amounts to running a single LASSO program with dp^2 predictors: $\left(\Sigma_\varepsilon^{-1/2} \otimes I\right) Y \sim \Sigma_\varepsilon^{-1/2} \otimes \mathscr{X}$ - cannot be implemented in parallel.

- σ_ε^{ij}:= $(i,j)^{th}$ entry of Σ_ε^{-1}. The objective function is

$$\frac{1}{N} \sum_{i=1}^{p} \sum_{j=1}^{p} \sigma_\varepsilon^{ij} (\mathscr{Y}_i - \mathscr{X} B_i)' (\mathscr{Y}_j - \mathscr{X} B_j) + \lambda_N \sum_{k=1}^{p} \|B_k\|_1$$

Block Coordinate Descent for ℓ_1-LL

1. pre-select d. Run ℓ_1-LS to get $\hat{B}, \hat{\Sigma}_\varepsilon^{-1}$.
2. iterate till convergence:
    1. For $i = 1, \dots, p$,
        - ⋆ set $r_i := (1/2\,\hat{\sigma}_\varepsilon^{ii}) \sum_{j \neq i} \hat{\sigma}_\varepsilon^{ij} \left(\mathscr{Y}_j - \mathscr{X} \hat{B}_j \right)$
        - ⋆ update $\hat{B}_i = arg \min_{B_i} \frac{\hat{\sigma}_\varepsilon^{ii}}{N} \|(\mathscr{Y}_i + r_i) - \mathscr{X} B_i\|^2 + \lambda_N \|B_i\|_1$

- each iteration amounts to running p separate LASSO programs, each with dp predictors: $\mathscr{Y}_i + r_i \sim \mathscr{X}$, $i = 1, \dots, p$.
- Can be implemented in parallel

Outline

1. Introduction
2. Modeling Framework
3. Theoretical Considerations
4. Implementation
5. Performance Evaluation

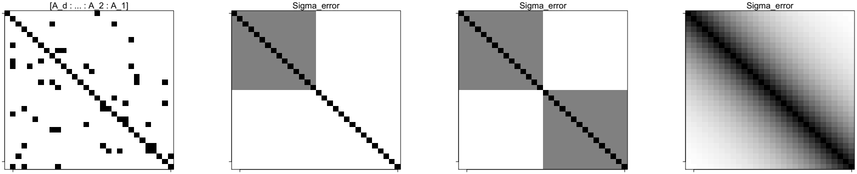
VAR models considered

- Small Size VAR, $p = 10, d = 1, T = 30, 50$
- Medium Size VAR, $p = 30, d = 1, T = 80, 120, 160$

In each setting, we generate an adjacency matrix A_1 with $5 \sim 10\%$ non-zero edges selected at random and rescale to ensure that the process is stable with $SNR = 2$.
We generate three different error processes with covariance matrix Σ_ε from one of the following families:

1. **Block-I:** $\Sigma_\varepsilon = ((\sigma_{\varepsilon,ij}))_{1 \le i,j \le p}$ with $\sigma_{\varepsilon,ii} = 1$, $\sigma_{\varepsilon,ij} = \rho$ if $1 \le i \ne j \le p/2$, 0 otherwise;
2. **Block-II:** $\Sigma_\varepsilon = ((\sigma_{\varepsilon,ij}))_{1 \le i,j \le p}$ with $\sigma_{\varepsilon,ii} = 1$, $\sigma_{\varepsilon,ij} = \rho$ if $1 \le i \ne j \le p/2$ or $p/2 < i \ne j \le p$, 0 otherwise;
3. |bf Toeplitz: $\Sigma_\varepsilon = ((\sigma_{\varepsilon,ij}))_{1 \le i,j \le p}$ with $\sigma_{\varepsilon,ij} = \rho^{|i-j|}$.

VAR models considered (ctd)



(a) A_1 (b) Σ_ϵ: Block-I (c) Σ_ϵ: Block-II (d) Σ_ϵ: Toeplitz

We let ρ vary in $\{0.5, 0.7, 0.9\}$.

Larger values of ρ indicate that the error processes are more strongly correlated.

Comparisons and Performance Criteria

Different methods for VAR estimation:

- OLS
- ℓ_1-LS
- ℓ_1-LL
- ℓ_1-LL-O (Oracle version, assuming Σ_ε known)
- Ridge

evaluated using the following performance metrics:

1. *Model Selection:* Area under receiving operator characteristic curve (AUROC)
2. *Estimation error:* Relative estimation accuracy measured by $\|\hat{B} - B\|_F / \|B\|_F$

Results I

Table: VAR(1) model with $p = 10$, $T = 30$

		BLOCK-I			BLOCK-II			Toeplitz		
	ρ	0.5	0.7	0.9	0.5	0.7	0.9	0.5	0.7	0.9
AUROC	ℓ_1-LS	0.77	0.74	0.7	0.79	0.76	0.74	0.82	0.79	0.77
	ℓ_1-LL	0.77	0.75	0.73	0.79	0.77	0.77	0.81	0.8	0.81
	ℓ_1-LL-O	0.8	0.79	0.76	0.82	0.8	0.81	0.85	0.84	0.84
Estimation Error	OLS	1.24	1.39	1.77	1.29	1.63	2.36	1.32	1.56	2.58
	ℓ_1-LS	0.68	0.72	0.76	0.64	0.67	0.7	0.63	0.66	0.69
	ℓ_1-LL	0.66	0.66	0.66	0.57	0.59	0.53	0.59	0.56	0.49
	ℓ_1-LL-O	0.61	0.62	0.62	0.53	0.54	0.47	0.53	0.51	0.42
	ridge	0.72	0.74	0.75	0.7	0.71	0.72	0.7	0.71	0.72

Results II

Table: VAR(1) model with $p = 30$, $T = 120$

		BLOCK-I			BLOCK-II			Toeplitz		
	ρ	0.5	0.7	0.9	0.5	0.7	0.9	0.5	0.7	0.9
AUROC	ℓ_1-LS	0.89	0.85	0.77	0.87	0.81	0.69	0.91	0.87	0.76
	ℓ_1-LL	0.89	0.87	0.82	0.9	0.89	0.88	0.91	0.91	0.89
	ℓ_1-LL-O	0.92	0.9	0.84	0.93	0.92	0.9	0.94	0.93	0.92
Estimation Error	OLS	1.73	2	2.93	1.95	2.53	4.28	1.82	2.28	3.88
	ℓ_1-LS	0.72	0.76	0.85	0.74	0.82	0.93	0.69	0.73	0.86
	ℓ_1-LL	0.71	0.71	0.72	0.68	0.68	0.65	0.67	0.63	0.6
	ℓ_1-LL-O	0.66	0.66	0.68	0.64	0.63	0.59	0.63	0.59	0.54
	Ridge	0.81	0.83	0.85	0.82	0.85	0.88	0.81	0.82	0.86

Summary/Discussion

- Investigated penalized VAR estimation in high-dimension
- Established estimation consistency for all stable VAR models, based on novel techniques using spectral representation of stationary processes
- Developed parallellizable algorithm for likelihood based VAR estimates

There is extensive work on characterizing univariate time series, through mixing conditions or functional dependence measures. However, thre is little work for multivariate series, which is needed to be able to provide results in the current setting.

References

- S. Basu and G. Michailidis, Estimation in High-dimensional Vector Autoregressive Models, arXiv: 1311.4175
- S. Basu, A. Shojaie and G. Michailidis, Network Granger Causality with Inherent Grouping Structure, revised for *JMLR*