Finite Time Analysis of Vector Autoregressive Models under Linear Restrictions

Yao Zheng

December 2019

Department of Statistics
University of Connecticut

A Broad Perspective

Low vs. High Dimensional Analysis of Time series

- Low dimensional setup
  - "fine-grained"

Conditional heteroscedasticity,
Heavy tails, Quantile inference,
Non-stationarity, ...

- High dimensional setup
  - "coarse"

Dimensionality reduction,
Non-asymptotic guarantees, ...

A "sharp" non-asymptotic analysis in high dimensions can uncover low dimensional phenomena.

Outline

1. Background
2. Small-Ball Method for Stochastic Regression
3. Application to VAR Models
4. Analysis of Lower Bounds
5. Conclusion and Discussion

Background

Vector Autoregressive (VAR) Model

For an observed d-dimensional time series $X_t \in \mathbb{R}^d$, VAR(1) model:

$$\begin{bmatrix} X_{t+1,1} \\ X_{t+1,2} \\ \vdots \\ X_{t+1,d} \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1d} \\ a_{21} & a_{22} & \cdots & a_{2d} \\ \vdots & \vdots & & \vdots \\ a_{d1} & a_{d2} & \cdots & a_{dd} \end{bmatrix} \begin{bmatrix} X_{t,1} \\ X_{t,2} \\ \vdots \\ X_{t,d} \end{bmatrix} + \begin{bmatrix} \eta_{t,1} \\ \eta_{t,2} \\ \vdots \\ \eta_{t,d} \end{bmatrix}, \quad t = 1, 2, \dots, n$$

$$X_{t+1} = A \quad X_t + \eta_t, \quad \eta_t \ i.i.d.\ E(\eta_t) = 0$$

where n is the sample size/time horizon (asymptotic analysis: $n \to \infty$).

Numerous applications: economics, finance, energy forecasting, ecological forecasting, neuroscience, health research, reinforcement learning, ...

Problem of Over-parameterization

- The unknown transition matrix A has d^2 parameters.
- For the general VAR(p) model

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

number of parameters $= O(pd^2)$.

- **Possible over-parametrization when d is even moderately large!**

$\Rightarrow$ **Cannot provide reliable estimates and forecasts without further restrictions.**

Literature: Taming the Dimensionality of Large VAR Models

(D). Direct reduction (our focus)

- Regularized estimation[a]
- Banded model[b]
- Network model[c]
- Other parameter restrictions motivated by specific applications

(I). Indirect reduction

- Reduced rank models
- Factor models
- ...

[a] Davis et al. (2015, JCGS), Han et al. (2015, JMLR), Basu and Michailidis (2015, AoS), ...
[b] Guo et al. (2016, Biometrika)
[c] Zhu et al. (2017, AoS)

Motivation of This Work

What most work in (D) has in common:

(i) **A particular sparsity or structural assumption is often imposed on the transition matrix A:** exact sparsity, banded/network structure, ...

(ii) **There is an almost exclusive focus on stable processes:**
i.e., imposing the spectral radius $\rho(A) < 1$, or even more stringently, the spectral norm $\|A\|_2 < 1$[a].

Motivation of This Work

What most work in (D) has in common:

(i) **A particular sparsity or structural assumption is often imposed on the transition matrix A:** exact sparsity, banded/network structure, ...

(ii) **There is an almost exclusive focus on stable processes:** i.e., imposing the spectral radius $\rho(A) < 1$, or even more stringently, the spectral norm $\|A\|_2 < 1$[a].

[a] 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 approach:

- **Linear restriction framework encompassing various existing models**
- **Allow unstable and even slightly explosive processes:**

$$\rho(A) \leq 1 + c/n$$

Our Objective

We 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 non-asymptotic analysis of the ordinary least squares (OLS) estimator for

- **possibly unstable and even slightly explosive** VAR models with $\rho(A) \leq 1 + c/n$
- under **linear restrictions** in the form of

$$\underbrace{\mathcal{C}}_{\textbf{known restriction matrix}} \quad \underbrace{\text{vec}(A^\top)}_{\textbf{stacking rows of A}} = \underbrace{\mu}_{\textbf{known vector}} ;$$

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

Linear Restriction Framework

For time-dependent pairs (X_t, Y_t), consider the unrestricted **multivariate stochastic regression:**

$$\underset{q\times 1}{Y_t} = \underset{q\times d}{A}\ \underset{d\times 1}{X_t} + \underset{q\times 1}{\eta_t}.$$

This includes VAR(p) models as special cases; VAR(1) if $Y_t = X_{t+1}$, $q = d$.

- Let $\beta = \underbrace{\text{vec}(A^\top)}_{\text{stacking rows of } A} \in \mathbb{R}^N$, where $N = qd$.
- Parameter space of a **linearly restricted** model: for $0 \leq m \leq N$,

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

where $\mathcal{C}$ and μ are known, and $\underbrace{\text{rank}(\mathcal{C}) = N - m}_{N - m \text{ independent restrictions}}$.

Equivalent Form

For simplicity, we restrict our attention to $\mu = 0$ in this talk. Note that

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

has **an equivalent, unrestricted parameterization:**

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

Specific relationship between $\mathcal{C}$ and R:

Let $\widetilde{\mathcal{C}}$ be an $m \times N$ complement of $\mathcal{C}$ such that $\mathcal{C}_{\text{full}} = (\widetilde{\mathcal{C}}^\top, \mathcal{C}^\top)^\top$ is invertible. Then let $\mathcal{C}_{\text{full}}^{-1} = (R, \widetilde{R})$, where R is an $N \times m$ matrix.

- 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, the above forms of $\mathcal{L}$ are equivalent.

Implications

- There exists a unique unrestricted $\theta_* \in \mathbb{R}^m$ such that

$$\underset{N\times 1}{\beta_*} = R \underset{m\times 1}{\theta_*}.$$

- 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 $\mathcal{C}$: Zero Restrictions

Recall:

$$\underset{N\times m}{R}\ \underset{m\times 1}{\theta} = \underset{N\times 1}{\beta} \quad \Leftrightarrow \quad \underset{(N-m)\times N}{\mathcal{C}}\ \underset{N\times 1}{\beta} = 0$$

Restricting the i-th element of β to zero: $\beta_i = 0$

- This can be encoded in R by setting the i-th row of R to zero.
- Alternatively, it can be encoded in $\mathcal{C}$ by setting a row of $\mathcal{C}$ to

$$(0, \ldots, 0, \underbrace{1}_{\text{the } i\text{-th entry}}, 0, \ldots, 0) \in \mathbb{R}^N.$$

How to Encode Restrictions via R or $\mathcal{C}$: Equality Restrictions

Recall:

$$\underset{N\times m}{R}\ \underset{m\times 1}{\theta} = \underset{N\times 1}{\beta} \quad \Leftrightarrow \quad \underset{(N-m)\times N}{\mathcal{C}}\ \underset{N\times 1}{\beta} = 0$$

Restricting that the i-th and j-th elements of β are equal: $\beta_i - \beta_j = 0$

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

$$(0, \ldots, 0, \underbrace{1}_{\text{the } k\text{-th entry}}, 0, \ldots, 0) \in \mathbb{R}^m.$$

- Alternatively, we may set a row of $\mathcal{C}$ to

$$(0, \ldots, 0, \underbrace{1}_{\text{the } i\text{-th entry}}, 0, \ldots, 0, \underbrace{-1}_{\text{the } j\text{-th entry}}, 0, \ldots, 0) \in \mathbb{R}^N.$$

Example 1: VAR(p) Models

VAR(p) model

$$\underset{d_0\times 1}{Z_{t+1}} = \underset{d_0\times d_0}{A_1} Z_t + \underset{d_0\times d_0}{A_2} Z_{t-1} + \cdots + \underset{d_0\times d_0}{A_p} Z_{t-p+1} + \varepsilon_t.$$

- Let $X_t = (Z_t^\top, Z_{t-1}^\top, \ldots, Z_{t-p+1}^\top)^\top \in \mathbb{R}^d$ and $\eta_t = (\varepsilon_t^\top, 0, \ldots, 0)^\top \in \mathbb{R}^d$, where $d = d_0 p$. Then

$$\underbrace{\begin{pmatrix} Z_{t+1} \\ Z_t \\ \vdots \\ Z_{t-p+2} \end{pmatrix}}_{X_{t+1}} = \underbrace{\begin{pmatrix} A_1 & \cdots & A_{p-1} & A_p \\ I_{d_0} & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & I_{d_0} & 0 \end{pmatrix}}_{A} \underbrace{\begin{pmatrix} Z_t \\ Z_{t-1} \\ \vdots \\ Z_{t-p+1} \end{pmatrix}}_{X_t} + \underbrace{\begin{pmatrix} \varepsilon_t \\ 0 \\ \vdots \\ 0 \end{pmatrix}}_{\eta_t}$$

- **Thus, VAR(p) models can be viewed as linearly restricted VAR(1) models.**

We may focus on VAR(1) models from now on.

Example 2: Banded VAR Model

Banded VAR model

In practice, it is often sufficient to collect information from "neighbors":

$$a_{ij} = 0 \quad \forall\, |i-j| > k_0.$$

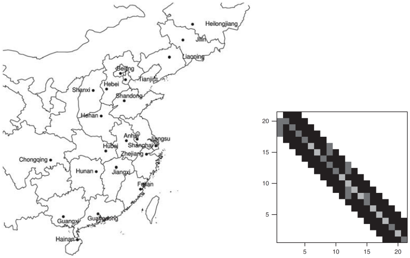


Figure 1: Location plot and estimated transition matrix $\hat{A}$ (Guo et al., 2016, Biometrika).

In this case, $\mu = 0$ and

$$R = \begin{pmatrix} R_{(1)} & & 0 \\ & \ddots & \\ 0 & & R_{(d)} \end{pmatrix}$$

is a $d^2 \times m$ block diagonal matrix.

Actually, the definition of "neighbors" can be more general.

Example 3: Network VAR Model

Network VAR model

To analyze users' time series data from large social networks, Zhu et al. (2017, AoS) imposes that

- $a_{11} = \cdots = a_{dd}$;
- the zero-nonzero pattern of A is known: $a_{ij} = 0$ **if individual** j **does not follow individual** i **on the social network**;
- all nonzero off-diagonal entries of A are equal.

This model is essentially low-dimensional.

Example 4: Pure Unit Root Process

Pure unit root process

$$A = \rho I_d, \quad \text{where} \quad \rho \in \mathbb{R}.$$

If $\rho = 1$, it is the pure unit root process, a classic unstable VAR process.

- If all restrictions are imposed (only ρ is unknown), then $R = (e_1^\top, \ldots, e_d^\top)^\top \in \mathbb{R}^{d^2}$, with $e_i = (0, \ldots, 0, \underbrace{1}_{\text{the } i\text{-th entry}}, 0, \ldots, 0)^\top \in \mathbb{R}^d$.
- Testing $H_0 : A_* = I_d$ (unit root testing in panel data) has been extensively studied in the asymptotic literature.[a]
- Our non-asymptotic approach can precisely characterize the behavior of $\widehat{\rho}$ over $|\rho| \in (0, 1 + c/n]$.

[a] See Chang (2004, JoE) and Zhang et al. (2018, AoS) for low and high dimensional cases, respectively.

Ordinary Least Squares (OLS) Estimation

- We can define the OLS estimator under the general **multivariate stochastic regression** framework:

$$\underset{q\times 1}{Y_t} = \underset{q\times d}{A_*} \underset{d\times 1}{X_t} + \underset{q\times 1}{\eta_t}, \tag{1}$$

where A_* is the true value. Then (1) has the matrix form

$$\underbrace{\begin{pmatrix} Y_1^\top \\ \vdots \\ Y_n^\top \end{pmatrix}}_{n\times q} = \underbrace{\begin{pmatrix} X_1^\top \\ \vdots \\ X_n^\top \end{pmatrix}}_{n\times d} \underbrace{A_*^\top}_{d\times q} + \underbrace{\begin{pmatrix} \eta_1^\top \\ \vdots \\ \eta_n^\top \end{pmatrix}}_{n\times q},$$

$$\text{i.e.,} \quad Y \quad = \quad X \quad A_*^\top \quad + \quad E.$$

- By vectorization, $\underbrace{\text{vec}(Y)}_{y} = (I_q \otimes X)\underbrace{\text{vec}(A_*^\top)}_{\beta_*} + \underbrace{\text{vec}(E)}_{\eta}$.

Ordinary Least Squares (OLS) Estimation

- Here we let

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

- By reparameterization, we further have

$$y = (I_q \otimes X)\beta_* + \eta = \underbrace{(I_q \otimes X)R}_{Z}\,\theta_* + \eta = Z\theta_* + \eta.$$

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

$$\widehat{\beta} = R\widehat{\theta}, \quad \text{where} \quad \widehat{\theta} = \arg\min_{\theta \in \mathbb{R}^m} \|y - \underbrace{Z}_{qn \times m}\theta\|^2. \tag{2}$$

[a]To ensure the feasibility of (2), we assume that $qn \geq m$. (But Z need not be full rank).

Ordinary Least Squares (OLS) Estimation

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

$$A_* = (I_q \otimes \theta_*^\top)\widetilde{R},$$

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

$$\widetilde{R} = (R_1, \ldots, R_q)^\top.$$

Hence, we can obtain the OLS estimator of A by

$$\widehat{A} = (I_q \otimes \widehat{\theta}^\top)\widetilde{R}.$$

- Note that $\|\widehat{\beta} - \beta_*\| = \|\widehat{A} - A_*\|_F$.

A Sneak Peek of Our Results

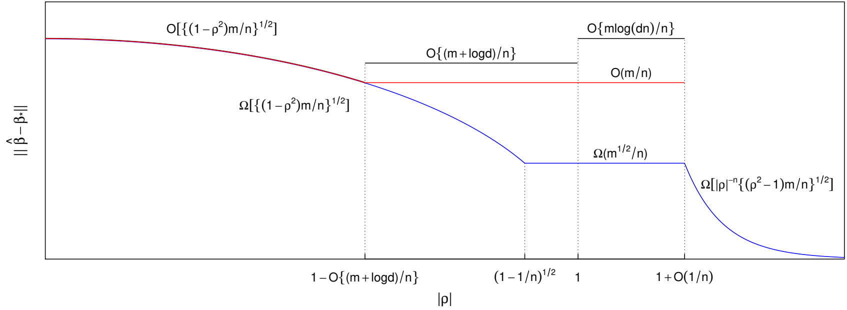


Figure 2: Illustration of theoretical upper (black) and lower (blue) bounds and actual rates (red) suggested by simulation results for VAR(1) model with $A_* = \rho I_d$ and Gaussian innovations.

Small-Ball Method for Stochastic Regression

Key Technical Tool for Upper Bound Analysis

Extension of Mendelson's small-ball method to time-dependent data[a]

Why using this method

The small-ball method helps us establish lower bounds of the Gram matrix $X^\top X$ (or $Z^\top Z$) under very mild conditions, while dropping the stability assumption and avoiding reliance on mixing properties.

How to use this method

(a) Formulate a (pointwise) small-ball condition

(b) Use this condition to control the lower tail behavior of the Gram matrix

(c) **Derive upper bounds for the estimation error**

(d) Verify the small-ball condition - in our context, **for VAR models**

[a]Simchowitz et al. (2018, COLT)

Main Idea of (a)→(b): Lower-Bounding $\lambda_{\min}(\sum_{t=1}^n X_t X_t^\mathsf{T})$

(1) Divide the data into **size-k blocks** along the time dimension, with the ℓ-th block being $\{X_{(\ell-1)k+1}, \dots, X_{\ell k}\}$.

(2) Lower-bound each $\sum_{i=1}^k \langle X_{(\ell-1)k+i}, \omega \rangle^2$ for $\omega \in \mathcal{S}^{d-1}$ with high probability by a small ball condition (defined in the next slide).

(3) Aggregate to get with probability at least $1 - \exp(-cn/k)$,

$$\frac{1}{n} \sum_{t=1}^n \langle X_t, \omega \rangle^2 \gtrsim \omega^\mathsf{T} \Gamma_{\mathrm{sb}} \omega.$$

(4) By the covering method, strengthen the pointwise bound into a lower bound on

$$\inf_{\omega \in \mathcal{S}^{d-1}} \sum_{t=1}^n \langle X_t, \omega \rangle^2,$$

where $\mathcal{S}^{d-1} = \{\omega \in \mathbb{R}^d : \|\omega\| = 1\}$ is the unit sphere in $\mathbb{R}^d$.

Small-Ball Condition for Dependent Data

Block martingale small ball (BMSB) condition: 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 $s \geq 0$,
>
> $$\frac{1}{k} \sum_{t=1}^{k} \mathbb{P}(|X_{s+t}| \geq \nu \mid \mathcal{F}_s) \geq \alpha$$
>
> with probability one.

Here, k is the block size.

Small-Ball Condition for Dependent Data

Block martingale small ball (BMSB) condition: 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

$$\{\omega^{\mathsf{T}} X_t, t = 1, 2, \dots\}$$

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

Regularity Conditions for Upper Bound Analysis

Assumptions for multivariate stochastic regression

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

A2. For any $\delta \in (0,1)$, there exists $\overline{\Gamma}_R$ dependent on δ such that

$$\mathbb{P}(Z^\top Z \npreceq n\overline{\Gamma}_R) \leq \delta.$$

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

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

Assumptions A1 and A2 will be verified (with specific Γ_{sb} and $\overline{\Gamma}_R$) for VAR models later.

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

Theorem 1 (General upper bound)

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

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

where $\underline{\Gamma}_R = R^{\prime}(I_q \otimes \Gamma_{\mathrm{sb}})R$. *Then, with probability at least* $1 - 3\delta$, *we have*

$$\|\widehat{\beta} - \beta_*\| \leq \frac{9\sigma}{\alpha}\sqrt{\frac{\lambda_{\max}(R\underline{\Gamma}_R^{-1}R^{\prime})}{n}\left\{12m\log\frac{14}{\alpha} + 9\log\det(\overline{\Gamma}_R\underline{\Gamma}_R^{-1}) + 6\log\frac{1}{\delta}\right\}}.$$

Similarly, we can also provide an upper bound for $\|\widehat{A} - A_*\|_2$.

Application to VAR Models

Properties of VAR(1) Model

$$X_{t+1} = A_* X_t + \eta_t, \qquad t = 1, \ldots, n,$$

subject to

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

where $\beta_* = \text{vec}(A_*^\intercal) \in \mathbb{R}^{d^2}$, $\theta_* \in \mathbb{R}^m$, and $R \in \mathbb{R}^{d^2 \times m}$. Then $\{X_t\}$ is adapted to the filtration $\mathcal{F}_t = \sigma\{\eta_1, \ldots, \eta_{t-1}\}$.

Assumptions for VAR model (Note: A4 $\Rightarrow$ A1–A3.)

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

(ii) The innovations $\{\eta_t\}$ are $i.i.d.$ with $E(\eta_t) = 0$ and $\text{var}(\eta_t) = \Sigma_\eta = \sigma^2 I_d$.

(iii) There is a universal constant $C_0 > 0$ such that, for every $\omega \in \mathcal{S}^{d-1}$, the density of $\omega^\intercal \Sigma_\eta^{-1/2} \eta_t$ is bounded by C_0 almost everywhere.

(iv) $\{\eta_t\}$ are σ^2-sub-Gaussian.

About Fixing X_0

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

Then

$$\mathrm{var}(X_t) = E(X_t X_t^\intercal) = \sigma^2 \Gamma_t,$$

where the **finite-time controllability Gramian**

$$\Gamma_t = \sum_{s=0}^{t-1} A_*^s (A_*^\intercal)^s.$$

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

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

About Fixing X_0

... which implies that

$$\mathrm{var}(X_t) < \infty \quad \textit{if and only if} \quad \rho(A_*) = \max\{|\lambda_1|, \ldots, |\lambda_d|\} < 1,$$

and if $\rho(A_*) < 1$, then

$$\mathrm{var}(X_t) = \sigma^2 \sum_{s=0}^{\infty} A_*^s (A_*^\top)^s = \sigma^2 \lim_{t\to\infty} \Gamma_t.$$

In contrast, by fixing X_0, we can provide a unified analysis of stable and unstable processes via the finite-time controllability Gramian Γ_t.

Assumption A4 $\Rightarrow$ A1

Lemma 1 (Verification of the BMSB condition)

Let $\{X_t\}_{t=1}^{n+1}$ be generated by the linearly restricted vector autoregressive model. Under Assumptions A4(ii) and (iii), for any $1 \le k \le \lfloor n/2 \rfloor$, $\{X_t\}_{t=1}^{n}$ satisfies the $(2k, \Gamma_{\mathrm{sb}}, 1/10)$-BMSB condition, where $\Gamma_{\mathrm{sb}} = \sigma^2 \Gamma_k / (4C_0)^2$.

By Lemma 1, for any $1 \le k \le \lfloor n/2 \rfloor$, the matrix $\underline{\Gamma}_R$ in Theorem 1 can be specified as

$$\underline{\Gamma}_R = \sigma^2 R^\top (I_d \otimes \Gamma_k) R / (4C_0)^2.$$

Assumption A4 $\Rightarrow$ A2: Two choices of $\overline{\Gamma}_R$

Lemma 2 (The first choice of $\overline{\Gamma}_R$)

Let $\{X_t\}_{t=1}^{n+1}$ be generated by the linearly restricted vector autoregressive model. Under Assumptions A4(i) and (ii), for any $\delta \in (0,1)$, it holds $\mathrm{pr}(Z^{\mathsf{T}}Z \not\succeq n\overline{\Gamma}_R) \leq \delta$, where $\overline{\Gamma}_R = R^{\mathsf{T}}(I_d \otimes \overline{\Gamma})R$, with $\overline{\Gamma} = \sigma^2 m \Gamma_n/\delta$.

By Lemma 2, the matrix $\overline{\Gamma}_R$ in Theorem 1 can be chosen as

$$\overline{\Gamma}_R = \overline{\Gamma}_R^{(1)} := \sigma^2 m R^{\mathsf{T}}(I_d \otimes \Gamma_n)R/\delta.$$

Assumption A4 $\Rightarrow$ A2: Two choices of $\overline{\Gamma}_R$

Let $\Sigma_X = [E(X_t X_s^\top)_{d\times d}]_{1\le t,s\le n}$ be the covariance matrix of the $dn \times 1$ vector $\mathrm{vec}(X^\top) = (X_1^\top, \ldots, X_n^\top)^\top$. Then, for a universal constant $C_1 > 0$, define $\psi(m,d,\delta) = C_1\{m\log 9 + \log d + \log(2/\delta)\}$, and

$$\xi = \xi(m,d,n,\delta) = 2\left\{\frac{\lambda_{\max}(\Gamma_n)\psi(m,d,\delta)\|\Sigma_X\|_2}{\sigma^2 n}\right\}^{1/2} + \frac{2\psi(m,d,\delta)\|\Sigma_X\|_2}{\sigma^2 n}.$$

Lemma 3 (The second choice of $\overline{\Gamma}_R$)

Let $\{X_t\}_{t=1}^{n+1}$ be generated by the linearly restricted vector autoregressive model. Under Assumptions A4(i) and (ii), if $\{\eta_t\}$ are normally distributed, then for any $\delta \in (0,1)$, it holds $\mathrm{pr}(Z^T Z \not\geq n\overline{\Gamma}_R) \le \delta$, where $\overline{\Gamma}_R = R^T(I_d \otimes \overline{\Gamma})R$, with $\overline{\Gamma} = \sigma^2\Gamma_n + \sigma^2\xi I_d$, and $\xi = \xi(m,d,n,\delta)$.

By Lemma 3, the matrix $\overline{\Gamma}_R$ in Theorem 1 can be chosen as

$$\overline{\Gamma}_R = \overline{\Gamma}_R^{(2)} := \sigma^2 R^\top(I_d \otimes \Gamma_n)R + \sigma^2\xi(m,d,n,\delta)R^\top R.$$

Theorem 1 applied to VAR(1) model)

Let $\{X_t\}_{t=1}^{n+1}$ be generated by the linearly restricted VAR model. Fix $\delta \in (0,1)$. Suppose that Assumption A4 hold and

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

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

$$\|\widehat{\beta} - \beta_*\| \leq \frac{9\sigma}{\alpha}\sqrt{\frac{\lambda_{\max}(R\underline{\Gamma}_R^{-1}R^T)}{n}\left\{12m \log \frac{14}{\alpha} + 9\log\det(\overline{\Gamma}_R \underline{\Gamma}_R^{-1}) + 6\log\frac{1}{\delta}\right\}}.$$

Here, $\underline{\Gamma}_R = \sigma^2 R^T(I_d \otimes \Gamma_k)R/(4C_0)^2$ with $1 \leq k \leq \lfloor n/2 \rfloor$, and $\overline{\Gamma}_R = \overline{\Gamma}_R^{(1)}$ or $\overline{\Gamma}_R^{(2)}$ (if $\{\eta_t\}$ are normally distributed), where

$$\overline{\Gamma}_R^{(1)} = \sigma^2 m R^T(I_d \otimes \Gamma_n)R/\delta,$$
$$\overline{\Gamma}_R^{(2)} = \sigma^2 R^T(I_d \otimes \Gamma_n)R + \sigma^2 \xi(m,d,n,\delta)R^T R.$$

Verifying the Existence of k in $(\star)$

- Obviously, without imposing normality on $\{\eta_t\}$, we can only choose $\overline{\Gamma}_R = \overline{\Gamma}_R^{(1)}$. However, if $\{\eta_t\}$ are normal, we can set $\overline{\Gamma}_R$ to whichever of $\overline{\Gamma}_R^{(1)}$ and $\overline{\Gamma}_R^{(2)}$ delivers the sharper upper bound.

- It can be shown that

$$\log\det(\overline{\Gamma}_R \underline{\Gamma}_R^{-1}) \lesssim \begin{cases} m\log(m/\delta) + \kappa, & \text{if} \quad \overline{\Gamma}_R = \overline{\Gamma}_R^{(1)} \\ m\log\{2\max(1,\xi)\} + \kappa, & \text{if} \quad \overline{\Gamma}_R = \overline{\Gamma}_R^{(2)} \end{cases},$$

  where $\xi = \xi(m, d, n, \delta)$ and $\kappa = \log\det\left\{R^\top (I_d \otimes \Gamma_n) R (R^\top R)^{-1}\right\}$.

Next goal: Derive explicit upper bounds for ξ and κ. Note that

- $\Gamma_n = \sum_{s=0}^{n-1} A_*^s (A_*^\top)^s \preceq \Gamma_\infty < \infty$ only if $\rho(A_*) < 1$.
- ξ depends on $\|\Sigma_X\|_2$, which also depends on n and is not necessarily bounded even if $\rho(A_*) < 1$. Recall $(\Sigma_X)_{t,s} = E(X_t X_s^\top) = \sigma^2 A_*^{t-s}\Gamma_s$ for $1 \le s \le t \le n$ (growing with s).

Upper Bounds on κ

Different cases of A_*:

A5. $\rho(A_*) \leq 1 + c/n$, where $c > 0$ is a universal constant.

A6. $\rho(A_*) \leq \bar{\rho} < 1$ and $\|A_*\|_2 \leq C$, where $C, \bar{\rho} > 0$ are universal constants.

Jordan decomposition: $A_* = SJS^{-1}$, where J has L blocks with maximum block size $b_{\max} = \max_{1 \leq \ell \leq L} b_\ell$, Let $\operatorname{cond}(S) = \{\lambda_{\max}(S^* S)/\lambda_{\min}(S^* S)\}^{1/2}$, where S^* is the conjugate transpose of S.

Lemma S7 (Upper bounds of κ)

For any $A_* \in \mathbb{R}^{d \times d}$, under Assumption A5,

$$\kappa \lesssim m \left[\log\{d \operatorname{cond}(S)\} + b_{\max} \log n\right].$$

Moreover, if Assumption A6 holds, then $\kappa \lesssim m$.

Simple example: $A_* = \rho I_d \quad \Rightarrow \quad b_{\max} = \operatorname{cond}(S) = 1.$

Upper Bounds on ξ

Different cases of A_*:

A5. $\rho(A_*) \leq 1 + c/n$, where $c > 0$ is a universal constant.

A6. $\rho(A_*) \leq \bar{\rho} < 1$ and $\|A_*\|_2 \leq C$, where $C, \bar{\rho} > 0$ are universal constants.

A7. $\rho(A_*) \leq \bar{\rho} < 1$, $\|A_*^t\|_2 \leq C\varrho^t$ for any integer $1 \leq t \leq n$, and $\mu_{\min}(\mathcal{A}) = \inf_{\|z\|=1} \lambda_{\min}(\mathcal{A}^*(z)\mathcal{A}(z)) \geq \mu_1$, where $C, \bar{\rho}, \mu_1 > 0$ and $\varrho \in (0,1)$ are universal constants, and $\mathcal{A}(z) = I_d - A_* z$ for $z \in \mathbb{C}$.

Lemma S8 (Upper bounds of ξ)

For any $A_* \in \mathbb{R}^{d\times d}$, under Assumption A5,

$$\log \xi \lesssim \log\{d \operatorname{cond}(S)\} + b_{\max} \log n.$$

Moreover, if Assumption A7 holds, then $\xi \lesssim 1$.

Feasible Region for k

Note: In Theorem 1, as the upper bound of $\log\det(\overline{\Gamma}_R \underline{\Gamma}_R^{-1})$ becomes smaller,

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

Sufficient condition for ($\star$):

$$k \lesssim \begin{cases} \frac{n}{m[\log\{d\,\mathrm{cond}(S)\} + b_{\max}\log n] + \log(1/\delta)}, & \text{if Assumption A5 holds} \\ \frac{n}{m\log(m/\delta) + \log d}, & \text{if Assumption A6 holds} \\ \frac{n}{m + \log(d/\delta)}, & \text{if Assumption A7 and } \{\eta_t\} \text{ are normal} \end{cases} . \quad (\bigstar)$$

Analysis of Upper Bounds for VAR Model

Denote $\Gamma_{R,k} = R\underline{\Gamma}_R^{-1}R^\top = R\left\{R^\top(I_d \otimes \Gamma_k)R\right\}^{-1}R^\top$ (decreasing in k).

Theorem 2 (Upper bounds for VAR model)

Let $\{X_t\}_{t=1}^{n+1}$ be generated by the linearly restricted VAR model. Fix $\delta \in (0,1)$. For any $1 \leq k \leq \lfloor n/2 \rfloor$ satisfying (★), under Assumption A4,

(i) *if Assumption A5 holds, with probability at least $1-3\delta$,*

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

(ii) *if Assumption A6 holds, with probability at least $1-3\delta$,*

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

(iii) *if Assumption A7 holds and $\{\eta_t\}$ are normal, with probability at least $1-3\delta$,*

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

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

$$n^{1/2}(\widehat{\beta} - \beta_*) \to N(0, \underbrace{R\{R^{\top}(I_d \otimes \Gamma_\infty)R\}^{-1}R^{\top}}_{\lim_{k\to\infty} \lambda_{\max}(\Gamma_{R,k})}) \qquad (3)$$

  in distribution as $n \to \infty$, where $\Gamma_\infty = \lim_{n\to\infty} \Gamma_n$.

- **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 $Z^{\top}Z$ and $Z^{\top}\eta$ through the Moore-Penrose pseudoinverse $Z^{\dagger}$. (Recall that $Z^{\dagger} = (Z^{\top}Z)^{-1}Z^{\top}$ if $Z^{\top}Z \succ 0$)

Insight from Theorem 2: Impact of Restrictions

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 (effective) model size will increase to $\widetilde{m}$, and the scale factor in the error bound will become $\lambda_{\max}(\Gamma_{R^{(1)},k})$,** where it can be shown that

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

Strengthening Theorem 2: Leveraging k

- Note that $\lambda_{\max}(\Gamma_{R,k})$ is monotonic decreasing in k.
- **By choosing the largest possible k satisfying ($\bigstar$), 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 phase transition from the slow to fast error rate regimes in terms of $\sigma_{\min}(A_*)$.**

A Sharper Analysis of Upper Bounds for VAR Model

Theorem 3 (Sharpened upper bounds for VAR model)

Suppose that the conditions of Theorem 2 hold. Fix $\delta \in (0,1)$, *and let* $c_1 > 0$ *be a universal constant.*

(i) *Under Assumption A5, if*

$$\sigma_{\min}(A_*) \leq 1 - \frac{c_1 \{m [\log\{d \operatorname{cond}(S)\} + b_{\max} \log n] + \log(1/\delta)\}}{n}, \quad \text{(A.1)}$$

then, with probability at least $1 - 3\delta$,

$$\|\widehat{\beta} - \beta_*\| \lesssim \sqrt{\frac{\{1 - \sigma_{\min}^2(A_*)\} \{m [\log\{d \operatorname{cond}(S)\} + b_{\max} \log n] + \log(1/\delta)\}}{n}}, \quad \text{(S.1)}$$

and if inequality (A.1) *holds in the reverse direction, then, with probability at least* $1 - 3\delta$,

$$\|\widehat{\beta} - \beta_*\| \lesssim \frac{m [\log\{d \operatorname{cond}(S)\} + b_{\max} \log n] + \log(1/\delta)}{n}. \quad \text{(F.1)}$$

A Sharper Analysis of Upper Bounds for VAR Model

Theorem 3 (Cont'd)

(ii) *Under Assumption A6, if*

$$\sigma_{\min}(A_*) \leq 1 - \frac{c_1\{m\log(m/\delta) + \log d\}}{n}, \tag{A.2}$$

then, with probability at least $1 - 3\delta$,

$$\|\widehat{\beta} - \beta_*\| \lesssim \sqrt{\frac{\{1 - \sigma^2_{\min}(A_*)\}m\log(m/\delta)}{n}}, \tag{S.2}$$

and if inequality (A.2) *holds in the reverse direction, then, with probability at least* $1 - 3\delta$,

$$\|\widehat{\beta} - \beta_*\| \lesssim \frac{m\log(m/\delta) + \log d}{n}. \tag{F.2}$$

A Sharper Analysis of Upper Bounds for VAR Model

Theorem 3 (Cont'd)

(ii) *Under Assumption A7, if*

$$\sigma_{\min}(A_*) \leq 1 - \frac{c_1\{m + \log(d/\delta)\}}{n}, \tag{A.3}$$

then, with probability at least $1 - 3\delta$*,*

$$\|\widehat{\beta} - \beta_*\| \lesssim \sqrt{\frac{\{1 - \sigma_{\min}^2(A_*)\}\{m + \log(1/\delta)\}}{n}}. \tag{S.3}$$

and if inequality (A.3) *holds in the reverse direction, then, with probability at least* $1 - 3\delta$*,*

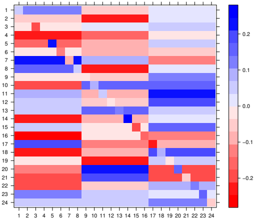
$$\|\widehat{\beta} - \beta_*\| \lesssim \frac{m + \log(d/\delta)}{n}. \tag{F.3}$$

Simulation Experiment

Three data generating processes (DGPs) with $\eta_t \overset{i.i.d.}{\sim} N(0, I_d)$:

- DGP1 (banded structure): $a_{*ij} = 0$ if $|i - j| > k_0$, where $k_0 \geq 1$ is the bandwidth parameter. $\Rightarrow m = d + (2d - 1)k_0 - k_0^2$

- DGP2 (group structure): X_t is equally partitioned into K groups. In each row of A_*, the off-diagonal entries a_{*ij} with j belonging to the same group are assumed to be equal. $\Rightarrow m = (K + 1)d$



- DGP3: $A_* = \rho I_d$, where $\rho \in \mathbb{R}$. $\Rightarrow m \geq 1$

Simulation Results

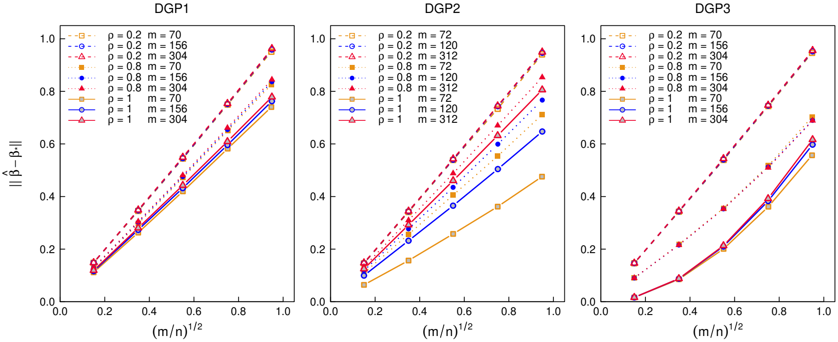


Figure 3: Plots of $\|\widehat{\beta} - \beta_*\|$ against $(m/n)^{1/2}$ for three data generating processes with $\rho(A_*) = 0.2$, 0.8 or 1 and different m. $\mathrm{DGP1}$ and $\mathrm{DGP3}$ were fitted as banded vector autoregressive models with $m = 70$, 156 or 304, and $\mathrm{DGP2}$ was fitted as grouped vector autoregressive models with $m = 72$, 120 or 312.

Simulation Results

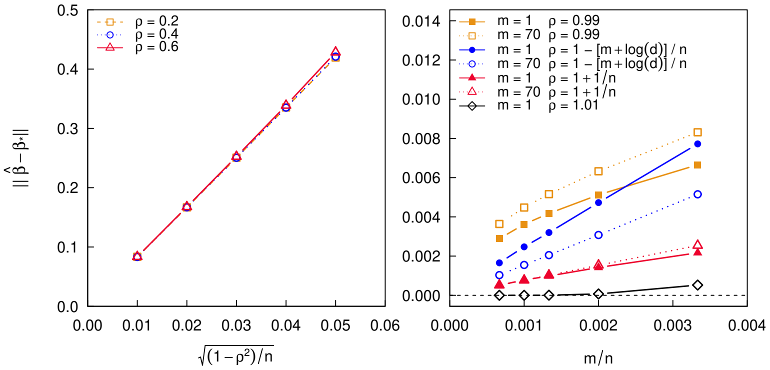


Figure 4: Error rates for $\mathrm{DGP3}$ as ρ is fixed or approaching one at different rates. Left panel: plot of $\|\widehat{\beta} - \beta_*\|$ against $\{(1-\rho^2)/n\}^{1/2}$ with $\rho = 0.2$, 0.4 or 0.6, and $m = 70$. Right panel: plot of $\|\widehat{\beta} - \beta_*\|$ against m/n with $\rho = 0.99$, $1 - (m + \log d)/n$, $1 + 1/n$ or 1.01, and $m = 1$ or 70. The case of $(m, \rho) = (70, 1.01)$ is omitted as the process becomes very explosive.

Analysis of Lower Bounds

Notations: For a fixed $\bar{\rho} > 0$, let $\Theta(\bar{\rho}) = \{\theta \in \mathbb{R}^m : \rho\{A(\theta)\} \leq \bar{\rho}\}$. so the linearly restricted subspace of β is $\mathcal{L}(\bar{\rho}) = \{R\theta : \theta \in \Theta(\bar{\rho})\}$. Denote by $\mathbb{P}_\theta^{(n)}$ the distribution of $(X_1, \ldots, X_{n+1})$ on $(\mathcal{X}^{n+1}, \mathcal{F}_{n+1})$.

Theorem 4 (Lower bounds for Gaussian VAR model)

Suppose that $\{X_t\}_{t=1}^{n+1}$ follow the VAR model $X_{t+1} = AX_t + \eta_t$ with linear restrictions defined previously. In addition, Assumptions A4(i) and (ii) hold, and $\{\eta_t\}$ are normal. Fix $\delta \in (0, 1/4)$ and $\bar{\rho} > 0$. Then, for any $\epsilon \in (0, \bar{\rho}/4]$, we have

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

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

$$n \sum_{s=0}^{n-1} \bar{\rho}^{2s} \lesssim \frac{m + \log(1/\delta)}{\epsilon^2}.$$

Minimax Rates Implied by Theorem 4

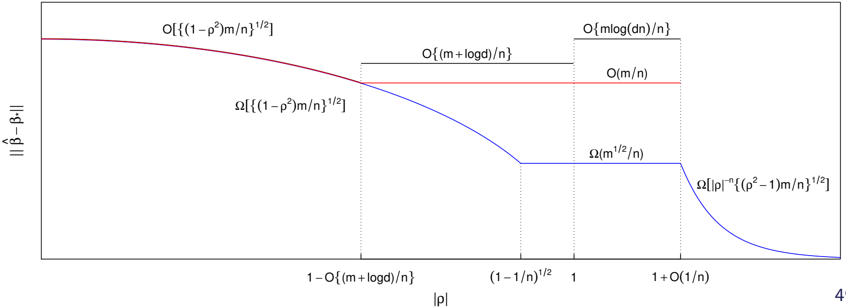
Corollary 2 (Minimax rates for Gaussian VAR model)

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

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

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

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



Conclusion and Discussion

Conclusion

- We develop a unified non-asymptotic theory for the OLS estimation of VAR models under linear restrictions, which is applicable to stable, unstable and even slightly explosive processes.
- The derived upper bounds reflect an interesting connection between asymptotic and non-asymptotic theory.
- Simulation results shed light on the sharpness of the error bounds and the actual phase transition behavior.

A "sharp" non-asymptotic analysis in high dimensions can uncover low dimensional phenomena.

Some future directions

- **Estimation with data-driven restrictions**:
  Such an estimation procedure would involve (1) suggesting possible linear restrictions based on subject knowledge and then (2) selecting the true restrictions by a data-driven approach.

- **Linear hypothesis testing**:
  Simultaneous tests for linear constraints of the VAR model

Manuscript: Yao Zheng and Guang Cheng (2019+). Finite time analysis of vector autoregressive models under linear restrictions. arXiv:1811.10197. Under revision for *Biometrika*.

Thank you!

Email: yao.zheng@uconn.edu

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

Chang, Y. (2004). Bootstrap unit root tests in panels with cross-sectional dependency. *Journal of Econometrics*, 120:263–293.

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.

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.

Zhang, B., Pan, G., and Gao, J. (2018). CLT for largest eigenvalues and unit root testing for high-dimensional nonstationary time series. *The Annals of Statistics*, 46:2186–2215.

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