Linear Convergence of Gradient Methods for Estimating Structured Transition Matrices in High-dimensional Vector Autoregressive Models

Xiao Lv [1] Wei Cui [1] Yulong Liu [2]

[1]School of Information and Electronics
Beijing Institute of Technology

[2]School of Physics
Beijing Institute of Technology

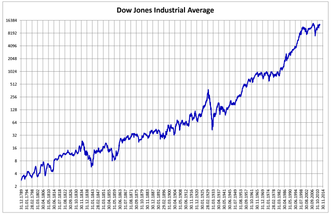
December, 2021

Outline

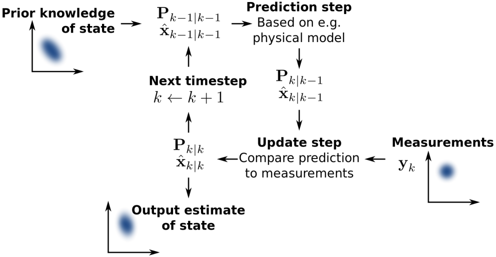
1. Motivation
   1.1. Why Time Series?
   1.2. How to Model Time Series? Vector Autoregressive Models
   1.3. Vector Autoregressive Models in High-dimensional Regime
   1.4. Related Work and Our Contribution
2. Main Results
   2.1. Single-structured Transition Matrices
   2.2. Superposition-structured Transition Matrices
3. Numerical Results
   3.1. Network Learning with a Sparse Transition Matrix
   3.2. Network Learning with a Low-rank Transition Matrix
   3.3. Network Learning with a Superposition-structured Transition Matrix
   3.4. Granger Causal Effects among Log-returns of Stocks in S&P 500 Index
   3.5. Background Modeling

Why Time Series?

- Diverse applications in forecasting, clustering, signal detecting, etc.



(a) Dow Jones Industrial Average. [1]



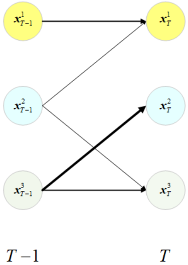
(b) Kalman filter. [2]

Figure: Applications of time series.

[1] Source: https://commons.wikimedia.org/wiki/File:Dow_Jones_Industrial_Average.png.
[2] Source: https://commons.wikimedia.org/wiki/File:Basic_concept_of_Kalman_filtering.svg.

- Capture the relationship between multiple time-varying quantities.

$$\boldsymbol{x}_{t+1} = \boldsymbol{\Gamma}_\star^T \boldsymbol{x}_t + \boldsymbol{e}_{t+1}, \quad \boldsymbol{e}_t \overset{iid}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_e), \quad t = 0, \cdots, n-1.$$



- Transform into the matrix form

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\Gamma}_\star + \boldsymbol{E},$$

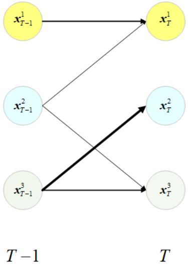
where $\boldsymbol{Y} = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n]^T$, $\boldsymbol{X} = [\boldsymbol{x}_0, \cdots, \boldsymbol{x}_{n-1}]^T$, and $\boldsymbol{E} = [\boldsymbol{e}_1, \cdots, \boldsymbol{e}_n]^T$.

- **Goal** of VAR models: estimate the transition matrix $\boldsymbol{\Gamma}_\star$.

How to Model Time Series? Vector Autoregressive Models

- Capture the relationship between multiple time-varying quantities.

$$\boldsymbol{x}_{t+1} = \boldsymbol{\Gamma}_\star^T \boldsymbol{x}_t + \boldsymbol{e}_{t+1}, \quad \boldsymbol{e}_t \overset{iid}{\sim} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_e), \quad t = 0, \cdots, n-1.$$



- Transform into the matrix form

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\Gamma}_\star + \boldsymbol{E},$$

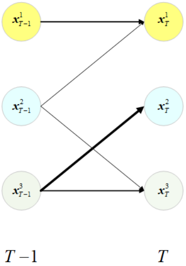
where $\boldsymbol{Y} = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n]^T$, $\boldsymbol{X} = [\boldsymbol{x}_0, \cdots, \boldsymbol{x}_{n-1}]^T$, and $\boldsymbol{E} = [\boldsymbol{e}_1, \cdots, \boldsymbol{e}_n]^T$.

- Goal of VAR models: estimate the transition matrix $\boldsymbol{\Gamma}_\star$.

How to Model Time Series? Vector Autoregressive Models

- Capture the relationship between multiple time-varying quantities.

$$\boldsymbol{x}_{t+1} = \boldsymbol{\Gamma}_\star^T \boldsymbol{x}_t + \boldsymbol{e}_{t+1}, \quad \boldsymbol{e}_t \overset{iid}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_e), \quad t = 0, \cdots, n-1.$$



- Transform into the matrix form

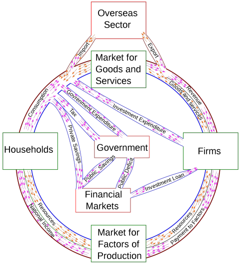
$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\Gamma}_\star + \boldsymbol{E},$$

where $\boldsymbol{Y} = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n]^T$, $\boldsymbol{X} = [\boldsymbol{x}_0, \cdots, \boldsymbol{x}_{n-1}]^T$, and $\boldsymbol{E} = [\boldsymbol{e}_1, \cdots, \boldsymbol{e}_n]^T$.

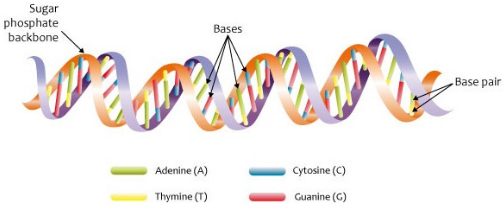
- **Goal** of VAR models: estimate the transition matrix $\boldsymbol{\Gamma}_\star$.

Vector Autoregressive Models in High-dimensional Regime

- VAR models in **high-dimensional** regime: macroeconomics, genomics, etc.



(a) Macroeconomics. [3]



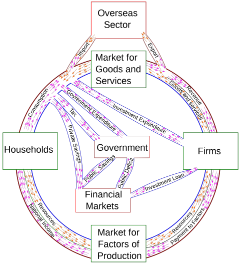
(b) Genomics. [4]

- **Challenge:** underdetermined problem.
- **Solution:** impose **structure priors**, such as sparsity, group-sparsity and low rank.

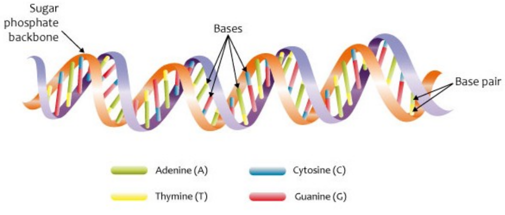
[3] Source: https://commons.wikimedia.org/wiki/File:CircularFlowEN-SVG.svg
[4] Source: https://commons.wikimedia.org/wiki/File:DNA_double_helix_(13081113544).jpg

Vector Autoregressive Models in High-dimensional Regime

- VAR models in **high-dimensional** regime: macroeconomics, genomics, etc.



(a) Macroeconomics. [3]



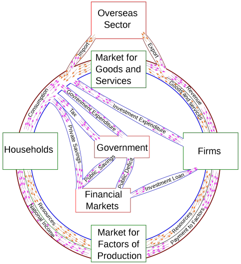
(b) Genomics. [4]

- **Challenge:** underdetermined problem.
- **Solution:** impose **structure priors**, such as sparsity, group-sparsity and low rank.

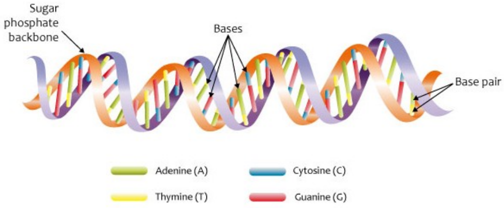
[3] Source: https://commons.wikimedia.org/wiki/File:CircularFlowEN-SVG.svg
[4] Source: https://commons.wikimedia.org/wiki/File:DNA_double_helix_(13081113544).jpg

Vector Autoregressive Models in High-dimensional Regime

- VAR models in **high-dimensional** regime: macroeconomics, genomics, etc.



(a) Macroeconomics. [3]



(b) Genomics. [4]

- **Challenge:** underdetermined problem.
- **Solution:** impose **structure priors**, such as sparsity, group-sparsity and low rank.

[3]Source: https://commons.wikimedia.org/wiki/File:CircularFlowEN-SVG.svg
[4]Source: https://commons.wikimedia.org/wiki/File:DNA_double_helix_(13081113544).jpg

Related Work and Our Contribution

- Theory in low-dimensional settings is well established in (Lütkepohl, 2005).
- Several literature about the **statistical analysis** in high-dimensional settings (Loh and Wainwright, 2012; Han and Liu, 2013; Basu and Michailidis, 2015; Melnyk and Banerjee, 2016; Basu et al., 2019).
- Other important questions:
  - Few literature from the **algorithmic view** in high-dimensional settings.
  - Single-structured assumption for parameters might be too simple in real applications.
- **Our contribution**:
  - Provide the non-asymptotic optimization guarantee for VAR models.
  - Consider both single-structured and superposition-structured transition matrices.

Related Work and Our Contribution

- Theory in low-dimensional settings is well established in (Lütkepohl, 2005).
- Several literature about the **statistical analysis** in high-dimensional settings (Loh and Wainwright, 2012; Han and Liu, 2013; Basu and Michailidis, 2015; Melnyk and Banerjee, 2016; Basu et al., 2019).
- Other important questions:
  - Few literature from the **algorithmic view** in high-dimensional settings.
  - Single-structured assumption for parameters might be too simple in real applications.
- **Our contribution**:
  - Provide the non-asymptotic optimization guarantee for VAR models.
  - Consider both single-structured and superposition-structured transition matrices.

Related Work and Our Contribution

- Theory in low-dimensional settings is well established in (Lütkepohl, 2005).
- Several literature about the **statistical analysis** in high-dimensional settings (Loh and Wainwright, 2012; Han and Liu, 2013; Basu and Michailidis, 2015; Melnyk and Banerjee, 2016; Basu et al., 2019).
- Other important questions:
  - Few literature from the **algorithmic view** in high-dimensional settings.
  - Single-structured assumption for parameters might be too simple in real applications.
- **Our contribution**:
  - Provide the non-asymptotic optimization guarantee for VAR models.
  - Consider both single-structured and superposition-structured transition matrices.

Related Work and Our Contribution

- Theory in low-dimensional settings is well established in (Lütkepohl, 2005).
- Several literature about the **statistical analysis** in high-dimensional settings (Loh and Wainwright, 2012; Han and Liu, 2013; Basu and Michailidis, 2015; Melnyk and Banerjee, 2016; Basu et al., 2019).
- Other important questions:
  - Few literature from the **algorithmic view** in high-dimensional settings.
  - Single-structured assumption for parameters might be too simple in real applications.
- **Our contribution**:
  - Provide the non-asymptotic optimization guarantee for VAR models.
  - Consider both single-structured and superposition-structured transition matrices.

Related Work and Our Contribution

- Theory in low-dimensional settings is well established in (Lütkepohl, 2005).
- Several literature about the **statistical analysis** in high-dimensional settings (Loh and Wainwright, 2012; Han and Liu, 2013; Basu and Michailidis, 2015; Melnyk and Banerjee, 2016; Basu et al., 2019).
- Other important questions:
  - Few literature from the **algorithmic view** in high-dimensional settings.
  - Single-structured assumption for parameters might be too simple in real applications.
- **Our contribution**:
  - Provide the non-asymptotic optimization guarantee for VAR models.
  - Consider both single-structured and superposition-structured transition matrices.

Related Work and Our Contribution

- Theory in low-dimensional settings is well established in (Lütkepohl, 2005).
- Several literature about the **statistical analysis** in high-dimensional settings (Loh and Wainwright, 2012; Han and Liu, 2013; Basu and Michailidis, 2015; Melnyk and Banerjee, 2016; Basu et al., 2019).
- Other important questions:
  - Few literature from the **algorithmic view** in high-dimensional settings.
  - Single-structured assumption for parameters might be too simple in real applications.
- **Our contribution**:
  - Provide the non-asymptotic optimization guarantee for VAR models.
  - Consider both single-structured and superposition-structured transition matrices.

Outline

1. Motivation
   1.1. Why Time Series?
   1.2. How to Model Time Series? Vector Autoregressive Models
   1.3. Vector Autoregressive Models in High-dimensional Regime
   1.4. Related Work and Our Contribution

2. Main Results
   2.1. Single-structured Transition Matrices
   2.2. Superposition-structured Transition Matrices

3. Numerical Results
   3.1. Network Learning with a Sparse Transition Matrix
   3.2. Network Learning with a Low-rank Transition Matrix
   3.3. Network Learning with a Superposition-structured Transition Matrix
   3.4. Granger Causal Effects among Log-returns of Stocks in S&P 500 Index
   3.5. Background Modeling

Single-structured Transition Matrices

- Promote the structure of $\mathbf{\Gamma}_\star$ by a convex regularizer $\mathcal{R}(\cdot)$.

Constrained least square problem for VAR models with single-structured transition matrices

$$\begin{aligned} \min_{\mathbf{\Gamma}} \quad & \frac{1}{2n}\|\boldsymbol{Y} - \boldsymbol{X}\mathbf{\Gamma}\|_{\mathrm{F}}^2 \\ \text{s.t.} \quad & \mathcal{R}(\mathbf{\Gamma}) \leq \mathcal{R}(\mathbf{\Gamma}_\star). \end{aligned}$$

- Optimize through projected gradient descent (PGD).

Algorithm 1 PGD for single-structured transition matrices estimation

Input: Initial point $\mathbf{\Gamma}_0$, step size μ, iteration number K.
for $k = 0$ **to** $K - 1$ **do**
 $\mathbf{\Gamma}_{k+1} = \mathcal{P}_{\mathcal{K}}(\mathbf{\Gamma}_k - \mu \nabla f_n(\mathbf{\Gamma}_k))$
end for
Output: $\mathbf{\Gamma}_K$

Here, $\mathcal{K} = \{\mathbf{\Gamma} \mid \mathcal{R}(\mathbf{\Gamma}) \leq \mathcal{R}(\mathbf{\Gamma}_\star)\}$ is the descent set.

Two Assumptions

Stability (Basu and Michailidis, 2015)

The characteristic polynomial of the VAR model satisfies $\det(\mathcal{A}(z)) \neq 0$ on the unit circle of the complex plane $\{z \in \mathbb{C} \colon |z| = 1\}$, where $\mathcal{A}(z) = \boldsymbol{I}_{d\times d} - \boldsymbol{\Gamma}_\star^T z$.

Boundness

Suppose there are positive constants $\kappa_{\min}$ and $\kappa_{\max}$ satisfying

$$0 < \frac{\kappa_{\min}}{2\pi} \leq \operatorname*{ess\,inf}_{\theta\in[-\pi,\pi]} \lambda_{\min}(f_x(\theta)) \leq \operatorname*{ess\,sup}_{\theta\in[-\pi,\pi]} \lambda_{\max}(f_x(\theta)) \leq \frac{\kappa_{\max}}{2\pi}.$$

where $f_x(\theta)$ is the spectral density function defined as (Basu and Michailidis, 2015)

$$f_x(\theta) := \frac{1}{2\pi}\sum_{l=-\infty}^{\infty} \boldsymbol{\Sigma}_x(l) e^{-il\theta}, \qquad \theta \in [-\pi,\pi].$$

Here we use $\boldsymbol{\Sigma}_x(l) = \mathbb{E}[\boldsymbol{x}_t \boldsymbol{x}_{t+l}^T], \ \ t, l \in \mathbb{Z}$.

Two Assumptions

Stability (Basu and Michailidis, 2015)

The characteristic polynomial of the VAR model satisfies $\det(\mathcal{A}(z)) \neq 0$ on the unit circle of the complex plane $\{z \in \mathbb{C} \colon |z| = 1\}$, where $\mathcal{A}(z) = \boldsymbol{I}_{d\times d} - \boldsymbol{\Gamma}_\star^T z$.

Boundness

Suppose there are positive constants $\kappa_{\min}$ and $\kappa_{\max}$ satisfying

$$0 < \frac{\kappa_{\min}}{2\pi} \leq \operatorname*{ess\,inf}_{\theta\in[-\pi,\pi]} \lambda_{\min}(f_x(\theta)) \leq \operatorname*{ess\,sup}_{\theta\in[-\pi,\pi]} \lambda_{\max}(f_x(\theta)) \leq \frac{\kappa_{\max}}{2\pi}.$$

where $f_x(\theta)$ is the spectral density function defined as (Basu and Michailidis, 2015)

$$f_x(\theta) \coloneqq \frac{1}{2\pi}\sum_{l=-\infty}^{\infty} \boldsymbol{\Sigma}_x(l) e^{-il\theta}, \qquad \theta \in [-\pi, \pi].$$

Here we use $\boldsymbol{\Sigma}_x(l) = \mathbb{E}[\boldsymbol{x}_t \boldsymbol{x}_{t+l}^T],\ \ t, l \in \mathbb{Z}$.

Theoretical Result

Linear convergence of PGD

Starting from a point $\mathbf{\Gamma}_0$ satisfying $\mathcal{R}(\mathbf{\Gamma}_0) \leq \mathcal{R}(\mathbf{\Gamma}_\star)$, we perform PGD with the step size $\mu = 1/\kappa_{\max}$. If the number of measurements satisfies

$$\sqrt{n} > 2C\frac{\kappa_{\max}}{\kappa_{\min}}(\omega(\mathcal{C} \cap \mathbb{S}_F) + u),$$

then with probability at least $1 - c\exp(-u^2)$, the PGD update would obey

$$\|\mathbf{\Gamma}_{k+1} - \mathbf{\Gamma}_\star\|_{\mathrm{F}} \leq \rho^{k+1}\|\mathbf{\Gamma}_0 - \mathbf{\Gamma}_\star\|_{\mathrm{F}} + \frac{\xi}{1-\rho}.$$

- When $\rho < 1$, PGD would enjoy a **linear** convergence rate.
- The requirement of samples is of order $\omega(\mathcal{C} \cap \mathbb{S}_F)^2$, which is **sharp** up to a constant factor.
- The **temporal dependency** could be characterized by $\kappa_{\min}$ and $\kappa_{\max}$.

Theoretical Result

Linear convergence of PGD

Starting from a point $\mathbf{\Gamma}_0$ satisfying $\mathcal{R}(\mathbf{\Gamma}_0) \leq \mathcal{R}(\mathbf{\Gamma}_\star)$, we perform PGD with the step size $\mu = 1/\kappa_{\max}$. If the number of measurements satisfies

$$\sqrt{n} > 2C\frac{\kappa_{\max}}{\kappa_{\min}}(\omega(\mathcal{C} \cap \mathbb{S}_F) + u),$$

then with probability at least $1 - c\exp(-u^2)$, the PGD update would obey

$$\|\mathbf{\Gamma}_{k+1} - \mathbf{\Gamma}_\star\|_{\mathrm{F}} \leq \rho^{k+1}\|\mathbf{\Gamma}_0 - \mathbf{\Gamma}_\star\|_{\mathrm{F}} + \frac{\xi}{1-\rho}.$$

- When $\rho < 1$, PGD would enjoy a **linear** convergence rate.
- The requirement of samples is of order $\omega(\mathcal{C} \cap \mathbb{S}_F)^2$, which is **sharp** up to a constant factor.
- The **temporal dependency** could be characterized by $\kappa_{\min}$ and $\kappa_{\max}$.

Theoretical Result

Linear convergence of PGD

Starting from a point $\boldsymbol{\Gamma}_0$ satisfying $\mathcal{R}(\boldsymbol{\Gamma}_0) \leq \mathcal{R}(\boldsymbol{\Gamma}_\star)$, we perform PGD with the step size $\mu = 1/\kappa_{\max}$. If the number of measurements satisfies

$$\sqrt{n} > 2C\frac{\kappa_{\max}}{\kappa_{\min}}(\omega(\mathcal{C} \cap \mathbb{S}_F) + u),$$

then with probability at least $1 - c\exp(-u^2)$, the PGD update would obey

$$\|\boldsymbol{\Gamma}_{k+1} - \boldsymbol{\Gamma}_\star\|_{\mathrm{F}} \leq \rho^{k+1}\|\boldsymbol{\Gamma}_0 - \boldsymbol{\Gamma}_\star\|_{\mathrm{F}} + \frac{\xi}{1-\rho}.$$

- When $\rho < 1$, PGD would enjoy a **linear** convergence rate.
- The requirement of samples is of order $\omega(\mathcal{C} \cap \mathbb{S}_F)^2$, which is **sharp** up to a constant factor.
- The **temporal dependency** could be characterized by $\kappa_{\min}$ and $\kappa_{\max}$.

Theoretical Result

Linear convergence of PGD

Starting from a point $\boldsymbol{\Gamma}_0$ satisfying $\mathcal{R}(\boldsymbol{\Gamma}_0) \leq \mathcal{R}(\boldsymbol{\Gamma}_\star)$, we perform PGD with the step size $\mu = 1/\kappa_{\max}$. If the number of measurements satisfies

$$\sqrt{n} > 2C\frac{\kappa_{\max}}{\kappa_{\min}}(\omega(\mathcal{C} \cap \mathbb{S}_F) + u),$$

then with probability at least $1 - c\exp(-u^2)$, the PGD update would obey

$$\|\boldsymbol{\Gamma}_{k+1} - \boldsymbol{\Gamma}_\star\|_{\mathrm{F}} \leq \rho^{k+1}\|\boldsymbol{\Gamma}_0 - \boldsymbol{\Gamma}_\star\|_{\mathrm{F}} + \frac{\xi}{1-\rho}.$$

- When $\rho < 1$, PGD would enjoy a **linear** convergence rate.
- The requirement of samples is of order $\omega(\mathcal{C} \cap \mathbb{S}_F)^2$, which is **sharp** up to a constant factor.
- The **temporal dependency** could be characterized by $\kappa_{\min}$ and $\kappa_{\max}$.

Reduce to Independent Samples

Apply to the case where the rows of $\boldsymbol{X}$ are generated from $\boldsymbol{x}_t \overset{iid}{\sim} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_x)$.

Linear convergence of the multi-task learning problem with independent samples

Suppose $\kappa_{\min} \le \lambda_{\min}(\boldsymbol{\Sigma}_x) \le \lambda_{\max}(\boldsymbol{\Sigma}_x) \le \kappa_{\max}$. We adopt PGD with the step size $\mu = 1/\kappa_{\max}$ and a starting point $\boldsymbol{\Gamma}_0$ satisfying $\mathcal{R}(\boldsymbol{\Gamma}_0) \le \mathcal{R}(\boldsymbol{\Gamma}_\star)$. If the number of measurements satisfies

$$\sqrt{n} > 2C\frac{\kappa_{\max}}{\kappa_{\min}}(\omega(\mathcal{C} \cap \mathbb{S}_F) + u),$$

then with probability at least $1 - c\exp(-u^2)$ the PGD update would obey

$$\|\boldsymbol{\Gamma}_{k+1} - \boldsymbol{\Gamma}_\star\|_{\mathrm{F}} \le \rho^{k+1}\|\boldsymbol{\Gamma}_0 - \boldsymbol{\Gamma}_\star\|_{\mathrm{F}} + \frac{\xi}{1-\rho}.$$

Our results provide **unified** estimation error bounds for both independent and correlated samples.

Reduce to Independent Samples

Apply to the case where the rows of $\boldsymbol{X}$ are generated from $\boldsymbol{x}_t \overset{iid}{\sim} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_x)$.

Linear convergence of the multi-task learning problem with independent samples

Suppose $\kappa_{\min} \leq \lambda_{\min}(\boldsymbol{\Sigma}_x) \leq \lambda_{\max}(\boldsymbol{\Sigma}_x) \leq \kappa_{\max}$. We adopt PGD with the step size $\mu = 1/\kappa_{\max}$ and a starting point $\boldsymbol{\Gamma}_0$ satisfying $\mathcal{R}(\boldsymbol{\Gamma}_0) \leq \mathcal{R}(\boldsymbol{\Gamma}_\star)$. If the number of measurements satisfies

$$\sqrt{n} > 2C \frac{\kappa_{\max}}{\kappa_{\min}} (\omega(\mathcal{C} \cap \mathbb{S}_F) + u),$$

then with probability at least $1 - c\exp(-u^2)$ the PGD update would obey

$$\|\boldsymbol{\Gamma}_{k+1} - \boldsymbol{\Gamma}_\star\|_{\mathrm{F}} \leq \rho^{k+1} \|\boldsymbol{\Gamma}_0 - \boldsymbol{\Gamma}_\star\|_{\mathrm{F}} + \frac{\xi}{1-\rho}.$$

Our results provide **unified** estimation error bounds for both independent and correlated samples.

Superposition-structured Transition Matrices

- Suppose $\boldsymbol{\Gamma}_\star = \boldsymbol{S}_\star + \boldsymbol{L}_\star$, whose structure is promoted by two decomposable norms $\mathcal{R}_S(\cdot)$, $\mathcal{R}_L(\cdot)$.

Constrained least square problem with superposition-structured transition matrices

$$\begin{aligned} \min_{\boldsymbol{S},\boldsymbol{L}} \quad & f_n(\boldsymbol{S},\boldsymbol{L}) = \frac{1}{2n}\|\boldsymbol{Y} - \boldsymbol{X}(\boldsymbol{S}+\boldsymbol{L})\|_{\mathrm{F}}^2 \\ \text{s.t.} \quad & \mathcal{R}_S(\boldsymbol{S}) \le \mathcal{R}_S(\boldsymbol{S}_\star) \\ & \mathcal{R}_L(\boldsymbol{L}) \le \mathcal{R}_L(\boldsymbol{L}_\star). \end{aligned}$$

- Optimize through alternating projected gradient descent (AltPGD).

Algorithm 2 AltPGD for superposition-structured transition matrices estimation

Input: Initial points $\boldsymbol{S}_0$ and $\boldsymbol{L}_0$, step size μ, iteration number K.
for $k = 0$ **to** $K-1$ **do**
 $\boldsymbol{S}_{k+1} = \mathcal{P}_{\mathcal{K}_S}(\boldsymbol{S}_k - \mu\nabla_S f_n(\boldsymbol{S}_k, \boldsymbol{L}_k))$
 $\boldsymbol{L}_{k+1} = \mathcal{P}_{\mathcal{K}_L}(\boldsymbol{L}_k - \mu\nabla_L f_n(\boldsymbol{S}_k, \boldsymbol{L}_k))$
end for
Output: $\boldsymbol{S}_K$ and $\boldsymbol{L}_K$

Two Assumptions

Decomposable norm (Negahban et al., 2012)

A regularization function $\mathcal{R}(\cdot)$ is decomposable with respect to a subspace pair $(\mathcal{M}, \overline{\mathcal{M}}^{\perp})$, if

$$\mathcal{R}(\alpha + \beta) = \mathcal{R}(\alpha) + \mathcal{R}(\beta), \quad \forall \alpha \in \mathcal{M},\ \beta \in \overline{\mathcal{M}}^{\perp}.$$

Guarantee for the separate estimation.

Structural incoherence (Yang and Ravikumar, 2013)

Given the subspace pairs $(\mathcal{M}_{\mathcal{S}}, \overline{\mathcal{M}}_{\mathcal{S}}^{\perp})$ and $(\mathcal{M}_{\mathcal{L}}, \overline{\mathcal{M}}_{\mathcal{L}}^{\perp})$ for the two parameters $\boldsymbol{S}_{\star}$, $\boldsymbol{L}_{\star}$. Suppose

$$\max\Big\{\bar{\sigma}_{\max}(\mathcal{P}_{\overline{\mathcal{M}}_S}\boldsymbol{\Sigma}_x\mathcal{P}_{\overline{\mathcal{M}}_L}), \bar{\sigma}_{\max}(\mathcal{P}_{\overline{\mathcal{M}}_S^{\perp}}\boldsymbol{\Sigma}_x\mathcal{P}_{\overline{\mathcal{M}}_L}), \bar{\sigma}_{\max}(\mathcal{P}_{\overline{\mathcal{M}}_S}\boldsymbol{\Sigma}_x\mathcal{P}_{\overline{\mathcal{M}}_L^{\perp}}), \bar{\sigma}_{\max}(\mathcal{P}_{\overline{\mathcal{M}}_S^{\perp}}\boldsymbol{\Sigma}_x\mathcal{P}_{\overline{\mathcal{M}}_L^{\perp}})\Big\} \leq \frac{\kappa_{\min}}{8},$$

where $\boldsymbol{\Sigma}_x = \boldsymbol{\Sigma}_x(0)$ and $\bar{\sigma}_{\max}(\boldsymbol{\Sigma}) = \sup_{\boldsymbol{V}, \boldsymbol{U} \in \mathbb{S}_F} \langle \boldsymbol{V}, \boldsymbol{\Sigma}\boldsymbol{U} \rangle$.

Two Assumptions

Decomposable norm (Negahban et al., 2012)

A regularization function $\mathcal{R}(\cdot)$ is decomposable with respect to a subspace pair $(\mathcal{M}, \overline{\mathcal{M}}^{\perp})$, if

$$\mathcal{R}(\alpha + \beta) = \mathcal{R}(\alpha) + \mathcal{R}(\beta), \quad \forall \alpha \in \mathcal{M},\ \beta \in \overline{\mathcal{M}}^{\perp}.$$

Guarantee for the separate estimation.

Structural incoherence (Yang and Ravikumar, 2013)

Given the subspace pairs $(\mathcal{M}_{\mathcal{S}}, \overline{\mathcal{M}}_{\mathcal{S}}^{\perp})$ and $(\mathcal{M}_{\mathcal{L}}, \overline{\mathcal{M}}_{\mathcal{L}}^{\perp})$ for the two parameters $\boldsymbol{S}_\star$, $\boldsymbol{L}_\star$. Suppose

$$\max\Big\{\bar{\sigma}_{\max}(\mathcal{P}_{\overline{\mathcal{M}}_S}\boldsymbol{\Sigma}_x\mathcal{P}_{\overline{\mathcal{M}}_L}), \bar{\sigma}_{\max}(\mathcal{P}_{\overline{\mathcal{M}}_S^{\perp}}\boldsymbol{\Sigma}_x\mathcal{P}_{\overline{\mathcal{M}}_L}),$$
$$\bar{\sigma}_{\max}(\mathcal{P}_{\overline{\mathcal{M}}_S}\boldsymbol{\Sigma}_x\mathcal{P}_{\overline{\mathcal{M}}_L^{\perp}}), \bar{\sigma}_{\max}(\mathcal{P}_{\overline{\mathcal{M}}_S^{\perp}}\boldsymbol{\Sigma}_x\mathcal{P}_{\overline{\mathcal{M}}_L^{\perp}})\Big\} \le \frac{\kappa_{\min}}{8},$$

where $\boldsymbol{\Sigma}_x = \boldsymbol{\Sigma}_x(0)$ and $\bar{\sigma}_{\max}(\boldsymbol{\Sigma}) = \sup_{\boldsymbol{V},\boldsymbol{U}\in\mathbb{S}_F} \langle \boldsymbol{V}, \boldsymbol{\Sigma}\boldsymbol{U}\rangle$.

Theoretical Result

Linear convergence of AltPGD

Suppose $\mathbf{\Gamma}_\star$ is superposition-structured and $\mathbf{\Gamma}_\star = \boldsymbol{S}_\star + \boldsymbol{L}_\star$. Starting from points $\boldsymbol{S}_0$ and $\boldsymbol{L}_0$ satisfying $\mathcal{R}_S(\boldsymbol{S}_0) \leq \mathcal{R}_S(\boldsymbol{S}_\star)$ and $\mathcal{R}_L(\boldsymbol{L}_0) \leq \mathcal{R}_L(\boldsymbol{L}_\star)$, we adopt AltPGD with the step size $\mu = 1/\kappa_{\max}$. If the number of measurements satisfies

$$\sqrt{n} > 4C\frac{\kappa_{\max}}{\kappa_{\min}}(\omega(\mathcal{C}_S \cap \mathbb{S}_F) + \omega(\mathcal{C}_L \cap \mathbb{S}_F) + u),$$

then with probability at least $1 - c\exp(-u^2)$ the update would obey

$$\|\boldsymbol{S}_{k+1} - \boldsymbol{S}_\star\|_{\mathrm{F}} + \|\boldsymbol{L}_{k+1} - \boldsymbol{L}_\star\|_{\mathrm{F}} \leq \rho^{k+1}(\|\boldsymbol{S}_0 - \boldsymbol{S}_\star\|_{\mathrm{F}} + \|\boldsymbol{L}_0 - \boldsymbol{L}_\star\|_{\mathrm{F}}) + \frac{\xi}{1-\rho}.$$

- When $\rho < 1$, AltPGD would enjoy a **linear** convergence rate.
- The estimation error converges to **zero**, when the number of samples approaches infinity.

Reduce to Robust PCA

Write the sample matrix from n i.i.d. sample $\boldsymbol{z}_i = \boldsymbol{u}_i + \boldsymbol{v}_i$, $\boldsymbol{u}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{L}_\star)$ and $\boldsymbol{v}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{S}_\star)$

$$\boldsymbol{Y} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{z}_i \boldsymbol{z}_i^T = \boldsymbol{L}_\star + \boldsymbol{S}_\star + \boldsymbol{E},$$

where $\boldsymbol{E} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{z}_i \boldsymbol{z}_i^T - (\boldsymbol{L}_\star + \boldsymbol{S}_\star)$ is a Wishart noise matrix.

Constrained problem for robust PCA

$$\begin{aligned} \min_{\boldsymbol{S},\boldsymbol{L}} \quad & \frac{1}{2}\|\boldsymbol{Y} - \boldsymbol{S} - \boldsymbol{L}\|_{\mathrm{F}}^2 \\ \text{s.t.} \quad & \|\mathrm{vec}(\boldsymbol{S}^T)\|_1 \le \|\mathrm{vec}(\boldsymbol{S}_\star^T)\|_1, \quad \|\boldsymbol{L}\|_\star \le \|\boldsymbol{L}_\star\|_\star. \end{aligned}$$

Reduce to Robust PCA

Write the sample matrix from n i.i.d. sample $\boldsymbol{z}_i = \boldsymbol{u}_i + \boldsymbol{v}_i$, $\boldsymbol{u}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{L}_\star)$ and $\boldsymbol{v}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{S}_\star)$

$$\boldsymbol{Y} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{z}_i \boldsymbol{z}_i^T = \boldsymbol{L}_\star + \boldsymbol{S}_\star + \boldsymbol{E},$$

where $\boldsymbol{E} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{z}_i \boldsymbol{z}_i^T - (\boldsymbol{L}_\star + \boldsymbol{S}_\star)$ is a Wishart noise matrix.

Constrained problem for robust PCA

$$\begin{aligned} \min_{\boldsymbol{S},\boldsymbol{L}} \quad & \frac{1}{2}\|\boldsymbol{Y} - \boldsymbol{S} - \boldsymbol{L}\|_{\mathrm{F}}^2 \\ \text{s.t.} \quad & \|\mathrm{vec}(\boldsymbol{S}^T)\|_1 \le \|\mathrm{vec}(\boldsymbol{S}_\star^T)\|_1, \quad \|\boldsymbol{L}\|_\star \le \|\boldsymbol{L}_\star\|_\star. \end{aligned}$$

Theoretical Results

Linear convergence of AltPGD for robust PCA

Consider the robust PCA model where $\boldsymbol{S}_\star$ is a sparse matrix with $s_\star$ non-zero entries and $\boldsymbol{L}_\star$ is a $r_\star$-rank matrix. We adopt AltPGD with the step size $\mu = 1$ and starting points $\boldsymbol{S}_0$ and $\boldsymbol{L}_0$ satisfying $\|\text{vec}(\boldsymbol{S}_0^T)\|_1 \le \|\text{vec}(\boldsymbol{S}_\star^T)\|_1$ and $\|\boldsymbol{L}_0\|_\star \le \|\boldsymbol{L}_\star\|_\star$. If the number of measurements satisfies

$$\sqrt{n} > C'(\sqrt{s_\star \log d} + \sqrt{r_\star d} + u),$$

then the update would obey

$$\begin{aligned}
&\|\boldsymbol{S}_{k+1} - \boldsymbol{S}_\star\|_{\mathrm{F}} + \|\boldsymbol{L}_{k+1} - \boldsymbol{L}_\star\|_{\mathrm{F}} \\
&\le (\frac{1}{4})^{k+1}(\|\boldsymbol{S}_0 - \boldsymbol{S}_\star\|_{\mathrm{F}} + \|\boldsymbol{L}_0 - \boldsymbol{L}_\star\|_{\mathrm{F}}) + \frac{4}{3}C\|\boldsymbol{S}_\star + \boldsymbol{L}_\star\|\frac{\sqrt{s_\star \log d} + \sqrt{r_\star d} + u}{\sqrt{n}},
\end{aligned}$$

with probability at least $1 - c\exp(-u^2)$.

Outline

1. Motivation
   1.1. Why Time Series?
   1.2. How to Model Time Series? Vector Autoregressive Models
   1.3. Vector Autoregressive Models in High-dimensional Regime
   1.4. Related Work and Our Contribution

2. Main Results
   2.1. Single-structured Transition Matrices
   2.2. Superposition-structured Transition Matrices

3. Numerical Results
   3.1. Network Learning with a Sparse Transition Matrix
   3.2. Network Learning with a Low-rank Transition Matrix
   3.3. Network Learning with a Superposition-structured Transition Matrix
   3.4. Granger Causal Effects among Log-returns of Stocks in S&P 500 Index
   3.5. Background Modeling

Network Learning with a Sparse Transition Matrix

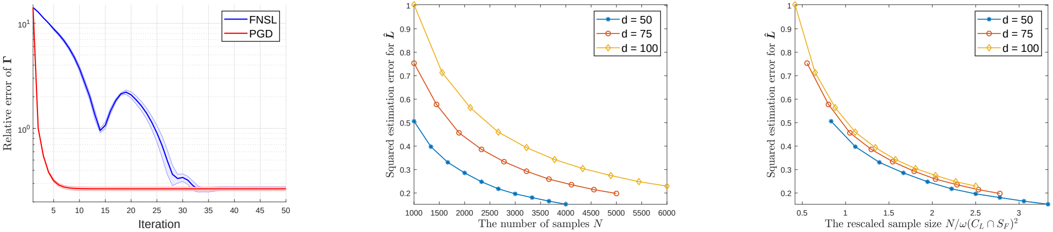
Use the true positive rate (TPR) and false alarm rate (FAR) as performance metrics.

$$\text{TPR} := \frac{\sharp\{\hat{\gamma}_{ij} \neq 0 \text{ and } \gamma^\star_{ij} \neq 0\}}{\sharp\{\gamma^\star_{ij} \neq 0\}}, \qquad \text{FAR} := \frac{\sharp\{\hat{\gamma}_{ij} \neq 0 \text{ and } \gamma^\star_{ij} = 0\}}{\sharp\{\gamma^\star_{ij} = 0\}}.$$

Table: Performance comparison between PGD and FNSL (Basu et al., 2019) on sparse network learning problems

$d = 100$	Method	TPR (%)	FAR (%)	EE	Total time (s)
$n = 1000$	PGD	**79.49**	**11.04**	**0.476**	**3.18**
	FNSL	73.64	14.19	0.489	75.59
$n = 1500$	PGD	**83.45**	**8.91**	**0.396**	**5.16**
	FNSL	78.43	11.62	0.417	140.16
$n = 2000$	PGD	**85.82**	**7.63**	**0.350**	**6.14**
	FNSL	81.30	10.07	0.373	183.79

Network Learning with a Low-rank Transition Matrix



(a) Convergence rate | (b) Squared estimation error | (c) Rescaled sample size

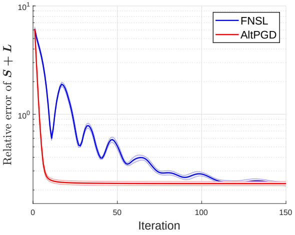
Figure: Convergence results of PGD for low-rank transition matrices estimation.

Network Learning with a Superposition-structured Transition Matrix

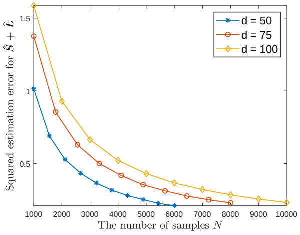
Table: Performance comparison between AltPGD and FNSL on estimation of sparse plus low-rank transition matrices

$d = 100$	Method	TPR (%)	FAR (%)	EE	Total time (s)
$n = 1500$	AltPGD	**78.26**	**11.70**	**0.475**	**19.16**
	FNSL	71.18	15.52	0.486	309.76
$n = 2000$	AltPGD	**81.06**	**10.20**	**0.421**	**26.05**
	FNSL	74.65	13.65	0.438	436.46
$n = 2500$	AltPGD	**83.19**	**9.05**	**0.379**	**32.27**
	FNSL	77.49	12.12	0.399	544.08

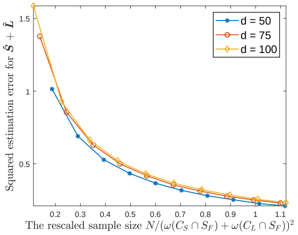
Network Learning with a Superposition-structured Transition Matrix



(a) Convergence rate



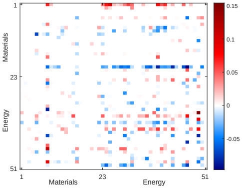
(b) Squared estimation error



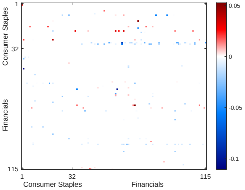
(c) Rescaled sample size

Figure: Convergence results of AltPGD for sparse plus low-rank transition matrices estimation.

Granger Causal Effects among Log-returns of Stocks in S&P 500 Index

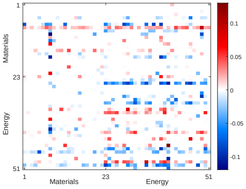


(a) Materials sector and energy sector

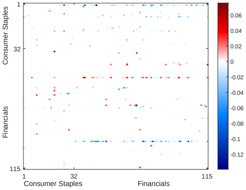


(b) Consumer staples sector and financials sector

Figure: Sparsity patterns of the transition matrix $\hat{\mathbf{\Gamma}}$ estimated by PGD.



(a) Materials sector and energy sector



(b) Consumer staples sector and financials sector

Figure: Sparsity patterns of the transition matrix $\hat{\mathbf{\Gamma}}$ estimated by FNSL.

Background Modeling

Reconstruct the static background through a sequence of video frames with moving objects in the foreground (Sobral et al., 2015).

(a) Original input frame

(b) Low-rank frame

Figure: Background modeling in the *Highway* video.

References I

Basu, S., Li, X., and Michailidis, G. (2019). Low rank and structured modeling of high-dimensional vector autoregressions. *IEEE Transactions on Signal Processing*, 67(5):1207–1222.

Basu, S. and Michailidis, G. (2015). Regularized estimation in sparse high-dimensional time series models. *The Annals of Statistics*, 43(4):1535–1567.

Han, F. and Liu, H. (2013). Transition matrix estimation in high dimensional time series. In Dasgupta, S. and McAllester, D., editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 172–180, Atlanta, Georgia, USA. PMLR.

Loh, P.-L. and Wainwright, M. J. (2012). High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *The Annals of Statistics*, 40(3):1637–1664.

Lütkepohl, H. (2005). *New Introduction to Multiple Time Series Analysis*. Springer Berlin Heidelberg.

References II

Melnyk, I. and Banerjee, A. (2016). Estimating structured vector autoregressive models. In Balcan, M. F. and Weinberger, K. Q., editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 830–839, New York, New York, USA. PMLR.

Negahban, S. N., Ravikumar, P., Wainwright, M. J., Yu, B., et al. (2012). A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557.

Sobral, A., Bouwmans, T., and Zahzah, E.-h. (2015). Lrslibrary: Low-rank and sparse tools for background modeling and subtraction in videos. In *Robust Low-Rank and Sparse Matrix Decomposition: Applications in Image and Video Processing*. CRC Press, Taylor and Francis Group.

Yang, E. and Ravikumar, P. K. (2013). Dirty statistical models. In Burges, C. J. C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Thank you!