Finite Time Analysis of Vector Autoregressive Models under Linear Restrictions

Yao Zheng

collaboration with Guang Cheng

Department of Statistics, Purdue University

Talk at Math, IUPUI

October 23, 2018

Introduction

The vector autoregressive (VAR) model

Vector autoregressive (VAR) model of order one:

$$X_{t+1} = AX_t + \eta_t, \qquad t = 1, \dots, T, \tag{1}$$

where

- $X_t \in \mathbb{R}^d$ is the observed d-dimensional time series
- $A \in \mathbb{R}^{d\times d}$ is the unknown transition matrix (possible over-parametrization when d is even moderately large!)
- $\eta_t \in \mathbb{R}^d$ are $i.i.d.$ innovations with mean zero
- T is the sample size/time horizon (asymptotic analysis: $T \to \infty$)
- Applications: e.g., economics and finance, energy forecasting, psychopathology, neuroscience, reinforcement learning, ...

The problem of over-parameterization

... is more severe for general VAR(p) models:

$$X_{t+1} = A_1 X_t + A_2 X_{t-1} + \cdots + A_p X_{t+1-p} + \eta_t,$$

Number of parameters $= O(pd^2)$

$\Rightarrow$ cannot provide reliable estimates and forecasts without further restrictions (Stock and Watson, 2001).

Literature: Taming the dimensionality of large VAR models

(D). Direct dimensionality reduction:

- Regularized estimation: Davis et al. (2015, JCGS), (Han et al., 2015, JMLR), (Basu and Michailidis, 2015, AoS), etc.
- Banded model: Guo et al. (2016, Biometrika)
- Network model: Zhu et al. (2017, AoS)
- Other parameter restrictions motivated by specific applications

(I). Indirect dimensionality reduction: low-rank structure, PCA, factor modelling, ...

We focus on direct dimensionality reduction in this paper.

What most existing work on (D) has in common:

(i) A particular sparsity or structural assumption is often imposed on the transition matrix A

e.g., exact sparsity, banded structure, certain network structure

(ii) There is an almost exclusive focus on stable processes

technically, this is to impose that the spectral radius $\rho(A) < 1$, or often even more stringently, the spectral norm $\|A\|_2 < 1$

*Denote the spectral radius of A by $\rho(A) := \max\{|\lambda_1|, \ldots, |\lambda_d|\}$, where λ_i are the eigenvalues of $A \in \mathbb{R}^{d \times d}$. Note that even when $\rho(A) < 1$, $\|A\|_2$ can be arbitrarily large for an asymmetric matrix A.

Our objective

– to study large VAR models from a more general viewpoint, without being confined to any particular sparsity structure or to the stable regime

We provide a novel non-asymptotic (finite-time) analysis of the ordinary least squares (OLS) estimator for

- possibly unstable VAR models (applicable region: $\rho(A) \leq 1 + c/T$)
- under linear restrictions in the form of

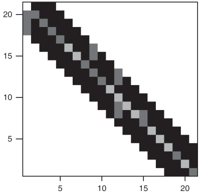
$$\underbrace{\mathcal{C}}_{\text{known restriction matrix}} \quad \underbrace{\text{vec}(A')}_{\text{stacking the rows of } A} = \underbrace{\mu}_{\text{known vector}}; \qquad (2)$$

often, we may simply use $\mu = 0$.

$\Rightarrow$ note that (2) encompasses zero and equality restrictions

Example 1: Banded VAR model of Guo et al. (2016, Biometrika)

Location plot and estimated transition matrix $\hat{A}$



- Motivation: in practice, it is often sufficient to collect information from "neighbors"

- Note that the same reasoning can be applied to general graphical structures: the zero-nonzero pattern of A can be determined according to any practically motivated graph with d nodes

Example 2: Network VAR model of Zhu et al. (2017, AoS)

An example of both zero and equality restrictions

- To analyze users' time series data from large social networks, the network VAR model of Zhu et al. (2017, AoS) imposes that

  (i) all diagonal entries of A are equal,

  (ii) all nonzero off-diagonal entries of A are equal, and

  (iii) the zero-nonzero pattern of A is known

  (e.g., a_{ij} is nonzero only if individual j follows individual i on the social network)

- But this model is essentially low-dimensional, as the number of unknown parameters is a fixed small number.

Problem formulation

General framework: Multivariate stochastic regression

– This includes VAR models as a special case

The unrestricted model:

$$\underbrace{Y_t}_{n\times 1} = \underbrace{A_*}_{n\times d}\underbrace{X_t}_{d\times 1} + \underbrace{\eta_t}_{n\times 1}. \tag{3}$$

- This becomes the VAR(1) model when $Y_t = X_{t+1}$ and $n = d$.
- Note that (X_t, Y_t) are time-dependent.

Imposing linear restrictions

- Let $\beta_* = \text{vec}(A'_*) \in \mathbb{R}^N$, where $N = nd$.
- Then the parameter space of a linearly restricted model can be defined as

$$\mathcal{L} = \{\beta \in \mathbb{R}^N : \underbrace{\mathcal{C}}_{(N-m)\times N} \beta = \underbrace{\mu}_{(N-m)\times 1}\},$$

  where $\mathcal{C}$ and μ are known, $\text{rank}(\mathcal{C}) = N - m$ (representing $N - m$ independent restrictions)
- To ease the notation, we restrict our attention to $\mu = 0$ in this talk.

An equivalent form

Note that

$$\mathcal{L} = \{\beta \in \mathbb{R}^N : \underbrace{\mathcal{C}}_{(N-m)\times N} \beta = \underbrace{0}_{(N-m)\times 1}\}$$

has an equivalent, unrestricted parameterization:

$$\mathcal{L} = \{\underbrace{R}_{N\times m}\theta : \theta \in \mathbb{R}^m\}.$$

Specifically:

- Let $\widetilde{\mathcal{C}}$ be an $m \times N$ complement of $\mathcal{C}$ such that $\mathcal{C}_{\text{full}} = (\widetilde{\mathcal{C}}', \mathcal{C}')'$ is invertible, and let $\mathcal{C}_{\text{full}}^{-1} = (R, \widetilde{R})$, where R is an $N \times m$ matrix.
- Note that if $\mathcal{C}\beta = 0$, then $\beta = \mathcal{C}_{\text{full}}^{-1}\mathcal{C}_{\text{full}}\beta = R\widetilde{\mathcal{C}}\beta + \widetilde{R}\mathcal{C}\beta = R\theta$, where $\theta = \widetilde{\mathcal{C}}\beta$. Conversely, if $\beta = R\theta$, then $\mathcal{C}\beta = \mathcal{C}R\theta = 0$. Thus, we have the above equivalence.

Implications

- There exists a unique unrestricted $\theta_* \in \mathbb{R}^m$ such that $\beta_* = R\theta_*$.
- Therefore, the original restricted model can be converted into a reparameterized unrestricted model.
- Special case: when $R = I_N$, there is no restriction at all, and $\beta_* = \theta_*$.

How to encode restrictions via R or $\mathcal{C}$?

Example 1 (Zero restriction):

- Suppose that the i-th element of β is restricted to zero: i.e., $\beta_i = 0$.
- Then this can be encoded in R by setting the i-th row of R to zero.
- Alternatively, it can be built into $\mathcal{C}$ by setting a row of $\mathcal{C}$ to

$$(0, \ldots, 0, 1, 0, \ldots, 0) \in \mathbb{R}^N,$$

where the i-th entry is one.

How to encode restrictions via R or $\mathcal{C}$?

Example 2 (Equality restriction):

- Consider the restriction that the i-th and j-th elements of β are equal: i.e., $\beta_i - \beta_j = 0$.

- Suppose that the value of $\beta_i = \beta_j$ is θ_k, the k-th element of θ. Then this restriction can be encoded in R by setting both the i-th and j-th rows of R to

$$(0, \ldots, 0, 1, 0, \ldots, 0) \in \mathbb{R}^m,$$

  where the k-th entry is one.

- We may set a row of $\mathcal{C}$ to the $1 \times N$ vector $c(i, j)$ whose ℓ-th entry is

$$[c(i, j)]_\ell = 1(\ell = i) - 1(\ell = j),$$

  where $1(\cdot)$ is the indicator function.

The ordinary least squares (OLS) estimator

- Define $T \times n$ matrices

$$Y = (Y_1, \ldots, Y_T)', \quad E = (\eta_1, \ldots, \eta_T)'$$

and $T \times d$ matrix

$$X = (X_1, \ldots, X_T)'.$$

Then (3) has the matrix form

$$Y = XA_*' + E.$$

The ordinary least squares (OLS) estimator

- Moreover, let

$$y = \text{vec}(Y), \quad \eta = \text{vec}(E) \quad \text{and} \quad Z = (I_n \otimes X)R.$$

- By vectorization and reparameterization, we can write the linearly restricted model in vector form as

$$y = (I_n \otimes X)\beta_* + \eta = Z\theta_* + \eta.$$

- As a result, the OLS estimator of β_* for the restricted model can be defined as

$$\widehat{\beta} = R\widehat{\theta}, \qquad \text{where} \qquad \widehat{\theta} = \underset{\theta \in \mathbb{R}^m}{\arg\min} \|y - Z\theta\|^2. \tag{4}$$

The ordinary least squares (OLS) estimator

- To ensure the feasibility of (4), we assume that $nT \geq m$.

  (note that $Z \in \mathbb{R}^{nT \times m}$; however, Z need not have full rank).

- Let $R = (R_1', \ldots, R_n')'$, where R_i are $d \times m$ matrices. Then,

$$A_* = (I_n \otimes \theta_*')\widetilde{R},$$

  where $\widetilde{R}$ is an $mn \times d$ matrix:

$$\widetilde{R} = (R_1, \ldots, R_n)'.$$

  Hence, we can obtain the OLS estimator of A by

$$\widehat{A} = (I_n \otimes \widehat{\theta'})\widetilde{R}.$$

General upper bound analysis

– will be applied to VAR models later...

Key technical tool for upper bound analysis

Mendelson's small-ball method for time-dependent data (Simchowitz et al., 2018, COLT)

Why using this method?

- Asymptotic tools require substantially different approach to deal with stable and unstable processes $\{X_t\}$.
- Nonasymptotic tools usually rely on mixing conditions, which suffer from error bound degradation for unstable processes.
- The small-ball method helps us establish lower bounds of the Gram matrix $X'X$ (or $Z'Z$) under very mild conditions, while dropping the stability assumption and avoiding reliance on mixing properties.

How to use the small-ball method?

- Formulate a small-ball condition
- Use this condition to control the lower tail behavior of the Gram matrix
- Derive estimation error bounds
- Verify the small-ball condition

Main idea of the small-ball method to lower-bound $\lambda_{\min}(\sum_{t=1}^{T} X_t X_t^{\mathsf{T}})$

a. Divide the data into size-k blocks, with the ℓ-th block being $\{X_{(\ell-1)k+1}, \ldots, X_{\ell k}\}$.

b. Lower-bound each $\sum_{i=1}^{k} \langle X_{(\ell-1)k+i}, w \rangle^2$ *w.h.p.* by (establishing) a *block martingale small ball condition*.

c. Aggregate to get with probability at least $1 - \exp(-cT/k)$,

$$\sum_{t=1}^{T} \langle X_t, w \rangle^2 \gtrsim T w^{\mathsf{T}} \Gamma_k w.$$

d. Strengthen the pointwise bound into a lower bound on $\inf_{w \in \mathcal{S}^{d-1}} \sum_{t=1}^{T} \langle X_t, w \rangle^2$ by the covering method.

Small-ball condition for dependent data

The block martingale small ball (BMSB) condition is defined as follows:

(i) Univariate case: For $\{X_t\}_{t \geq 1}$ taking values in $\mathbb{R}$ adapted to the filtration $\{\mathcal{F}_t\}$, we say that $\{X_t\}$ satisfies the (k, ν, α)-BMSB condition if:

there exist an integer $k \geq 1$ and universal constants $\nu > 0$ and $\alpha \in (0, 1)$ such that for every integer $s \geq 0$,

$$k^{-1} \sum_{t=1}^{k} \mathbb{P}(|X_{s+t}| \geq \nu \mid \mathcal{F}_s) \geq \alpha$$

with probability one.

Here, k is the block size.

Small-ball condition for dependent data

(ii) Multivariate case: For $\{X_t\}_{t\geq 1}$ taking values in $\mathbb{R}^d$, we say that $\{X_t\}$ satisfies the $(k, \Gamma_{\mathrm{sb}}, \alpha)$-BMSB condition if:

there exists

$$0 \prec \Gamma_{\mathrm{sb}} \in \mathbb{R}^{d\times d}$$

such that, for every $\omega \in \mathcal{S}^{d-1}$, the univariate time series

$$\{\omega' X_t, t = 1, 2, \dots\}$$

satisfies the $(k, \sqrt{w'\Gamma_{\mathrm{sb}}w}, \alpha)$-BMSB condition.

Regularity conditions for upper-bound analysis

A1. $\{X_t\}_{t=1}^T$ satisfies the $(k, \Gamma_{\mathrm{sb}}, \alpha)$-BMSB condition.

A2. For any $\delta \in (0,1)$, there exists $\overline{\Gamma}_R = R'(I_n \otimes \overline{\Gamma})R$ dependent on δ such that $\mathbb{P}(Z'Z \npreceq T\overline{\Gamma}_R) \leq \delta$.

A3. For every integer $t \geq 1$, $\eta_t \mid \mathcal{F}_t$ is mean-zero and σ^2-sub-Gaussian, where

$$\mathcal{F}_t = \sigma\{\eta_1, \dots, \eta_{t-1}, X_1, \dots, X_t\}.$$

Note that $X_t \in \mathcal{F}_t$.

General upper bound for $\|\widehat{\beta} - \beta_*\| (= \|\widehat{A} - A_*\|_F)$

Theorem 1: *Let $\{(X_t, Y_t)\}_{t=1}^T$ be generated by the linearly restricted stochastic regression model. Fix $\delta \in (0,1)$. Suppose that Assumptions A1–A3 hold, $0 \prec \Gamma_{\text{sb}} \preceq \overline{\Gamma}$, and*

$$T \geq \frac{9k}{\alpha^2}\left\{m \log \frac{27}{\alpha} + \frac{1}{2}\log\det(\overline{\Gamma}_R \underline{\Gamma}_R^{-1}) + \log n + \log\frac{1}{\delta}\right\}. \qquad (\star)$$

Then, with probability at least $1 - 3\delta$, we have

$$\|\widehat{\beta} - \beta_*\|$$

$$\leq \frac{9\sigma}{\alpha}\left[\frac{\lambda_{\max}(R\underline{\Gamma}_R^{-1}R')}{T}\left\{12m\log\frac{14}{\alpha} + 9\log\det(\overline{\Gamma}_R\underline{\Gamma}_R^{-1}) + 6\log\frac{1}{\delta}\right\}\right]^{1/2}.$$

General upper bound for $\|\widehat{A} - A_*\|_2$

Proposition 1: *Under the conditions of Theorem 1, with probability at least* $1 - 3\delta$*, we have*

$$\|\widehat{A} - A_*\|_2 \leq \frac{9\sigma}{\alpha} \left[\frac{\lambda_{\max}\left(\sum_{i=1}^{n} R_i \underline{\Gamma}_R^{-1} R_i'\right)}{T} \left\{ 12m \log \frac{14}{\alpha} + 9 \log \det(\overline{\Gamma}_R \underline{\Gamma}_R^{-1}) + 6 \log \frac{1}{\delta} \right\} \right]^{1/2}.$$

Linearly restricted VAR models

Notations for the VAR(1) representation

- We consider the model with $Y_t = X_{t+1} \in \mathbb{R}^d$, i.e., $\{X_t\}_{t=1}^{T+1}$ generated by

$$X_{t+1} = A_* X_t + \eta_t, \tag{5}$$

subject to

$$\beta_* = R\theta_*,$$

where $\beta_* = \text{vec}(A_*') \in \mathbb{R}^{d^2}$, $\theta_* \in \mathbb{R}^m$, $R = (R_1', \ldots, R_d')' \in \mathbb{R}^{d^2 \times m}$, and R_i are $d \times m$ matrices. $\{X_t\}$ is adapted to the filtration

$$\mathcal{F}_t = \sigma\{\eta_1, \ldots, \eta_{t-1}\}.$$

Representative examples

Example 1 (VAR(p) model)

- Interestingly, VAR models of order $p < \infty$ can be viewed as linearly restricted VAR(1) models. Consider the VAR(p) model

$$Z_{t+1} = A_{*1} Z_t + A_{*2} Z_{t-1} + \cdots + A_{*p} Z_{t-p+1} + \varepsilon_t, \tag{6}$$

where $Z_t, \varepsilon_t \in \mathbb{R}^{d_0}$, and $A_{*i} \in \mathbb{R}^{d_0 \times d_0}$ for $i = 1, \ldots, p$.

- Let $X_t = (Z_t', Z_{t-1}', \ldots, Z_{t-p+1}')' \in \mathbb{R}^d$, $\eta_t = (\varepsilon_t', 0, \ldots, 0)' \in \mathbb{R}^d$, and

$$A_* = \begin{pmatrix} A_{*1} & \cdots & A_{*p-1} & A_{*p} \\ I_{d_0} & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & I_{d_0} & 0 \end{pmatrix} \in \mathbb{R}^{d \times d}, \tag{7}$$

where $d = d_0 p$. As a result, (6) can be written exactly as the VAR(1) model in the previous slide.

Representative examples

Example 2 (Banded VAR model)

- Zero restrictions:

$$a_{*ij} = 0, \qquad |i-j| > k_0, \tag{8}$$

  where the integer $1 \le k_0 \le \lfloor (d-1)/2 \rfloor$ is called the bandwidth parameter.

- In this case, R is a block diagonal matrix:

$$R = \begin{pmatrix} R_{(1)} & & 0 \\ & \ddots & \\ 0 & & R_{(d)} \end{pmatrix} \in \mathbb{R}^{d^2 \times m}, \tag{9}$$

Example 3 (Network VAR model)

Representative examples

Example 4 (Pure unit-root process)

- Consider $A_* = \rho I$, where $\rho \in \mathbb{R}$ is the only unknown parameter.
- This can be imposed by setting $R = (e_1', \dots, e_d')' \in \mathbb{R}^{d^2}$, where e_i is the $d \times 1$ unit vector with the i-th being one.
- When $\rho = 1$, it becomes the pure unit-root process, a classic example of unstable VAR processes; e.g., the problem of testing $A_* = I$ has been studied extensively in the asymptotic literature.
- Our non-asymptotic approach can precisely characterize the behavior of the estimator $\widehat{\rho}$ over a continuous range of $|\rho| \in [0, 1 + c/T]$.

Verification of regularity conditions in Theorem 1

We will replace Assumptions A1–A3 with the following:

A4. (i) The process $\{X_t\}$ starts at $t = 0$, with $X_0 = 0$.

(ii) The innovations $\{\eta_t\}$ are independent and $N(0, \sigma^2 I_d)$.

Assumption A4 paves the way to the unified analysis of stable and unstable processes via the finite-time controllability Gramian

$$\Gamma_t = \sum_{s=0}^{t-1} A_*^s (A_*')^s, \tag{10}$$

a key quantity closely related to $\mathrm{var}(X_t)$.

Why do we need to fix $X_0 = 0$?

- Under this assumption, it holds

$$X_t = \eta_{t-1} + A_* \eta_{t-2} + \cdots + A_*^{t-1}\eta_0 + A_*^t X_0 = \sum_{s=0}^{t-1} A_*^s \eta_{t-s-1}, \qquad t \geq 1,$$

which yields

$$\mathrm{var}(X_t) = E(X_t X_t') = \sigma^2 \Gamma_t. \tag{11}$$

- This highlights a subtle but critical difference from the typical set-up in the asymptotic theory where X_t starts at $t = -\infty$, so that

$$X_t = \sum_{s=0}^{\infty} A_*^s \eta_{t-s-1}, \qquad t \in \mathbb{Z},$$

which implies that $\mathrm{var}(X_t) < \infty$ *if and only if* the spectral radius $\rho(A_*) = \max\{|\lambda_1|, \ldots, |\lambda_d|\} < 1$ (when the process is stable), and if $\rho(A_*) < 1$, then $\mathrm{var}(X_t) = \sigma^2 \sum_{s=0}^{\infty} A_*^s (A_*')^s = \sigma^2 \lim_{t\to\infty} \Gamma_t$.

Verifying Assumptions A1–A3

Lemma 1: *Let* $\{X_t\}_{t=1}^{T+1}$ *be generated by the linearly restricted VAR model. Under Assumption A4, we have the following results:*

(i) *for any* $1 \leq k \leq \lfloor T/2 \rfloor$, $\{X_t\}_{t=1}^{T}$ *satisfies the* $(2k, \Gamma_{\mathrm{sb}}, 3/20)$*-BMSB condition, where* $\Gamma_{\mathrm{sb}} = \sigma^2 \Gamma_k$*; and*

(ii) *for any* $\delta \in (0,1)$*, it holds that* $\mathbb{P}(Z'Z \npreceq T\overline{\Gamma}_R) \leq \delta$*, where* $\overline{\Gamma}_R$ *is defined as before with* $n = d$ *and* $\overline{\Gamma} = \sigma^2 m \Gamma_T / \delta$*.*

Applying the general result in Theorem 1

Theorem 1 revisited: *Let* $\{(X_t, Y_t)\}_{t=1}^{T}$ *be generated by the linearly restricted stochastic regression model. Fix* $\delta \in (0,1)$*. Suppose that Assumptions A1–A3 hold,* $0 \prec \Gamma_{\text{sb}} \preceq \overline{\Gamma}$*, and*

$$T \geq \frac{9k}{\alpha^2}\left\{m \log \frac{27}{\alpha} + \frac{1}{2}\log\det(\overline{\Gamma}_R \underline{\Gamma}_R^{-1}) + \log n + \log\frac{1}{\delta}\right\}. \qquad (\star)$$

Then, with probability at least $1-3\delta$*, we have*

$$\|\widehat{\beta} - \beta_*\|$$

$$\leq \frac{9\sigma}{\alpha}\left[\frac{\lambda_{\max}(R\underline{\Gamma}_R^{-1}R')}{T}\left\{12m\log\frac{14}{\alpha} + 9\log\det(\overline{\Gamma}_R \underline{\Gamma}_R^{-1}) + 6\log\frac{1}{\delta}\right\}\right]^{1/2}.$$

By Lemma 1, the matrices $\overline{\Gamma}_R$ and $\underline{\Gamma}_R$ in Theorem 1 become

$$\overline{\Gamma}_R = \sigma^2 m R'(I_d \otimes \Gamma_T)R/\delta \quad \text{and} \quad \underline{\Gamma}_R = \sigma^2 R'(I_d \otimes \Gamma_k)R,$$

where $1 \leq k \leq \lfloor T/2 \rfloor$. We need to verify the existence of k satisfying $(\star)$.

Verifying the existence of k

$$\log\det(\overline{\Gamma}_R \underline{\Gamma}_R^{-1}) = m\log(m/\delta) + \underbrace{\log\det\left[R'(I_d\otimes\Gamma_T)R\{R'(I_d\otimes\Gamma_k)R\}^{-1}\right]}_{\kappa_R(T,k)}.$$

We need to derive an explicit upper bound for $\kappa_R(T,k)$. Recall that

$$\Gamma_t = \sum_{s=0}^{t-1} A_*^s (A_*')^s.$$

Main idea:

- Since $0 \prec I_d \preceq \Gamma_k \preceq \Gamma_T$, we have $\kappa_R(T,k) \leq \kappa_R(T,1)$.
- Note that Γ_T behaves differently in stable and unstable regimes: if $\rho(A_*) < 1$, then $\Gamma_T \preceq \Gamma_\infty = \lim_{T\to\infty}\Gamma_T < \infty$, and therefore

  $$\kappa_R(T,1) \leq \kappa_R(\infty,1).$$

  However, if $\rho(A_*) \geq 1$, then Γ_∞ no longer exists, so, we need to carefully control the growth rate of Γ_T as T increases.

Verifying the existence of k

- ... to do so, we consider the Jordan decomposition:

$$A_* = SJS^{-1}, \tag{12}$$

where J has L blocks with sizes

$$1 \leq b_1, \ldots, b_L \leq d,$$

and both J and S are $d \times d$ complex matrices. Let

$$b_{\max} = \max_{1 \leq \ell \leq L} b_\ell,$$

and denote the condition number of S by

$$\mathrm{cond}(S) = \left\{\lambda_{\max}(S^* S)/\lambda_{\min}(S^* S)\right\}^{1/2},$$

where S^* is the conjugate transpose of S.

Upper bound on $\kappa_R(\infty, 1)(\geq \kappa_R(\infty, k))$

Proposition 1: *For any $A_* \in \mathbb{R}^{d\times d}$, we have the following results:*

(i) *If $\rho(A_*) \leq 1 + c/T$ for a fixed $c > 0$, then*

$$\kappa_R(T, 1) \lesssim m \left\{\log \operatorname{cond}(S) + \log d + b_{\max} \log T\right\}.$$

(ii) *In particular, if $\rho(A_*) < 1$ and $\sigma_{\max}(A_*) \leq C$ for a fixed $C > 0$, then*

$$\kappa_R(T, 1) \lesssim m.$$

Implication: Provided that $\sigma_{\max}(A_*) \leq C$, the results from Theorem 1 will be different for the stable regime ($\rho(A_*) < 1$) and the unstable regime ($1 \leq \rho(A_*) \leq 1 + c/T$) in both

- the feasible region for k (becomes larger)
- and the upper bound of $\|\widehat{\beta} - \beta_*\|$ (becomes smaller)

(as the upper bound of $\kappa_R(T, k)$ becomes smaller)

Feasible region for k

By Proposition 1, we obtain the following sufficient conditions for ($\star$):

$$k \lesssim \begin{cases} \dfrac{T}{m\left[\log\{md\,\mathrm{cond}(S)/\delta\} + b_{\max}\log T\right]}, & \text{if } \rho(A_*) \leq 1 + c/T, \\ \dfrac{T}{m\log(m/\delta) + \log d}, & \text{if } \rho(A_*) < 1 \text{ and } \sigma_{\max}(A_*) \leq C. \end{cases}$$

We refer to this condition as ($\bigstar$) in the following slides.

Analysis of upper bounds in VAR model

Denote

$$\Gamma_{R,k} = R\left\{R'(I_d \otimes \Gamma_k)R\right\}^{-1} R'.$$

Theorem 2: *Let $\{X_t\}_{t=1}^{T+1}$ be generated by the linearly restricted VAR model. Fix $\delta \in (0,1)$. For any $1 \le k \le \lfloor T/2 \rfloor$ satisfying (★), under Assumption A4, we have the following results:*

(i) *If $\rho(A_*) \le 1 + c/T$ for a fixed $c > 0$, then, with probability at least $1 - 3\delta$, we have*

$$\|\widehat{\beta} - \beta_*\| \lesssim \left(\lambda_{\max}(\Gamma_{R,k}) \frac{m\left[\log\left\{md\,\mathrm{cond}(S)/\delta\right\} + b_{\max}\log T\right]}{T}\right)^{1/2}.$$

(ii) *In particular, if $\rho(A_*) < 1$ and $\sigma_{\max}(A_*) \le C$ for a fixed $C > 0$, then, with probability at least $1 - 3\delta$, we have*

$$\|\widehat{\beta} - \beta_*\| \lesssim \left\{\lambda_{\max}(\Gamma_{R,k}) \frac{m\log(m/\delta)}{T}\right\}^{1/2}.$$

Understanding the scale factor $\lambda_{\max}(\Gamma_{R,k})$

This scale factor may be viewed as a low-dimensional property:

- The limiting distribution of $\widehat{\beta}$ under the assumptions that d is fixed (and so are m and A_*) and $\rho(A_*) < 1$ is

$$T^{1/2}(\widehat{\beta} - \beta_*) \to N(0, \underbrace{R\{R'(I_d \otimes \Gamma_\infty)R\}^{-1}R'}_{\lim_{k\to\infty} \lambda_{\max}(\Gamma_{R,k})}) \tag{13}$$

  in distribution as $T \to \infty$, where $\Gamma_\infty = \lim_{k\to\infty} \Gamma_k$.

- The strength of our non-asymptotic approach is signified by the preservation of this scale factor in the error bounds.

  The key is to *simultaneously* bound $Z'Z$ and $Z'\eta$ through the Moore-Penrose pseudoinverse $Z^\dagger$. (Recall that $Z^\dagger = (Z'Z)^{-1}Z'$ if $Z'Z \succ 0$)

Insight from Theorem 2

Adding more restrictions will reduce the error bounds through not only the reduced model size m, but also the reduced scale factor $\lambda_{\max}(\Gamma_{R,k})$.

- To illustrate this, suppose that $\beta_* = R\theta_* = R^{(1)}R^{(2)}\theta_*$, where $R^{(1)} \in \mathbb{R}^{d^2 \times \widetilde{m}}$ has rank $\widetilde{m}$, and $R^{(2)} \in \mathbb{R}^{\widetilde{m} \times m}$ has rank m, with $\widetilde{m} \geq m + 1$.

- Then $\mathcal{L}^{(1)} = \{R^{(1)}\theta : \theta \in \mathbb{R}^{\widetilde{m}}\} \supseteq \mathcal{L} = \{R\theta : \theta \in \mathbb{R}^m\}$.

- If the estimation is conducted on the larger parameter space $\mathcal{L}^{(1)}$, then the scale factor in the error bound will become $\lambda_{\max}(\Gamma_{R^{(1)},k})$, and the (effective) model size will increase to $\widetilde{m}$.

- it can be shown that

$$\lambda_{\max}(\Gamma_{R,k}) \leq \lambda_{\max}(\Gamma_{R^{(1)},k}).$$

Asymptotic rates implied by Theorem 2

Note that

$$\lambda_{\max}(\Gamma_{R,k}) \le \lambda_{\max}\{R(R'R)^{-1}R'\} = \lambda_{\max}\{(R'R)^{-1}R'R\} = 1.$$

Corollary 1: *Under the conditions of Theorem 2, the following results hold:*

(i) *If* $\rho(A_*) \le 1 + c/T$ *for a fixed* $c > 0$, *then*

$$\|\widehat{\beta} - \beta_*\| = O_p\left\{\left(\frac{m\left[\log\{md\,\mathrm{cond}(S)\} + b_{\max}\log T\right]}{T}\right)^{1/2}\right\}.$$

(ii) *In particular, if* $\rho(A_*) < 1$ *and* $\sigma_{\max}(A_*) \le C$ *for a fixed* $C > 0$, *then*

$$\|\widehat{\beta} - \beta_*\| = O_p\left\{\left(\frac{m\log m}{T}\right)^{1/2}\right\}.$$

Strengthening Theorem 2: leveraging k

- Note that $\lambda_{\max}(\Gamma_{R,k})$ is monotonic decreasing in k.
- By choosing the largest possible k, we can obtain the sharpest possible result from Theorem 2.
- We will capture the magnitude of $\lambda_{\max}(\Gamma_{R,k})$ via $\sigma_{\min}(A_*)$, a measure of the least excitable mode of the underlying dynamics.
- This allows us to uncover a split between the slow and fast error rate regimes in terms of $\sigma_{\min}(A_*)$.

Theorem 3

Fix $\delta \in (0,1)$, *and suppose that the conditions of Theorem 2 hold.*

(i) *If* $\rho(A_*) \leq 1 + c/T$ *for a fixed* $c > 0$, *then we have the following results:*

When

$$\sigma_{\min}(A_*) \leq 1 - \frac{c_1 m \left[\log\{md\,\mathrm{cond}(S)/\delta\} + b_{\max}\log T\right]}{T}, \qquad \text{(A1)}$$

where $c_1 > 0$ *is fixed, with probability at least* $1 - 3\delta$, *we have*

$$\|\widehat{\beta} - \beta_*\| \lesssim \left(\frac{\{1 - \sigma^2_{\min}(A_*)\} m \left[\log\{md\,\mathrm{cond}(S)/\delta\} + b_{\max}\log T\right]}{T}\right)^{1/2}; \qquad \text{(S1)}$$

and when the inequality in (A1) *holds in the reverse direction, with probability at least* $1 - 3\delta$, *we have*

$$\|\widehat{\beta} - \beta_*\| \lesssim \frac{m\left[\log\{md\,\mathrm{cond}(S)/\delta\} + b_{\max}\log T\right]}{T}. \qquad \text{(F1)}$$

Theorem 3 cont'd

(ii) *In particular, if* $\rho(A_*) < 1$ *and* $\sigma_{\max}(A_*) \leq C$ *for a fixed* $C > 0$*, then we have the following results:*
When

$$\sigma_{\min}(A_*) \leq 1 - \frac{c_2\{m\log(m/\delta) + \log d\}}{T}, \tag{A2}$$

where $c_2 > 0$ *is fixed, with probability at least* $1 - 3\delta$*, we have*

$$\|\widehat{\beta} - \beta_*\| \lesssim \left[\frac{\{1 - \sigma^2_{\min}(A_*)\}m\log(m/\delta)}{T}\right]^{1/2}; \tag{S2}$$

and when the inequality in (A2) *holds in the reverse direction, with probability at least* $1 - 3\delta$*, we have*

$$\|\widehat{\beta} - \beta_*\| \lesssim \frac{m\log(m/\delta)}{T}. \tag{F2}$$

A simple example: $A_* = \rho I_d$

Note that the smallest true model has size one, and hence we may fit any larger model with $m \geq 1$. Moreover, we have

$$\rho(A_*) = \sigma_{\min}(A_*) = |\rho|, \quad \mathrm{cond}(S) = 1 \quad \text{and} \quad b_{\max} = 1.$$

Then, by Theorem 3:

(a) If $|\rho| \leq 1 - O\{(m \log m + \log d)/T\}$, then $\|\widehat{\beta} - \beta_*\| \lesssim O\{\sqrt{(1-\rho^2) m \log m / T}\}$, w.h.p.; see (S2).

(b) If $1 - O\{(m \log m + \log d)/T\} \leq |\rho| < 1$, then $\|\widehat{\beta} - \beta_*\| \lesssim O(T^{-1} m \log m)$, w.h.p.; see (F2).

(c) If $1 \leq |\rho| \leq 1 + O(1/T)$, then $\|\widehat{\beta} - \beta_*\| \lesssim O\{T^{-1} m \log(mdT)\}$, w.h.p.; see (F1).

Analysis of lower bounds

Notations

For a fixed $\bar{\rho} > 0$, we consider the subspace of θ such that the spectral radius of $A(\theta)$ is bounded above by $\bar{\rho}$, i.e.,

$$\Theta(\bar{\rho}) = \{\theta \in \mathbb{R}^m : \rho\{A(\theta)\} \leq \bar{\rho}\}.$$

Then, the corresponding linearly restricted subspace of β is

$$\mathcal{L}(\bar{\rho}) = \{R\theta : \theta \in \Theta(\bar{\rho})\}.$$

Denote by $\mathbb{P}_\theta^{(T)}$ the distribution of the sample $(X_1, \ldots, X_{T+1})$ on the space $(\mathcal{X}^{T+1}, \mathcal{F}_{T+1})$.

Analysis of lower bounds

Theorem 4: *Suppose that $\{X_t\}_{t=1}^{T+1}$ follow the VAR model $X_{t+1} = AX_t + \eta_t$, with linear restrictions defined previously, and Assumption A4 holds. Fix $\delta \in (0, 1/4)$ and $\bar{\rho} > 0$. Let*

$$\gamma_T(\bar{\rho}) = \sum_{s=0}^{T-1} \bar{\rho}^{2s}.$$

Then, for any $\epsilon \in (0, \bar{\rho}/4]$, we have

$$\inf_{\widehat{\beta}} \sup_{\theta \in \Theta(\bar{\rho})} \mathbb{P}_\theta^{(T)} \left\{ \|\widehat{\beta} - \beta\| \geq \epsilon \right\} \geq \delta,$$

where the infimum is taken over all estimators of β subject to $\beta \in \{R\theta : \theta \in \mathbb{R}^m\}$, for any T such that

$$T\gamma_T(\bar{\rho}) \lesssim \frac{m + \log(1/\delta)}{\epsilon^2}.$$

Asymptotic rates implied by Theorem 4

Corollary 2: *The minimax rates of estimation over* $\beta \in \mathcal{L}(\bar{\rho})$ *in different stability regimes are as follows:*

(i) $\sqrt{(1-\bar{\rho}^2)m/T}$, *if* $\bar{\rho} \in (0, \sqrt{1-1/T})$;

(ii) $T^{-1}\sqrt{m}$, *if* $\bar{\rho} \in [\sqrt{1-1/T}, 1+c/T]$ *for a fixed* $c > 0$; *and*

(iii) $\bar{\rho}^{-T}\sqrt{(\bar{\rho}^2-1)m/T}$, *if* $\bar{\rho} \in (1+c/T, \infty)$.

Discussion

The following directions are worth exploring in the future:

- The small-ball method is known for its capability to accommodate heavy tailed data. It may be possible to drop the normality assumption of the innovations.
- In addition, one may consider the recovery of unknown restriction patterns by methods such as information criteria or regularization, e.g., the fussed lasso (Ke et al., 2015).
- Similar non-asymptotic theory for possibly unstable, low rank (Ahn and Reinsel, 1988; Negahban and Wainwright, 2011) or cointegrated (Onatski and Wang, 2018) VAR models, which would be useful for high dimensional inference.
- ...

References I

Ahn, S. K. and Reinsel, G. C. (1988). Nested reduced-rank autogressive models for multiple time series. *Journal of the American Statistical Association*, 83:849–856.

Basu, S. and Michailidis, G. (2015). Regularized estimation in sparse high-dimensional time series models. *The Annals of Statistics*, 43:1535–1567.

Davis, R. A., Zang, P., and Zheng, T. (2015). Sparse vector autoregressive modeling. *Journal of Computational and Graphical Statistics*, 25:1077–1096.

Guo, S., Wang, Y., and Yao, Q. (2016). High-dimensional and banded vector autoregressions. *Biometrika*, 103:889–903.

Han, F., Lu, H., and Liu, H. (2015). A direct estimation of high dimensional stationary vector autoregressions. *Journal of Machine Learning Research*, 16:3115–3150.

Ke, Z. T., Fan, J., and Wu, Y. (2015). Homogeneity pursuit. *Journal of the American Statistical Association*, 110:175–194.

Negahban, S. and Wainwright, M. J. (2011). Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics*, 39:1069–1097.

Onatski, A. and Wang, C. (2018). Alternative asymptotics for cointegration tests in large VARs. *Econometrica*, 86:1465–1478.

Simchowitz, M., Mania, H., Tu, S., Jordan, M., and Recht, B. (2018). Learning without mixing: Towards a sharp analysis of linear system identification. In *Proceedings of Machine Learning Research*, volume 75, pages 439–473. 31st Annual Conference on Learning Theory.

Stock, J. H. and Watson, M. W. (2001). Vector autoregressions. *Journal of Economic Perspectives*, 15:101–115.

Zhu, X., Pan, R., Li, G., Liu, Y., and Wang, H. (2017). Network vector autoregression. *The Annals of Statistics*, 45:1096–1123.

Thank you!