JOINT ESTIMATION OF MULTIPLE GRANGER GRAPHICAL MODELS USING NON-CONVEX PENALTIES

Thesis Proposal

Parinthorn Manomaisaowapak

Advisor: Assist. Prof. Dr. Jitkomut Songsiri

Department of Electrical Engineering, Faculty of Engineering

Chulalongkorn University

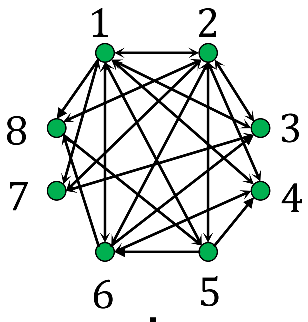
OUTLINE

- Introduction
- Overview
- Related works
- Work plan
- Background
- Methodology
- Preliminary results
- Future works

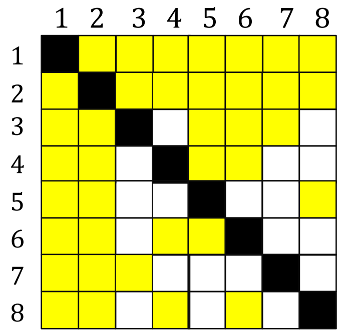
INTRODUCTION

Graphical representation

Causality network



Causality matrix



How to study relationship of variables?

Causality analysis

A strength of evidence

Granger graphical model

Granger causality(GC)

- Based on dynamical models
- Directly related to sparsity of model coefficient

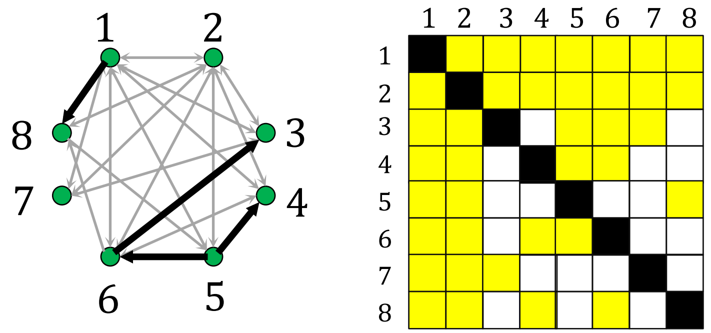
Graphical representation

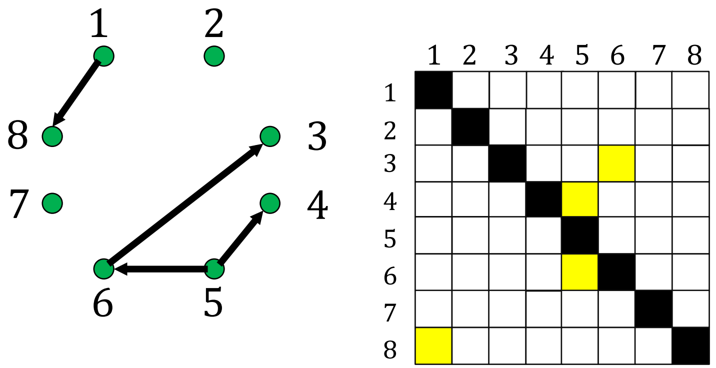
High dimensional GC network

- GC network has **large amount of connections**
- We aim to extract only **significant connections**

Sparse representation of GC network

Causality network | Causality matrix





INTRODUCTION

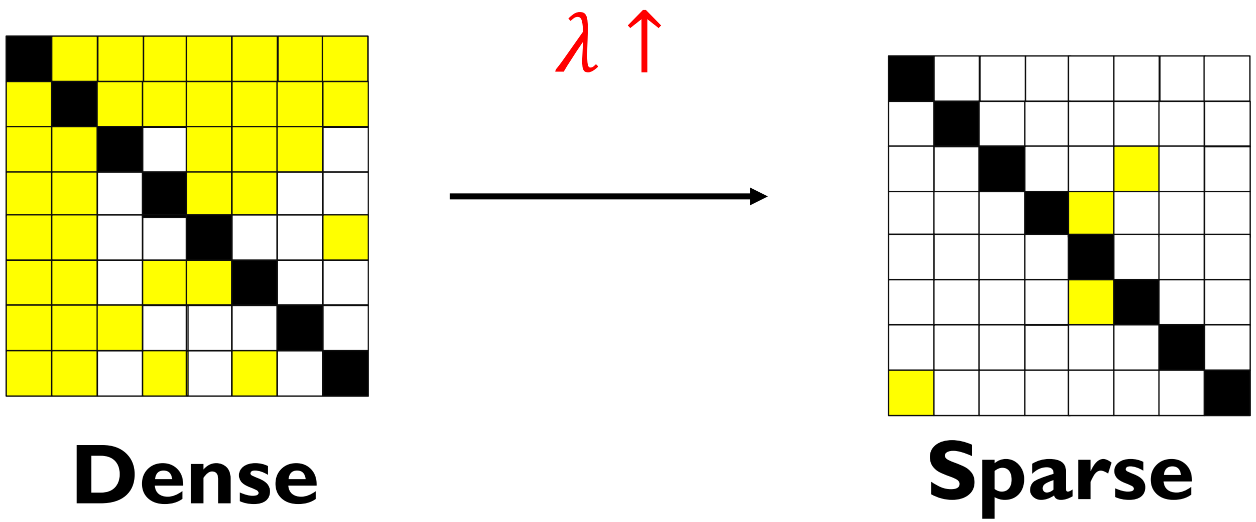
Sparse estimation formulation in general form.

$$\min_{\theta} f(\theta) + \lambda g(\theta)$$

Model parameter

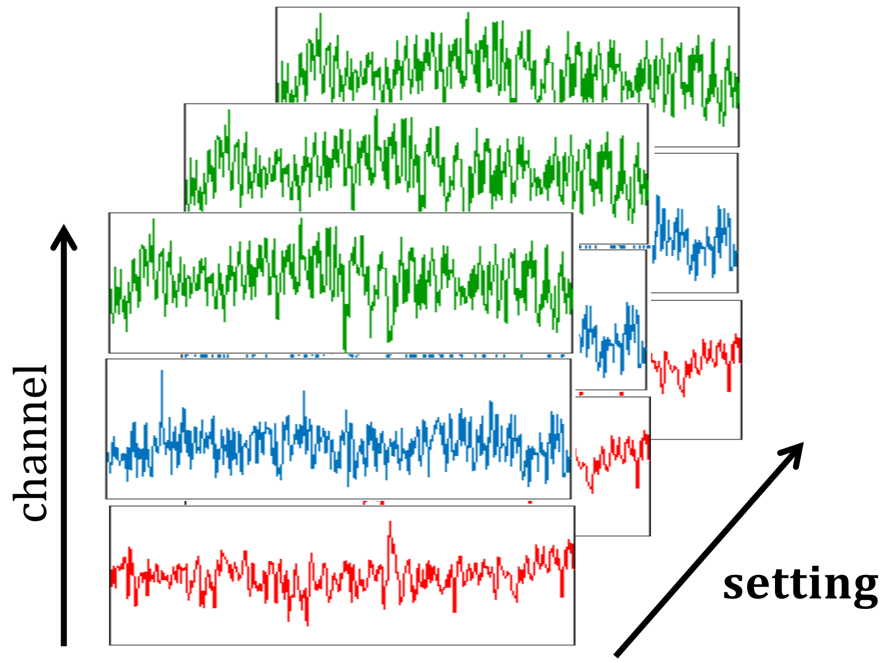
Fitting term

Sparsity inducing penalty



INTRODUCTION

consider when the same multivariate time-series are measured in different settings



analyze each **setting** separately → the relations will show if the samples are large enough

jointly analyze **all setting** → allow low sample size estimation By adding **prior information**

Joint estimation of multiple models

INTRODUCTION

Goal: Find important connections of multiple networks with prior knowledge

$$\min_{\theta_1,\ldots,\theta_K} \sum_i [f_i(\theta_i) + \boldsymbol{\lambda_1} h_i(\theta_i)] + \boldsymbol{\lambda_2} g(\theta_1, \ldots, \theta_K)$$

where h_i aims to promote differential sparsity in each model.

g aims to promote common sparsity across all models.

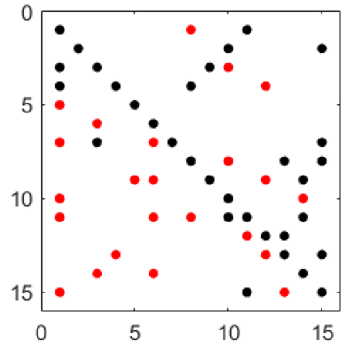
Require **definition of similarity**

Sparsity inducing function

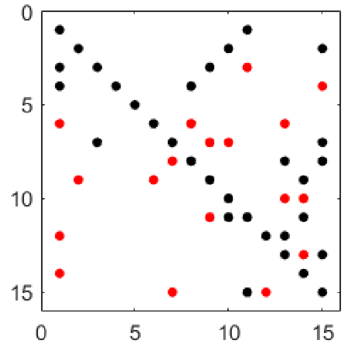
Example

group lasso → $\theta_1, \ldots, \theta_K$ has **same** non-zero **pattern**

fused lasso → $\theta_m - \theta_\ell$ is sparse → **Some** model **coefficients are identical**

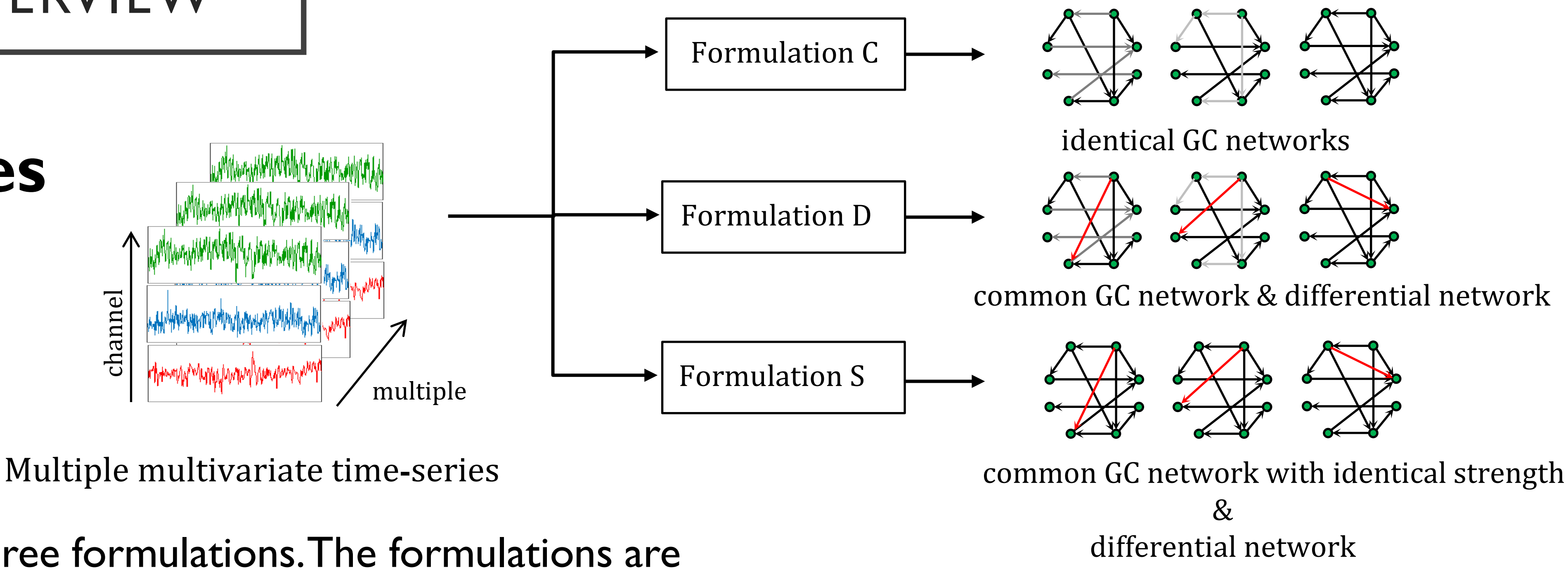


Model 1



Model 2

Objectives



- To propose three formulations. The formulations are
  - Formulation C: The estimated networks have an **identical sparsity pattern**
  - Formulation D: The estimated networks have some common parts and some different parts.
  - Formulation S: The estimated networks have some common parts and some different parts. The common parts also **share model parameters**.

- To provide efficient numerical methods for solving the proposed estimation methods in a large-scale setting.

Scope of work

- The proposed framework will be verified intensively in a simulated data sets and one real-world data set
- The usefulness of the methods will be illustrated on brain network application

Expected outcome

- Estimation formulations of multiple Granger graphical models
- A computer program that has input as a set of multivariate time-series and return group and individual Granger graphical model of the multiple time-series

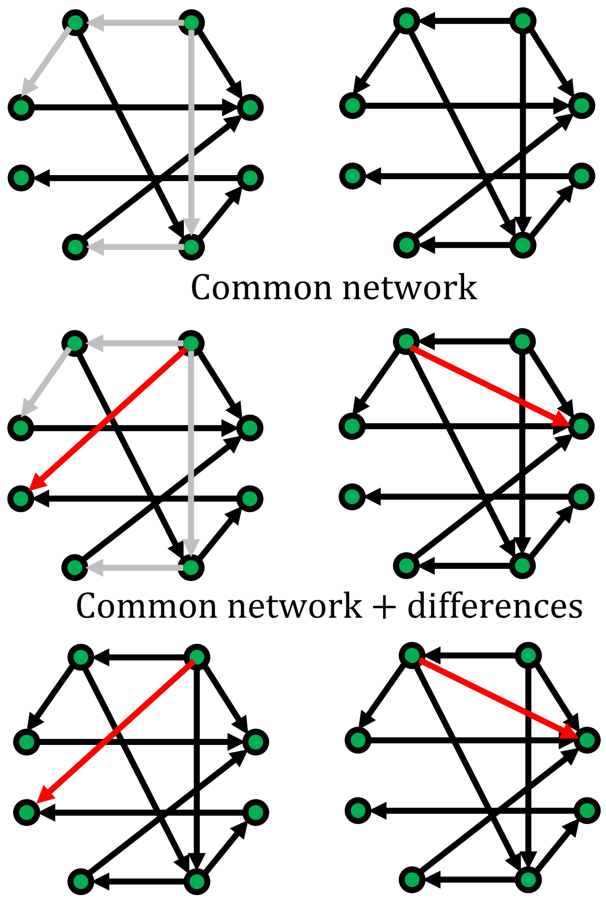
RELATED WORKS

non-convex group penalties [Our work]

C

D

S



Common network

Common network + differences

Common network has identical strength + differences

[Songsiri, 2017] group lasso, fused lasso

extension

[Wilms, 18] group lasso

[Gregorova, 2015] group lasso+Tikhonov

[Skripnikov, 2019] group lasso, two-stages

[Guo, 2011], [Chun, 2015] **non-convex penalty** Gaussian graphical model

[Skripnikov, 2019] sparse fused-lasso

[Tuck, 2020] sparse fused-lasso Gaussian graphical model

[Bore, 2020] **non-convex penalty** single Granger model

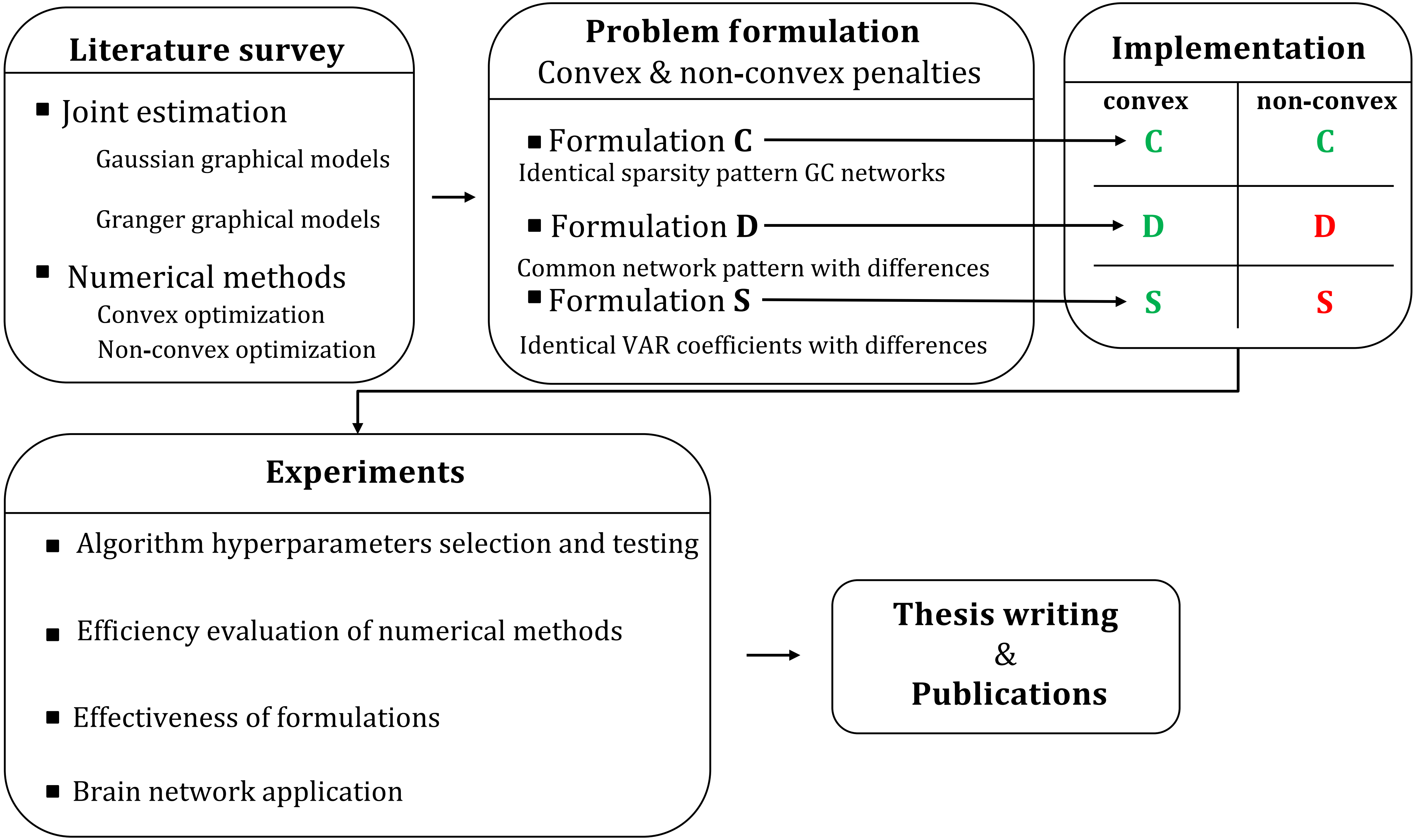
non-convex group norm penalty
$\sum \|x_{G_i}\|_p^q \quad p \geq 1, 0 < q < 1$

[Hu, 2017]

non-convex penalty
$\sum |x_i|^q \quad 0 < q < 1$

[Chartrand, 2008]

WORK PLAN

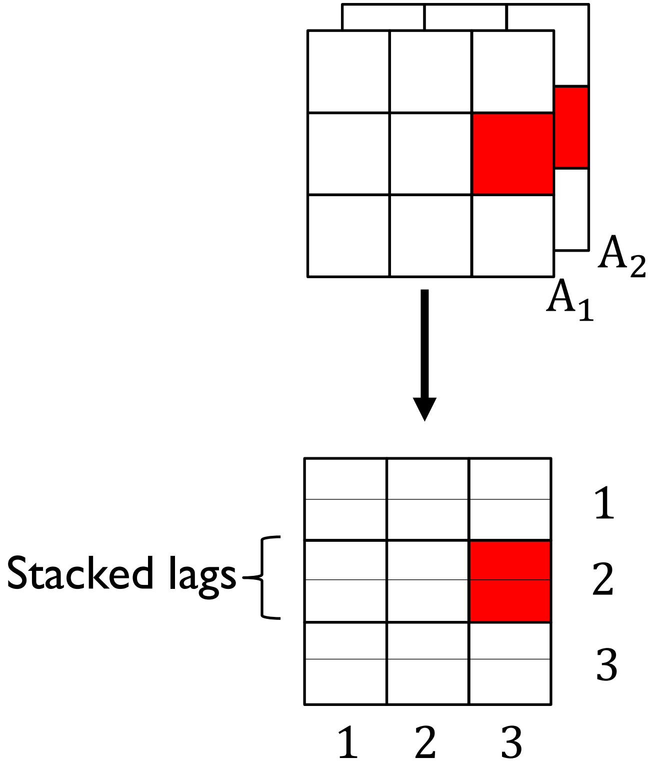

Literature survey
Joint estimation
Gaussian graphical models
Granger graphical models
Numerical methods
Convex optimization
Non-convex optimization
Problem formulation
Convex & non-convex penalties
Formulation C
Identical sparsity pattern GC networks
Formulation D
Common network pattern with differences
Formulation S
Identical VAR coefficients with differences
Implementation
convex
non-convex
C
C
D
D
S
S
Experiments
Algorithm hyperparameters selection and testing
Efficiency evaluation of numerical methods
Effectiveness of formulations
Brain network application
Thesis writing
&
Publications

BACKGROUND

Vector autoregressive model (VAR)

$$y(t) = \sum_{r=1}^{p} A_r y(t-r) + \eta(t)$$

$$A_r \in \mathbf{R}^{\boldsymbol{n}\times\boldsymbol{n}} \qquad y = (y_1, \dots, y_n) \in \mathbf{R}^{\boldsymbol{n}}$$



Granger causality on VAR models

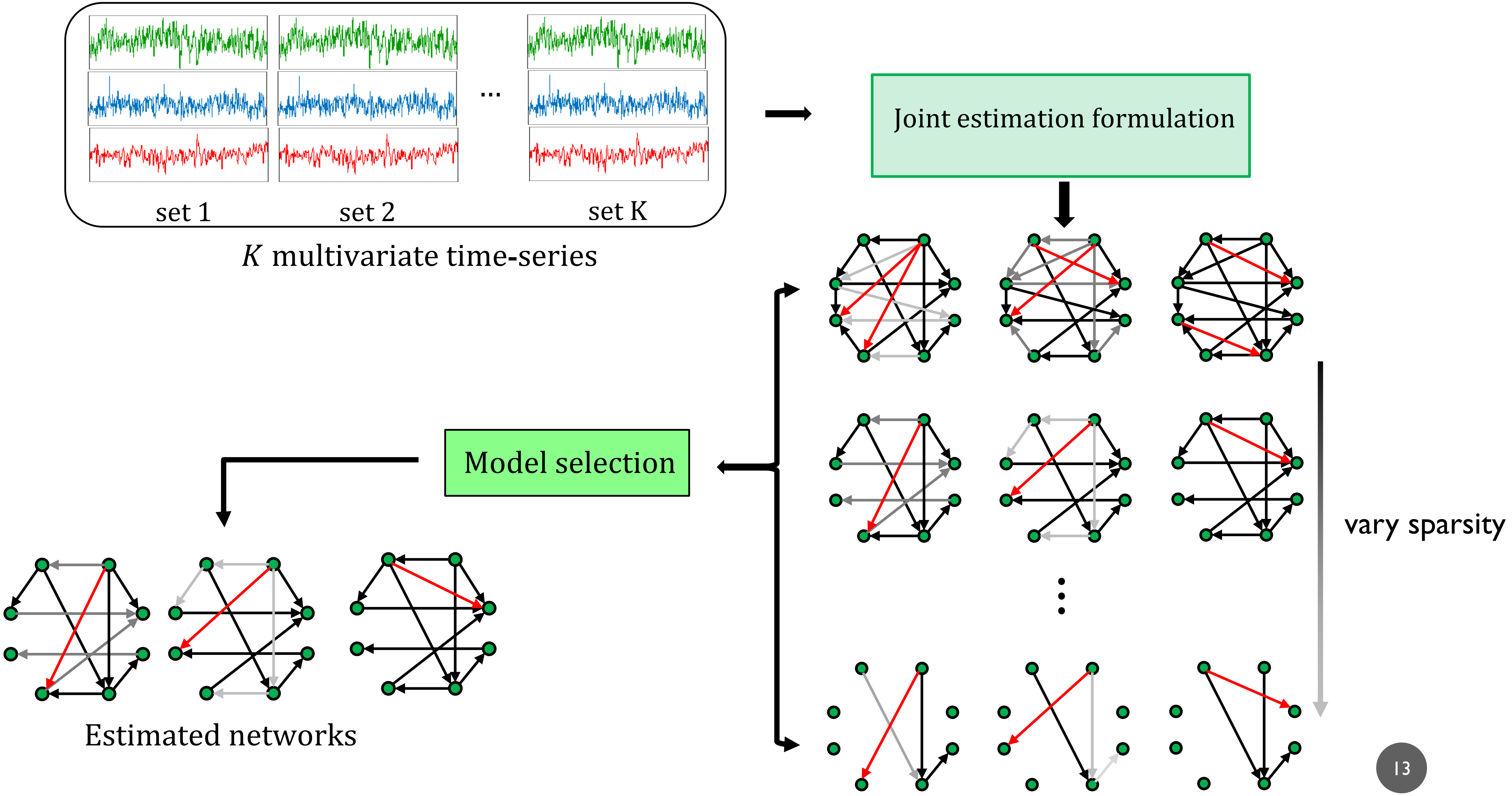
- Granger causality(GC, F_{ij}) is a strength of evidence
- Absence of GC connection can be investigated by the relation

$$F_{ij} = 0 \Leftrightarrow (A_r)_{ij} = 0; r = 1,2, \dots p$$ [Granger, 1980]

↓

Sparsity inducing penalty can be designed using this prior knowledge

METHODOLOGY OVERVIEW



We proposed three formulations,

Formulation C — **Common**

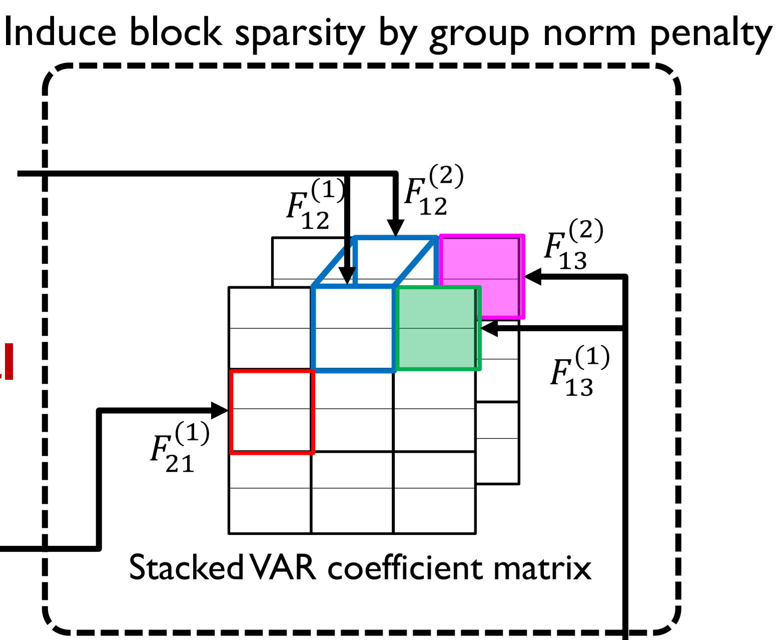
- **<u>Common</u>** pattern.

Formulation D — **Differential**

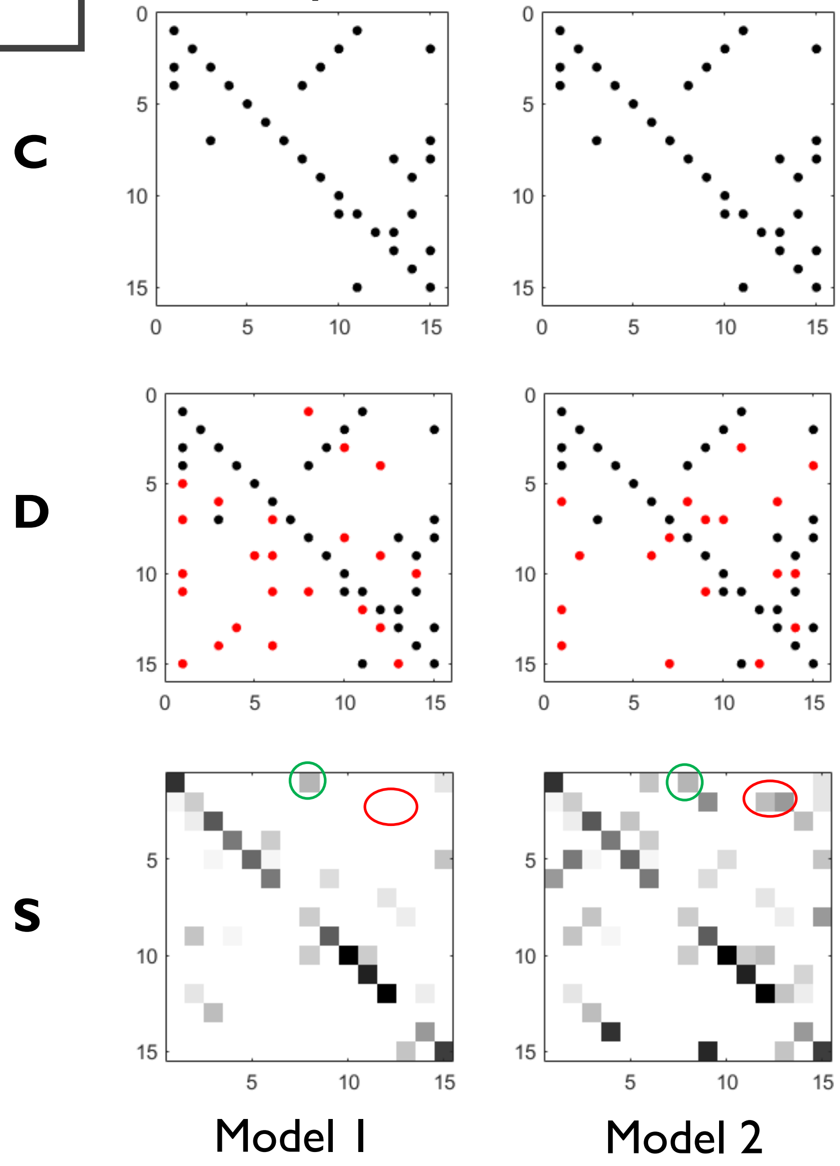
- **<u>Common</u>** pattern
- **<u>Different</u>** pattern

Formulation S — **Similar**

- **<u>Shared VAR coefficients</u>** value in all models
- **<u>Different</u>** pattern



Expected outcome



We used BIC criteria to find optimal tuning-parameters

$$BIC(\lambda_1, \lambda_2) = -2\,\mathcal{L}(\lambda_1, \lambda_2) + \log(\mathrm{N}) \cdot \mathrm{df}(\lambda_1, \lambda_2)$$

Log-likelihood of VAR model. (Fitness of models)

Effective degree of freedom (Complexity of models).

→ # off-diagonal nonzero estimated parameters

There are other choices.

Problem properties (Formulation C, D, S)

- The problem is in the form of $\min_{x} f(x) + g(x)$
- ∇f is Lipschitz-continuous.
- Function g is not differentiable at zero while we prefer sparse solutions
- We aim to solve high-dimensional problem or in a large-scale setting.

↓

First order algorithm should be considered first

↓

Proximal gradient methods unify the framework that solve this problem

- ISTA
- FISTA
- Iterative hard-thresholding algorithm
- Half-thresholding algorithm
- GIST
- GISA

Proximal algorithms

- require evaluation of **proximal operator**

  Definition: proximal operator of function g

$$\text{prox}_{\alpha g}(v) = \underset{x}{\text{argmin}}\ \ g(x) + \frac{1}{2\alpha}\|x - v\|_2^2$$

- are widely used in sparse estimation using lasso, group lasso for a convex case

- proximal operator has a closed-form expression for some functions, such as

ℓ_1 norm $\quad (\text{prox}_{\lambda\|x\|_1}(v))_i = \text{sign}(v_i)\max\{0, v_i - \lambda\}$ Soft thresholding operator

ℓ_2 norm $\quad \text{prox}_{\lambda\|x\|_2}(v) = \max\left\{0, 1 - \frac{\lambda}{\|v\|_2}\right\} v$ Block-soft thresholding operator

METHODOLOGY | ALGORITHM

proximal gradient algorithm

Solve $\min_x F(x) = f(x) + g(x)$

↓

$x^+ = \text{prox}_{\alpha g}(x - \alpha \nabla f(x))$

↓

Have a sufficient descent property in F

↓

Globally converge for our formulation
in non-convex setting
by virtue from its sufficient descent property

But it is slow

its special case is FISTA

accelerated proximal gradient(APG)

- $y = x^- + \text{correction term}$
- $x^+ = \text{prox}_{\alpha g}(y - \alpha \nabla f(y))$
- $\text{update correction term}$

not a sufficient descent method

↓

no convergence guaranteed for our formulation

↓

add a monitoring scheme [Li, 2015]

↓

monotone APG | **non-monotone APG**

↓

Globally converges in our formulations

proximal gradient methods

solve $\min_x F(x) = f(x) + g(x)$

↓

require proximal operator of
$g(x) = h_1(L_1 x) + h_2(L_2 x)$

↓

only $h_1(x), h_2(x)$
have closed-form proximal operator
[Hu, 2017]

→

Alternating direction methods of multipliers(ADMM)

solve $\min_{x,z} F(x) = f(x) + g(z) + \rho\|Ax + Bz - c\|_2^2$
subjected to $Ax+Bz=c$

↓

set $A = \begin{bmatrix} L_1 \\ L_2 \end{bmatrix}, B = -I, c = 0$

↓

Evaluation of $\mathrm{prox}_{\alpha g}(v)$
Reduce to $\mathrm{prox}_{\alpha h_1}(v), \mathrm{prox}_{\alpha h_2}(v)$

↓

no convergence guaranteed in non-convex formulations

↓

a convergence to critical point can be obtained
by **selecting a proper penalty parameter ρ**

adaptive ρ may solve the problem → **Spectral ADMM [Xu, 2017]**

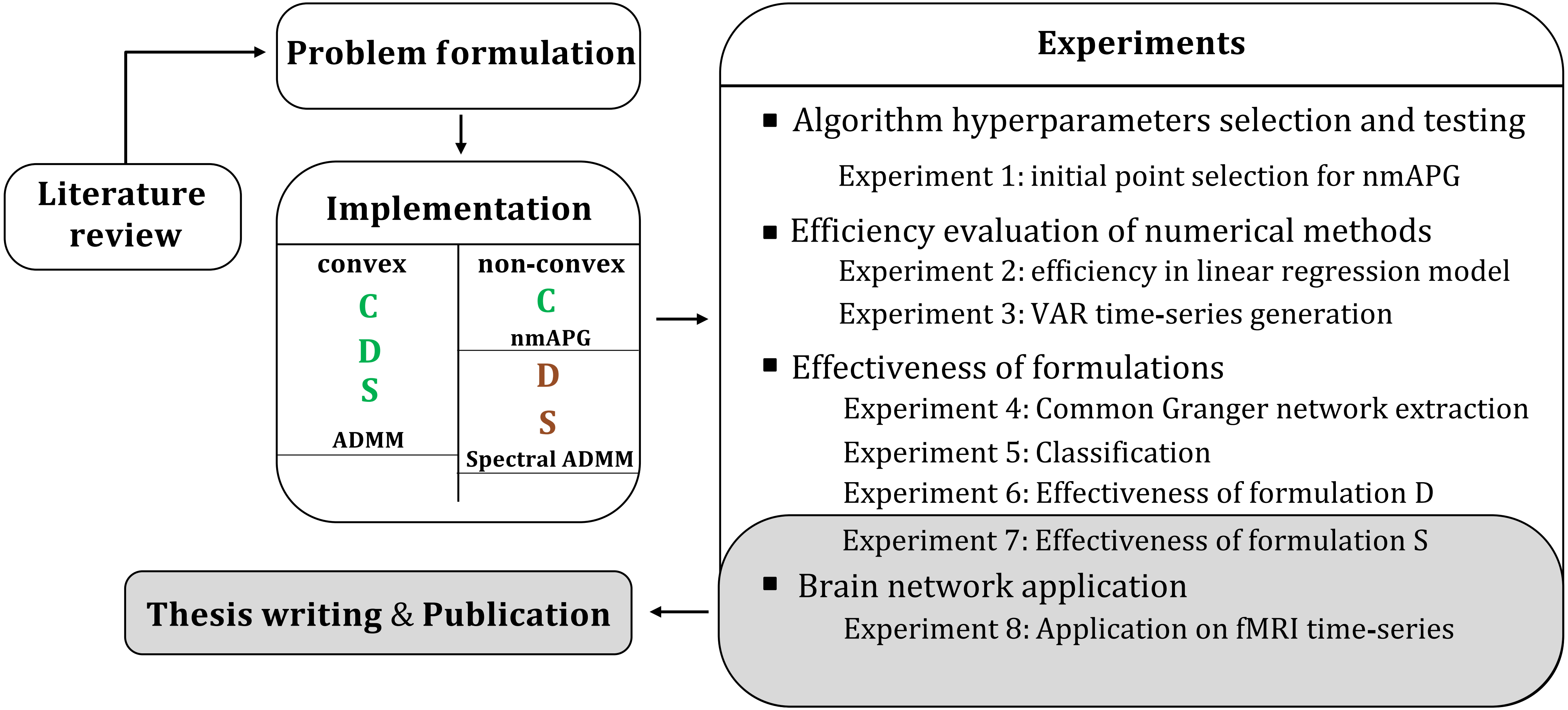
related work: **SDMM** [Combettes, 2011]

set $A = \begin{bmatrix} I \\ L_1 \\ L_2 \end{bmatrix}, B = -I, c = 0$

↓

Higher
computation complexity
than ours

PRELIMINARY RESULTS

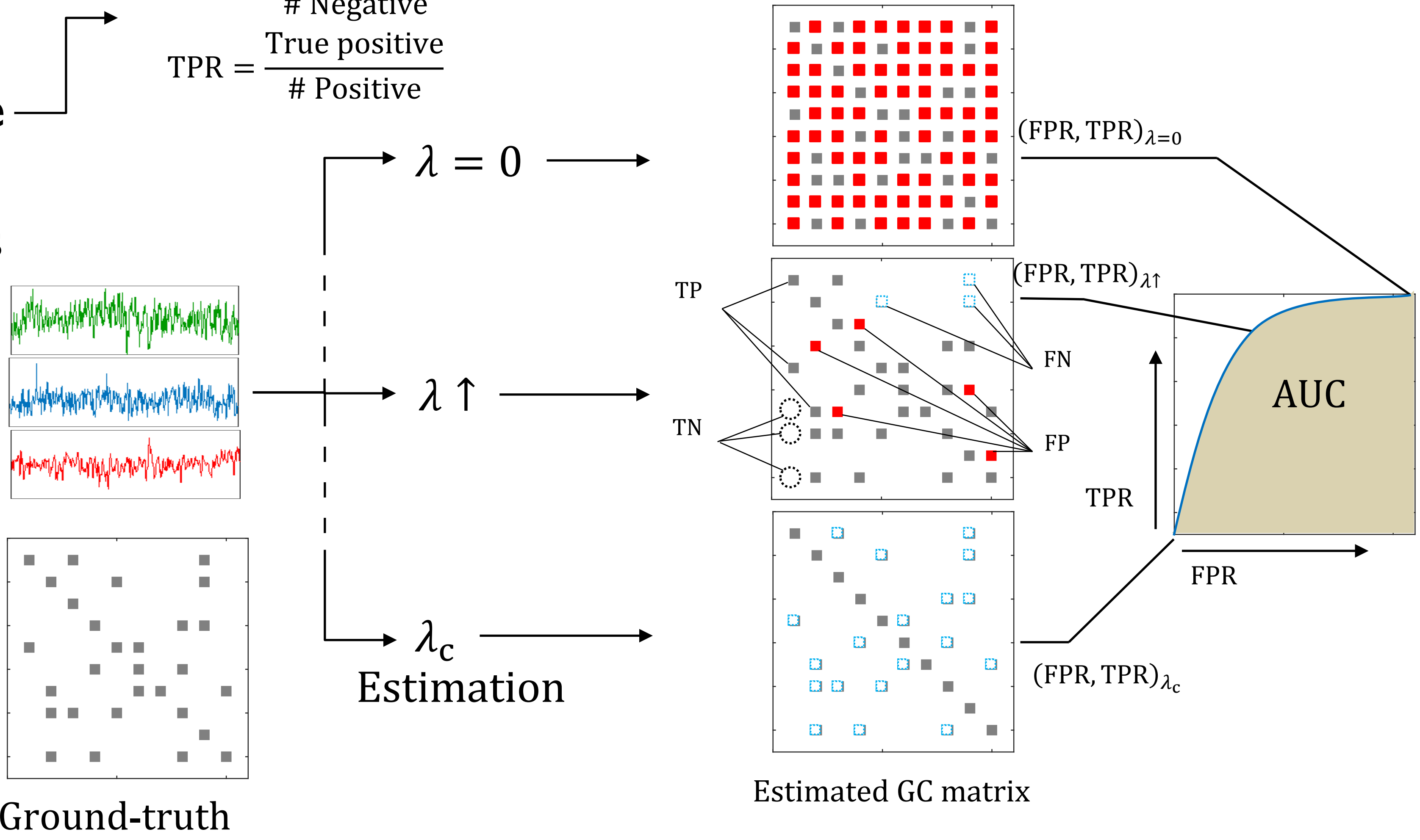


Performance index

- Area under ROC curve

$$\mathrm{FPR} = \frac{\text{False positive}}{\#\ \text{Negative}} \qquad \mathrm{TPR} = \frac{\text{True positive}}{\#\ \text{Positive}}$$

- Relative parameter bias
  - $\dfrac{\|\hat{x}-x_{\mathrm{true}}\|_2}{\|x_{\mathrm{true}}\|_2}$



PRELIMINARY RESULTS

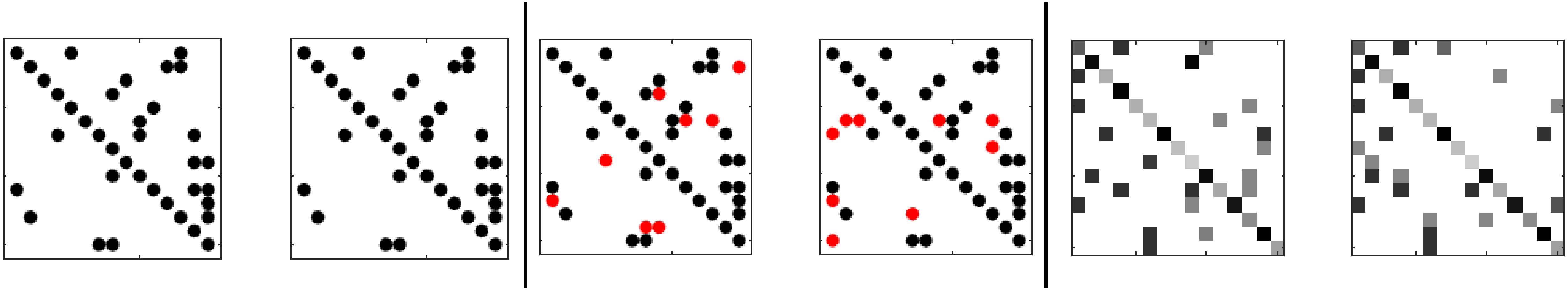
Experiment 3: VAR time-series with prespecified GC patterns generation

Objective: To test the formulations with known given structure

We randomized stable VAR coefficients that the Granger causality patterns are

1. **Common type** ground truth 2. **Differential type** ground truth 3. **Similar type** ground truth

Examples of generated GC matrix topology

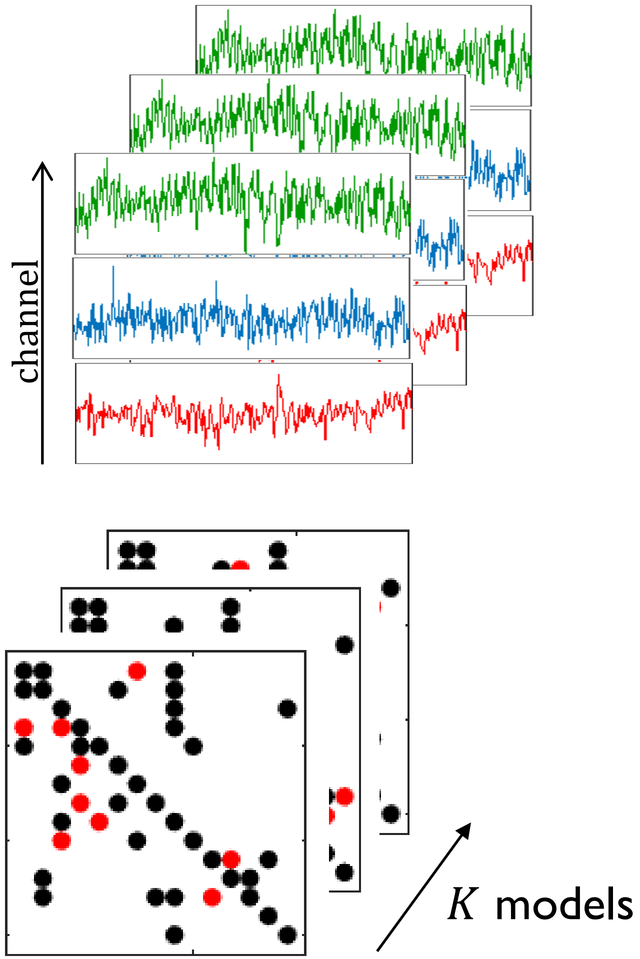


Common network density and **differential network** density can be set.

PRELIMINARY RESULTS

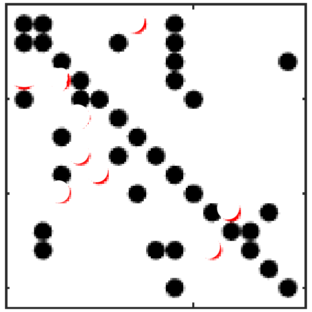
Experiment 4: Group level Granger network extraction

Objective: To extract common GC network with a presence of heterogeneous connections



Groundtruth networks

Formulation C



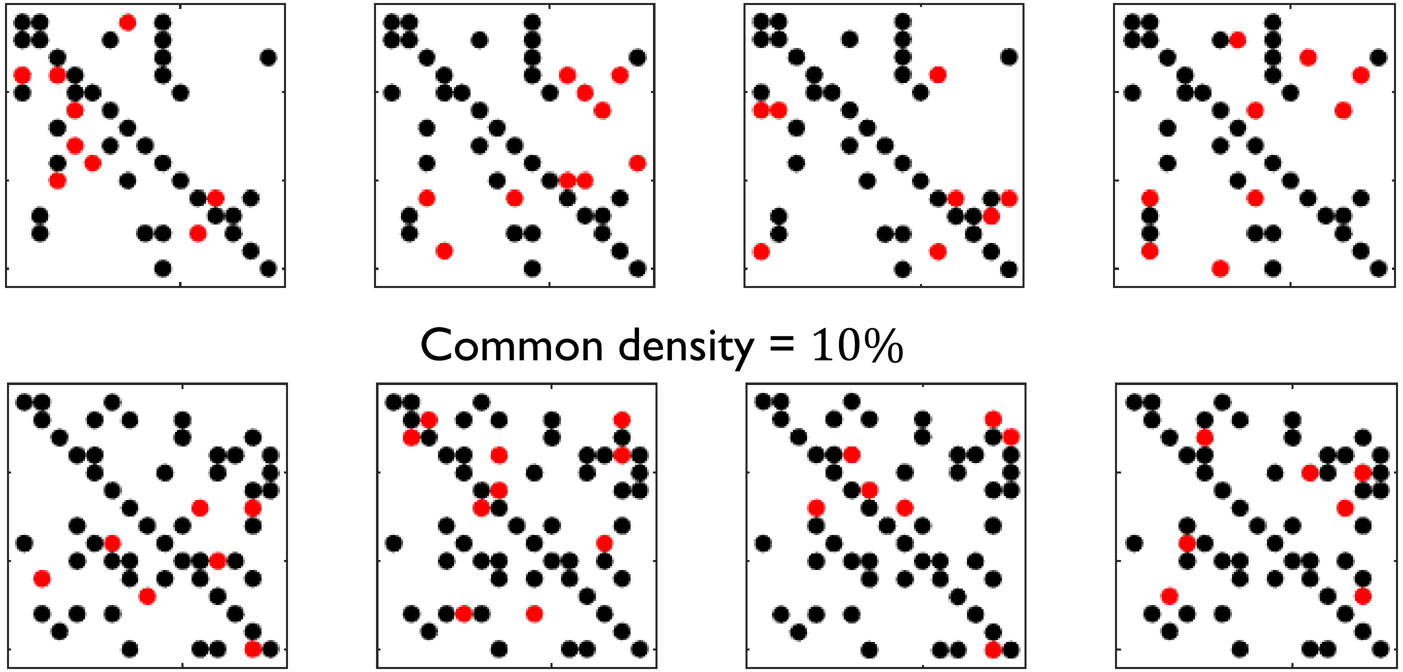
Extracted common network

PRELIMINARY RESULTS

Experiment 4: Group level Granger network extraction

Objective: To extract common GC network with a presence of heterogeneous connections

- 4 sets of 15-dimensional 2nd-order-VAR models
- Common density : 10%, 20%
- Differential density : 5%



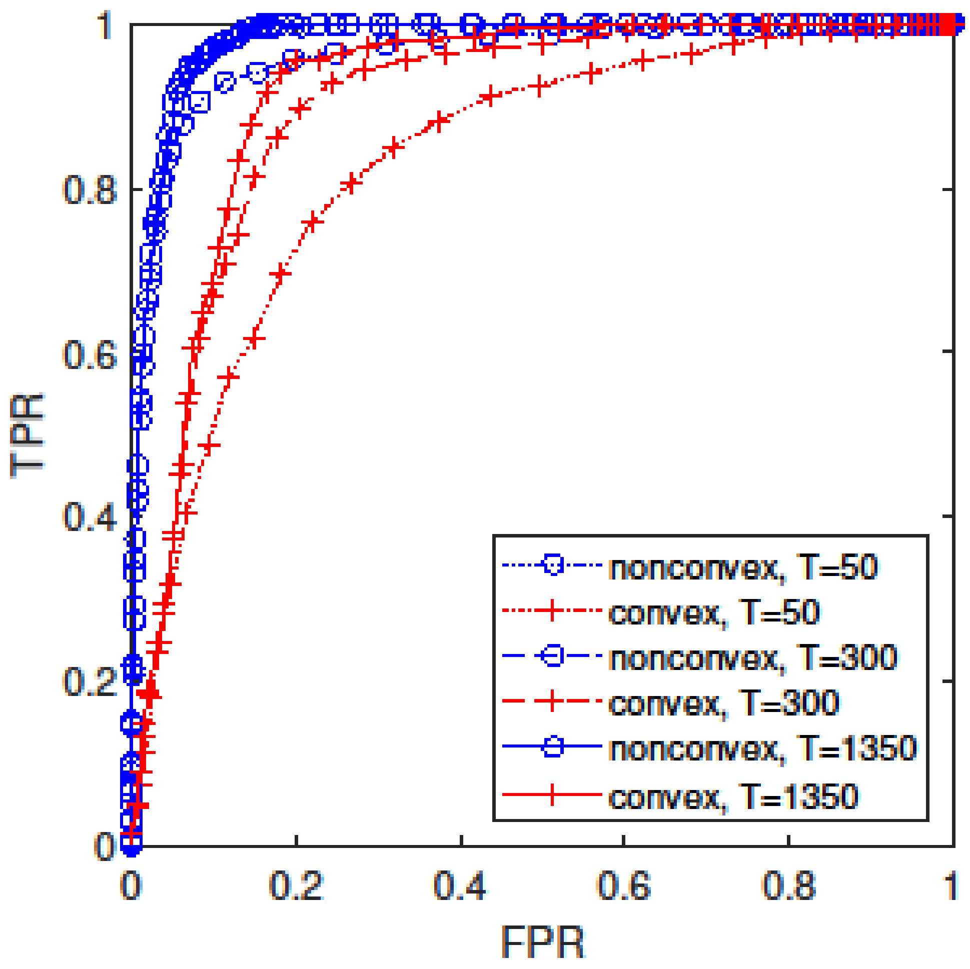
Common density = 10%

Common density = 20%

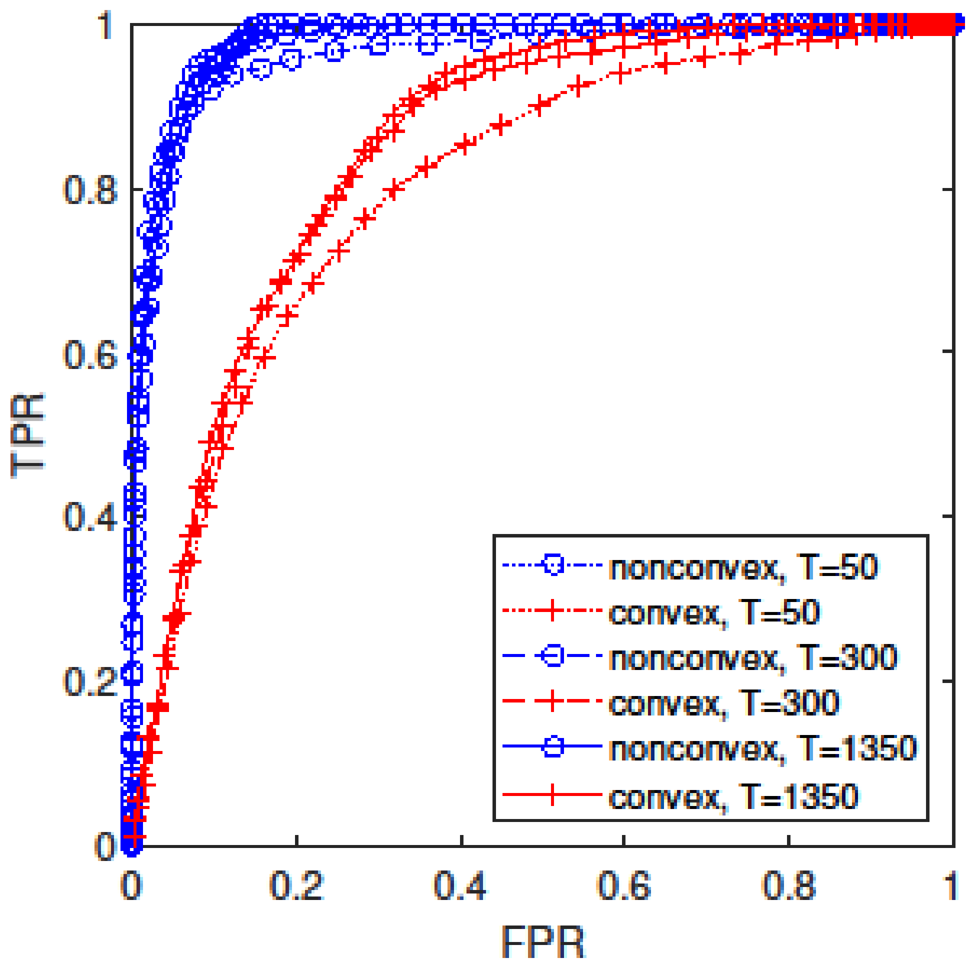
➡ Generate time-series with unit variance Gaussian noise.

PRELIMINARY RESULTS

Experiment 4: Group level Granger network extraction



Common density : $\mathbf{0.1}$

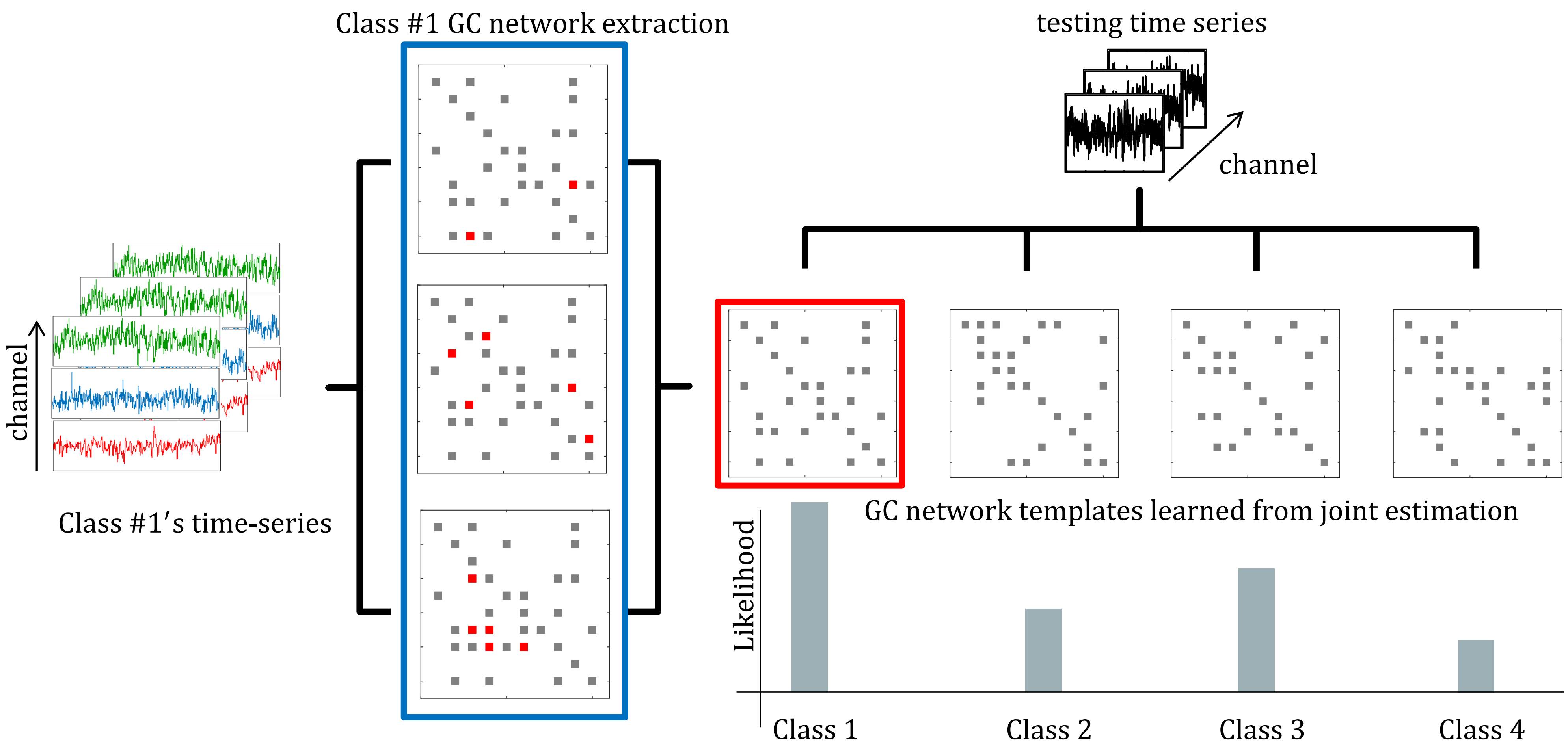


Common density : $\mathbf{0.2}$

PRELIMINARY RESULTS

Experiment 5: Supervised-classification using learned common Granger network

Objective: To illustrate the application of common Granger network extraction



Experiment 5: Supervised-classification using learned common Granger network

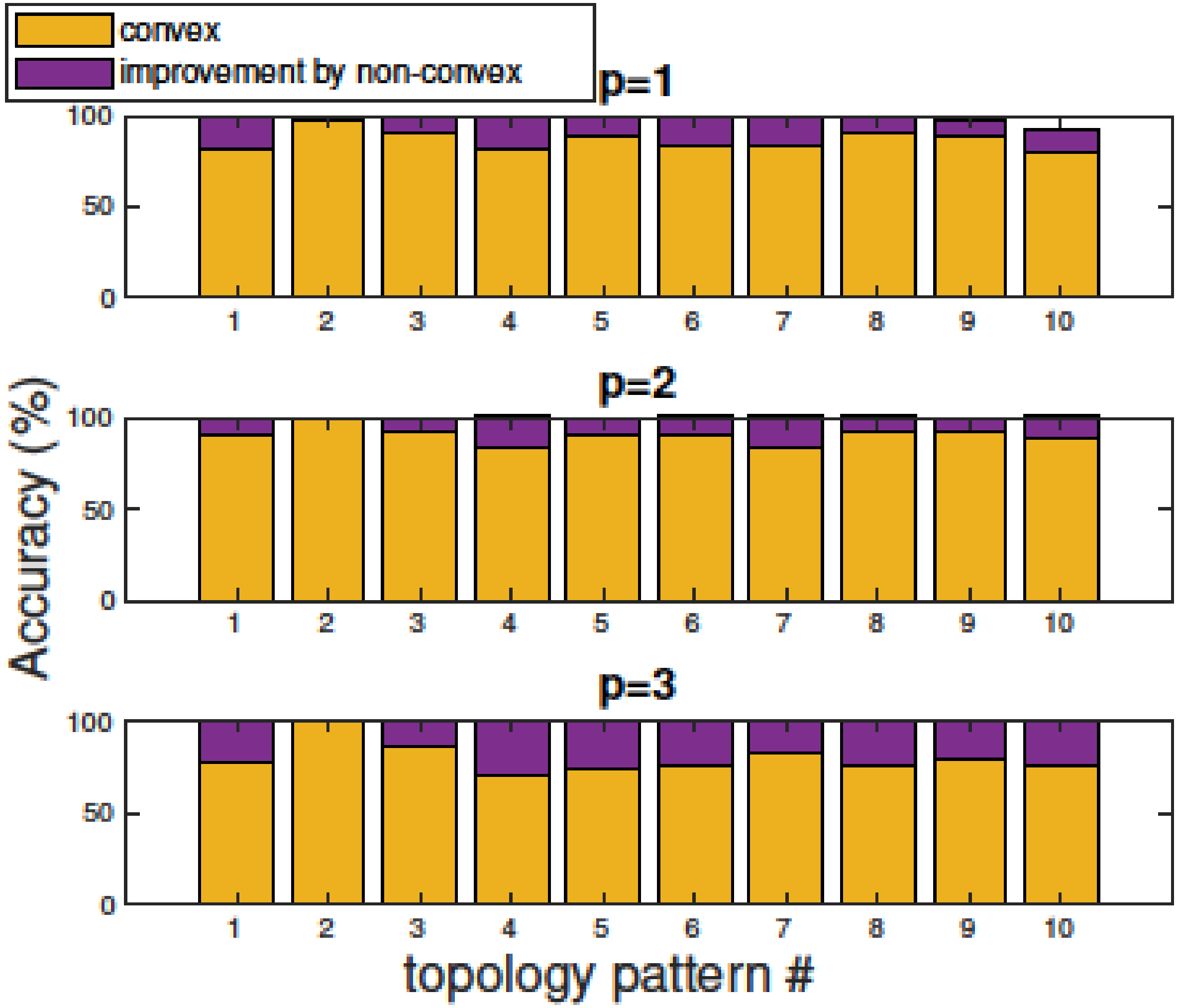
Setting

- 10 GC networks defined on 2nd order 15-dimensional VAR models
- The GC matrix of classifying time-series has sparsity pattern same as one of classes
- Common network density is set to 20%
- vary VAR lag order to test the performance when model order is wrongly chosen

PRELIMINARY RESULTS

Experiment 5: Supervised-classification using learned common Granger network

Result



- Near perfect classification rate in non-convex case
- Non-convex case did not deteriorate much when model order is wrong compared to convex case.

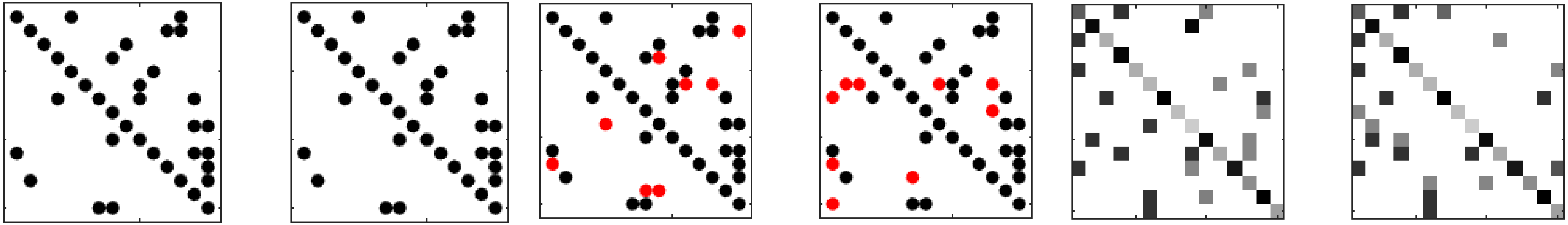
PRELIMINARY RESULTS

Experiment 6: Performance of differential priors

Objective: To illustrate the performance of formulation D

Setting

- 4 sets of 15-dimensional 2nd-order-VAR models
- Common network density is set to 20%
- Differential network density is set to 5%
- The ground-truth types are **common type**, **differential type**, **similar type**



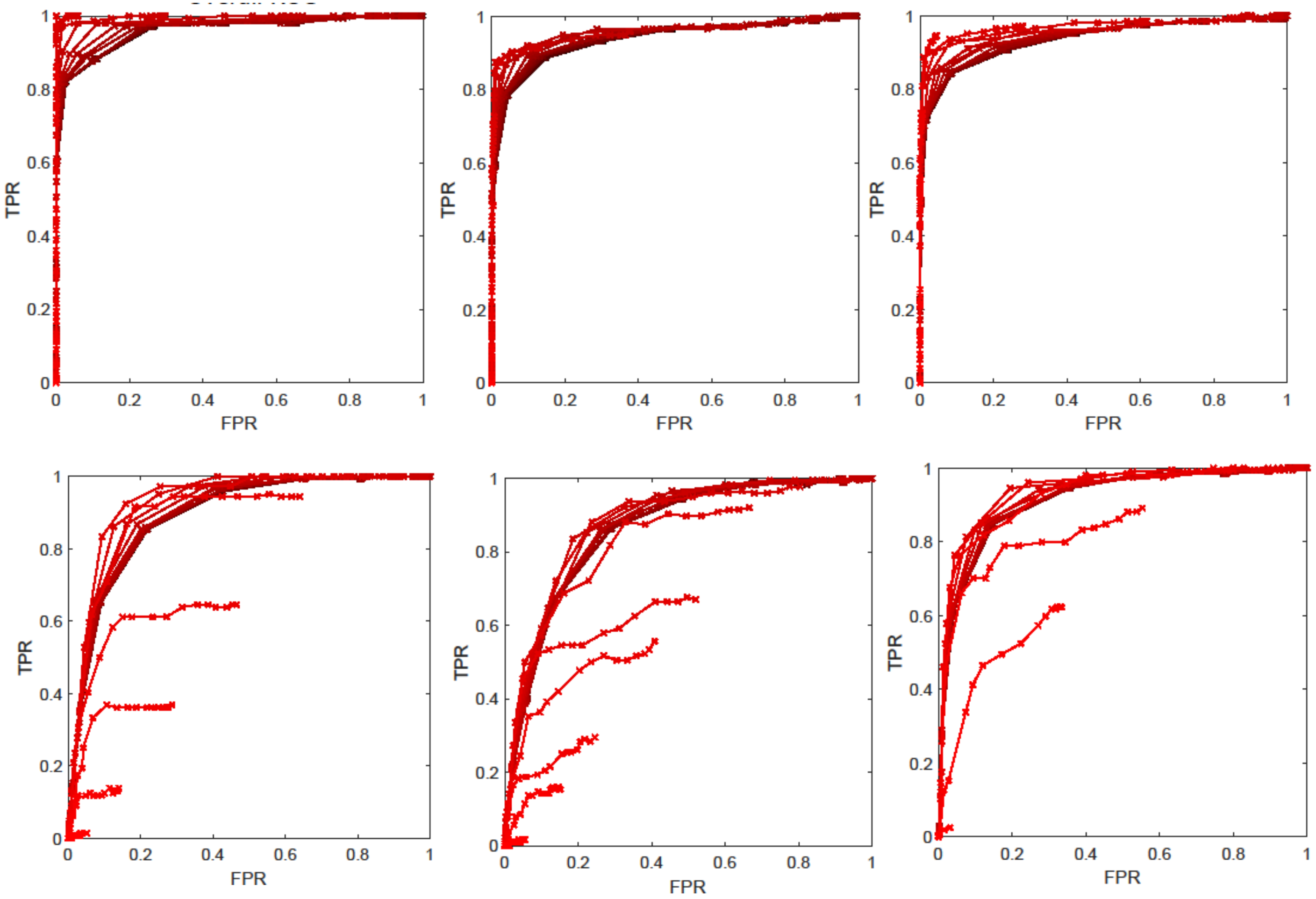
common type differential type similar type

PRELIMINARY RESULTS

Experiment 6: Performance of differential priors

Ground truth type

C D S



Overall ROC

non-convex formulation D
using
Group norm penalty

convex formulation D
using
Group lasso

FUTURE WORK

Experiments

- Effectiveness of formulations
  - Experiment 4: Common Granger network extraction
  - Experiment 5: Classification
  - Experiment 6: Effectiveness of formulation D
  - Experiment 7: Effectiveness of formulation S
- Brain network application
  - Experiment 8: Application on fMRI time-series

Thesis writing & Publication

"Learning A Common Granger Causality Network Using A Non−Convex Regularization", ICASSP-2020

Formulation C (non-convex)

Goal

- Control the convergence of ADMM algorithm to solve formulation D, S
- Increase performance of algorithms
- Apply formulations on fMRI data

Q&A

REFERENCES

[Bore20] J. C. Bore, P. Li, D. J. Harmah, F. Li, D. Yao, P. Xu, Directed EEG neural network analysis by LAPPS (p≤1) Penalized sparse Granger approach, Neural Networks, Volume 124, 2020, Pages 213-222,

[Boyd11] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers", Foundation and Trends in Machine Learning, vol. 3, no. 1, pp. 1-122, Jan. 2011.

[Chartrand08] R. Chartrand, V. Staneva,. (2008). Restricted isometry properties and nonconvex compressive sensing. Inverse Problems 24(3).

[Chun15] H. Chun, X. Zhang, and H. Zhao, "Gene regulation network inference with joint sparse Gaussian graphical models," Journal of Computational and Graphical Statistics, vol. 24, no. 4, pp. 954–974, 2015.

[Combettes11] P. L. Combettes and J. C. Pesquet. Proximal Splitting Methods in Signal Processing, pages 185-212. Springer New York, New York, NY, 2011.

[Granger1980] C.W.J. Granger, Testing for causality: A personal viewpoint, Journal of Economic Dynamics and Control, Volume 2, 1980, Pages 329-352, ISSN 0165-1889,

[Gregorova15] M. Gregorova, A. Kalousis, and S. Marchand-Maillet. Learning coherent Granger causality in panel vector autoregressive models. In Proceedings of the Demand Forecasting Workshop of the 32nd International Conference on Machine Learning. ICML, 2015.

[Guo11] J. Guo, E. Levina, G. Michailidis, and J. Zhu, "Joint estimation of multiple graphical models," Biometrika, vol. 98, no. 1, pp. 1–15, 2011.

[Hu17] Hu, C. Li, K. Meng, J. Qin, and X. Yang, "Group sparse optimization via $\ell_{p,q}$ regularization," Journal of Machine Learning Research, vol. 18, no. 30, pp. 1–52, 2017.

REFERENCES

[Huang15] F. Huang and S. Chen, "Joint learning of multiple sparse matrix Gaussian graphical models," IEEE Transactions on Neural Networks and Learning Systems, vol. 26, no. 11, pp. 2606–2620, 2015.

[Li15] H. Li and Z. Lin, "Accelerated proximal gradient methods for nonconvex programming," in Advances in Neural Information Processing Systems 28, pp. 379–387, 2015.

[Skrip19] A. Skripnikov and G. Michailidis, "Joint estimation of multiple network Granger causal models," Econometrics and Statistics, vol. 10, pp. 120–133, 2019.

[Skrip19] A. Skripnikov and G. Michailidis, "Regularized joint estimation of related vector autoregressive models," Computational Statistics & Data Analysis, vol. 139, pp. 164–177, 2019.

[Songsiri 17] J. Songsiri. Estimations in Learning Granger Graphical Models with Application to fMRI Time Series. Technical report, Chulalongkorn University, Department of Electrical engineering, July 2017.

[Teboulle18] M. Teboulle. A simplified view of first order methods for optimization. Math. Program. 170, 1 (2018), 67-96.

[Tuck20] J. Tuck and S. Boyd. Fitting Laplacian regularized stratified Gaussian models. ArXiv, abs/2005.01752, 2020.

[Wang19] Y. Wang, W. Yin, J. Zeng. Global Convergence of ADMM in Nonconvex Nonsmooth Optimization. Journal of Scientific Computing 78, 29–63 (2019).

[Wilms18] I. Wilms, L. Barbaglia, and C. Croux, "Multiclass vector auto-regressive models for multistore sales data," Journal of the Royal Statistical Society: Series C (Applied Statistics), vol. 67, no. 2, pp. 435–452, 2018.

[Xu17] Z. Xu, M. Figueiredo, T. Goldstein, "Adaptive ADMM with Spectral Penalty Parameter Selection," Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, PMLR 54:718-727, 2017.

Objective	TASK	PHASE ONE: June 2019	PHASE ONE: July 2019	PHASE ONE: August 2019	PHASE TWO: September 2019	PHASE TWO: October 2019	PHASE TWO: November 2019	PHASE TWO: December 2019	PHASE TWO: January 2020	PHASE TWO: Febuary 2020	PHASE THREE: March 2020	PHASE THREE: April 2020	PHASE THREE: May 2020
1	**Literature review on sparse formulation in graphical models**												
1.1	Joint sparse estimation of Gaussian graphical models	3, 4											
1.2	Joint sparse estimation of Granger graphical models	4	1, 2, 3, 4										
1.3	Survey of non-smooth, non-convex optimization algorithms		3, 4	1, 2, 3, 4					1, 2, 3, 4	1, 2, 3, 4	1, 2, 3, 4	1, 2, 3, 4	1, 2
1.5	Survey of non-convex regularization		3, 4	1, 2, 3, 4					1, 2, 3, 4	1, 2, 3, 4	1, 2, 3, 4	1, 2, 3, 4	1, 2
2	**Algorithm implementation & Experiment design**												
2.1	ADMM algorithm implementation for non-convex optimization			4	1								
2.2	nmAPG algorithm implementation				1, 2								
2.3	Coding optimization			4	1, 2, 3, 4	1, 2, 3, 4	1, 2, 3, 4	1, 2, 3, 4	1, 2, 3, 4	1, 2, 3, 4	1, 2, 3, 4	1, 2, 3, 4	1, 2
2.4	Experiments planning				3, 4	1, 2							
3	**Perform simulated data experiments**												
3.1	Experiment 1: initial point selection in linear regression model				1, 2								
3.2	Experiment 2: non-convex group norm regularization in linear regression models				1, 2								
3.3	Experiment 3: VAR time-series with pre-specified GC patterns generation	1, 2			2, 3, 4	1, 2							
3.4	Experiment 4: Common Granger network extraction				2, 3, 4	1, 2							
3.5	Experiment 5: Supervised-classification using learned common Granger network				2, 3, 4	1, 2							
3.6	Submit part of the proposal work to ICASSP					3, 4							
3.7	Experiment 6: Effectiveness of differential priors								1, 2, 3, 4	1, 2, 3, 4			
3.8	Preparing proposal										1, 2, 3, 4	1, 2, 3, 4	1, 2, 3, 4

Objective	TASK	PHASE FOUR: June 2020	PHASE FOUR: July 2020	PHASE FOUR: August 2020	PHASE FIVE: September 2020	PHASE FIVE: October 2020	PHASE FIVE: November 2020	PHASE SIX: December 2020	PHASE SIX: January 2021
4	**Future works**								
4.1	Spectral ADMM implementation & Optimization	1, 2, 3, 4							
4.2	Literature review on fMRI data preprocessing	2, 3, 4							
4.3	Perform real data experiments		1, 2, 3, 4	1, 2, 3, 4	1, 2, 3, 4	1, 2, 3, 4	1, 2, 3, 4		
4.4	Conclude result & preparing thesis defense							1, 2, 3, 4	1, 2, 3, 4

SUPPLEMENTARY: FORMULATION COST FUNCTION

$$(1/2)\ \left\|Y^{(k)} - A^{(k)}H^{(k)}\right\|_2^2$$

Least square (individual)

$$\sum_{k=1}^{K}(1/2)\ \left\|Y^{(k)} - A^{(k)}H^{(k)}\right\|_2^2$$

Least square (joint)

$$A^{(k)} = \left[\hat{A}_1^{(k)} \dots, \hat{A}_p^{(k)}\right]$$

$$B_{ij}^{(k)} = [(A_1^{(k)})_{ij} \dots (A_p^{(k)})_{ij}]$$

$$C_{ij} = [B_{ij}^{(1)} \dots B_{ij}^{(K)}]$$

Regularization

$$\sum_{k=1}^{K}\sum_{i \neq j}\left\|C_{ij}\right\|_2$$

Formulation C

$$\sum_{k=1}^{K}\sum_{i \neq j}\left\|B_{ij}^{(k)}\right\|_2$$

$$\sum_{k=1}^{K}\sum_{i \neq j}\left\|C_{ij}\right\|_2$$

Formulation D

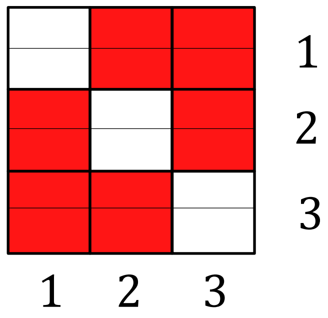
$$\sum_{k=1}^{K}\sum_{i \neq j}\left\|B_{ij}^{(k)}\right\|_2$$

$$\sum_{k<k'}\sum_{i \neq j}\left\|B_{ij}^{k} - B_{ij}^{k'}\right\|_2$$

Formulation S

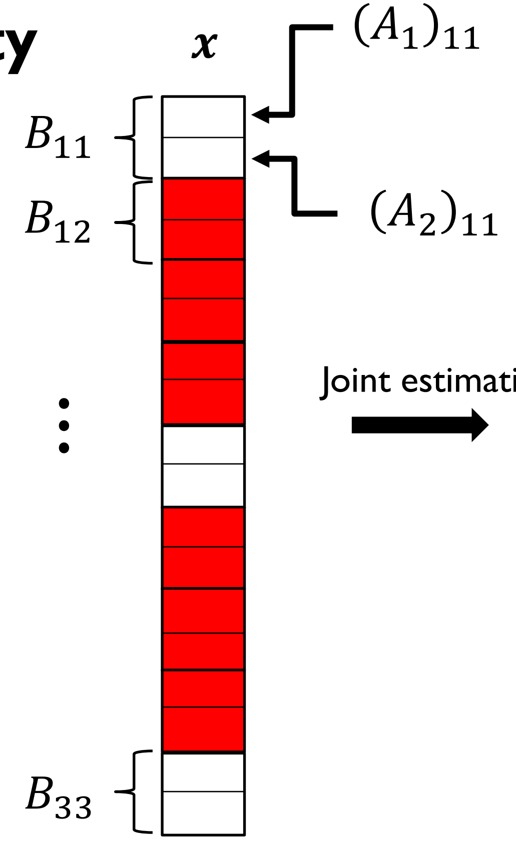
SUPPLEMENTARY

Group norm penalty

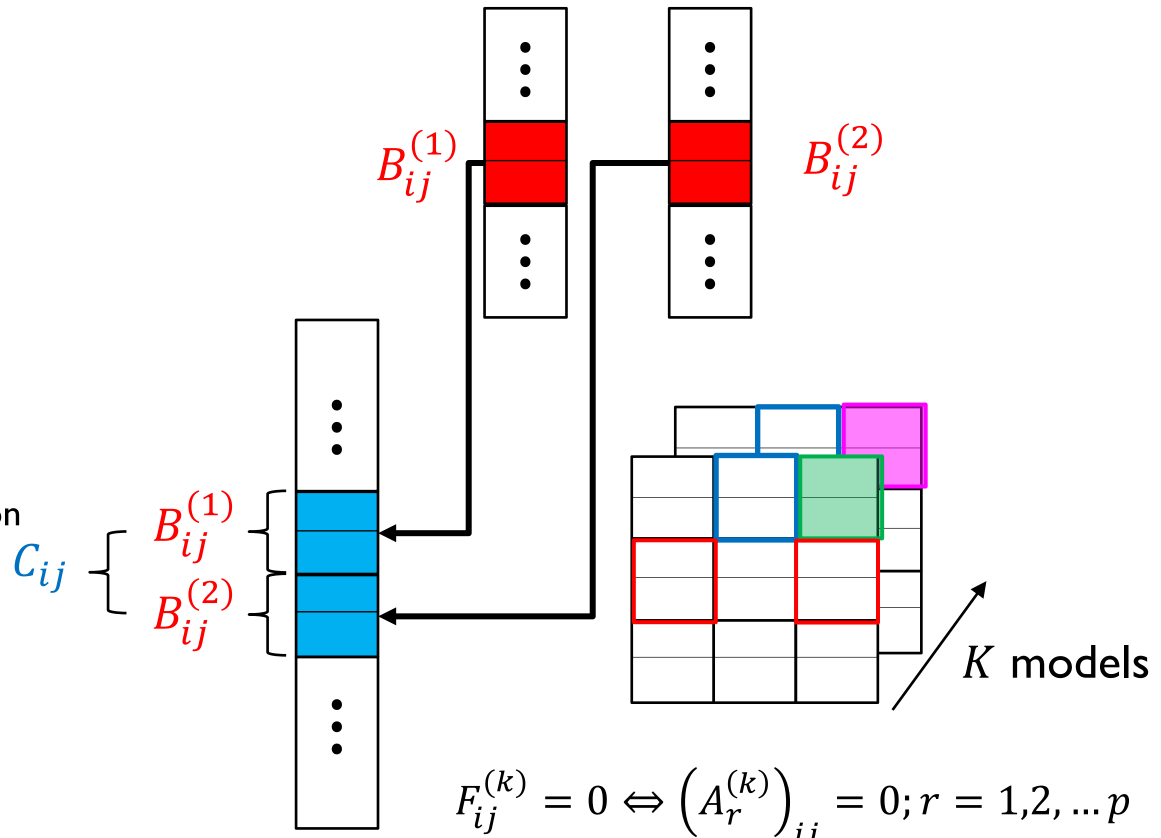


Stacked VAR coefficient matrix

vectorize



Joint estimation



$F_{ij}^{(k)} = 0 \Leftrightarrow \left(A_r^{(k)}\right)_{ij} = 0; r = 1,2, \dots p$

Knowing that the sparsity must be a block of size p

Penalize $\sum_{i \neq j} \left\| B_{ij} \right\|_2^q \rightarrow \|\boldsymbol{Px}\|_{2,q}^{(p)}$

$q = 1$, Group lasso

$q = 1/2$, Our non-convex extension

Penalize $\sum_{i \neq j} \left\| C_{ij} \right\|_2^q \rightarrow \|\boldsymbol{Px}\|_{2,q}^{(pK)}$

Penalize $\sum_{m \neq l} \sum_{i \neq j} \left\| B_{ij}^{(m)} - B_{ij}^{(l)} \right\|_2^q \rightarrow \|\boldsymbol{Dx}\|_{2,q}^{(p)}$

Penalize $\sum_k \sum_{i \neq j} \left\| B_{ij}^{(k)} \right\|_2^q \rightarrow \|\boldsymbol{Px}\|_{2,q}^{(p)}$

SUPPLEMENTARY | VECTORIZATION

C. $$\min_{x} \|y - Gx\|_2^2 + \lambda \|Px\|_{2,q}^{(pK)}$$

D. $$\min_{x} \|y - Gx\|_2^2 + \lambda_1 \|Px\|_{2,q}^{(p)} + \lambda_2 \|Px\|_{2,q}^{(pK)}$$

S. $$\min_{x} \|y - Gx\|_2^2 + \underbrace{\lambda_1 \|Px\|_{2,q}^{(p)} + \lambda_2 \|Dx\|_{2,q}^{(p)}}_{g(z)}$$

$$\begin{bmatrix} T_1 \\ T_2 \end{bmatrix} x - z = 0$$

SUPPLEMENTARY : ADMM STEP

$$x^+ = \underset{x}{\operatorname{argmin}} \|Gx - b\|_2^2 + \frac{\rho}{2} \left\| \begin{bmatrix} L_1 \\ L_2 \end{bmatrix} x - z + \frac{y}{\rho} \right\|_2^2$$

$$z^+ = \min_{z_1, z_2} \lambda_1 \|z_1\|_{2,q}^{(M_1)} + \lambda_2 \|z_2\|_{2,q}^{(M_2)} + \frac{\rho}{2} \left\| \begin{bmatrix} L_1 \\ L_2 \end{bmatrix} x - z + \frac{y}{\rho} \right\|_2^2$$

$$y^+ = y + \rho\left(\begin{bmatrix} L_1 \\ L_2 \end{bmatrix} x - z\right)$$

Monotone accelerated proximal gradient (mAPG)

Beck & Teboulle

Descent

$$y_k = x_k + \frac{t_{k-1} - 1}{t_k}(x_k - x_{k-1}) + \frac{t_{k-1}}{t_k}(z_k - x_k)$$

$$t_{k+1} = 0.5(1 + \sqrt{1 + 4t_k^2})$$

$$z_{k+1} = \mathrm{prox}_{\lambda g}(y_k - \lambda \nabla f(y_k))$$

$$x_{k+1} = \mathrm{argmin}\{F(x_k), F(z_{k+1})\}$$ **Monitoring step**

↓

Does not generate sufficient decreasing sequence.

SUPPLEMENTARY

Monotone accelerated proximal gradient (mAPG)

Beck & Teboulle	→	Li & Lin
Descent	→	Sufficient descent

$$y_k = x_k + \frac{t_{k-1} - 1}{t_k}(x_k - x_{k-1}) + \frac{t_{k-1}}{t_k}(z_k - x_k)$$

$$t_{k+1} = 0.5(1 + \sqrt{1 + 4t_k^2})$$

$$z_{k+1} = \mathrm{prox}_{\lambda g}(y_k - \lambda \nabla f(y_k))$$

→ Compute original proximal gradient step

$$v_{k+1} = \mathrm{prox}_{\lambda g}(x_k - \lambda \nabla f(x_k))$$

$$x_{k+1} = \mathrm{argmin}\{F(\boldsymbol{v_{k+1}}), F(z_{k+1})\}$$

Monitoring step → Sufficient descent

Monotone APG is proved to converge in some **non-convex** problems.

However, monitoring step is **too conservative**.

SUPPLEMENTARY

Monotone accelerated proximal gradient (mAPG)

Beck & Teboulle → Li & Lin

Descent → Sufficient descent

$$y_k = x_k + \frac{t_{k-1} - 1}{t_k}(x_k - x_{k-1}) + \frac{t_{k-1}}{t_k}(z_k - x_k)$$

$$t_{k+1} = 0.5(1 + \sqrt{1 + 4t_k^2})$$

$$z_{k+1} = \text{prox}_{\lambda g}(y_k - \lambda \nabla f(y_k))$$

$$F(z_{k+1}) \leq F(x_k) - \delta \|z_{k+1} - x_k\|_2^2 ?$$

YES: $x_{k+1} = z_{k+1}$

No proximal step

NO: $x_{k+1} = \text{prox}_{\lambda g}(x_k - \lambda \nabla f(x_k))$

Can **sufficient descent property** can be dropped in non-convex setting ?

There is a trick.

Non-monotone accelerated proximal gradient (nmAPG)

In objective function

Beck & Teboulle → Li & Lin

Descent → Sufficient descent

$$y_k = x_k + \frac{t_{k-1} - 1}{t_k}(x_k - x_{k-1}) + \frac{t_{k-1}}{t_k}(z_k - x_k)$$

$$t_{k+1} = 0.5(1 + \sqrt{1 + 4t_k^2})$$

$$z_{k+1} = \text{prox}_{\lambda g}(y_k - \lambda \nabla f(y_k))$$

→ $F(z_{k+1}) \leq \boldsymbol{c_k} - \delta \|z_{k+1} - y_k\|_2^2$?

YES: $x_{k+1} = z_{k+1}$ (No proximal step)

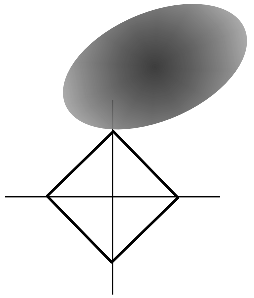
NO: Monitoring step in mAPG → $x_{k+1} = \text{argmin}\{F(\boldsymbol{v_{k+1}}), F(z_{k+1})\}$

c_k is weighted average of objective function in iterations $1, \dots, k$.

Sequence c_k is strictly monotone decreasing while $\boldsymbol{F(x_k)}$ **may not.**

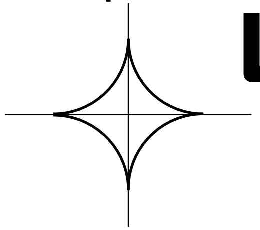
SUPPLEMENTARY

Choices of sparsity inducing-penalty



Group lasso → Penalize → Sum of 2-norm of variables components → Interpreted as → 2-norm composite with 1-norm

1-norm → Sparsity inducing part → 2-norm composite with ½ quasi-norm



2-norm composite with ½ quasi-norm → **NON-CONVEX**

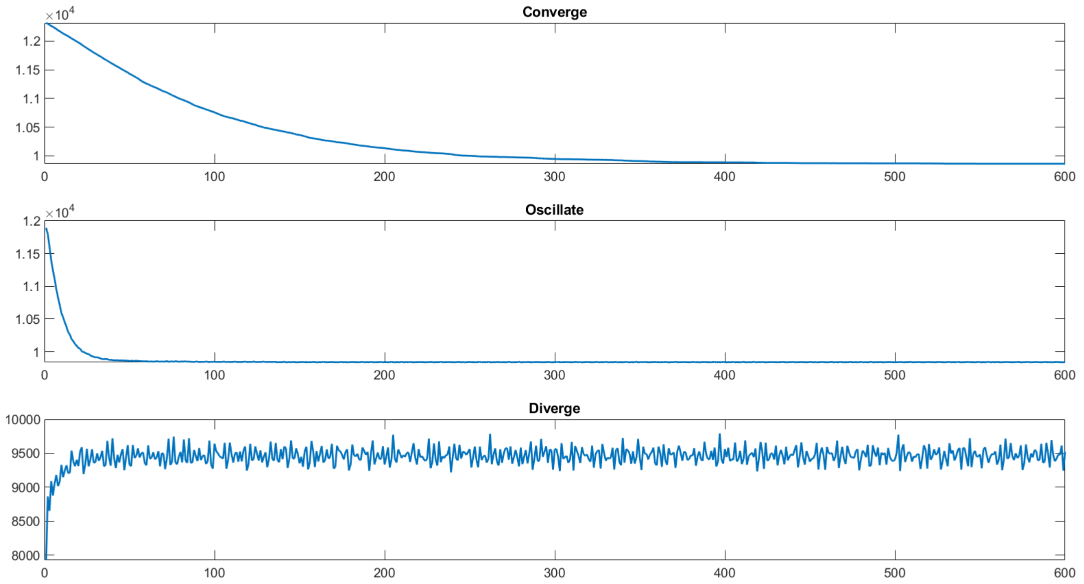
2-norm composite with 1-norm → **CONVEX**

2-norm composite with ½ quasi-norm → **Weaker true support recovery condition** than Group lasso

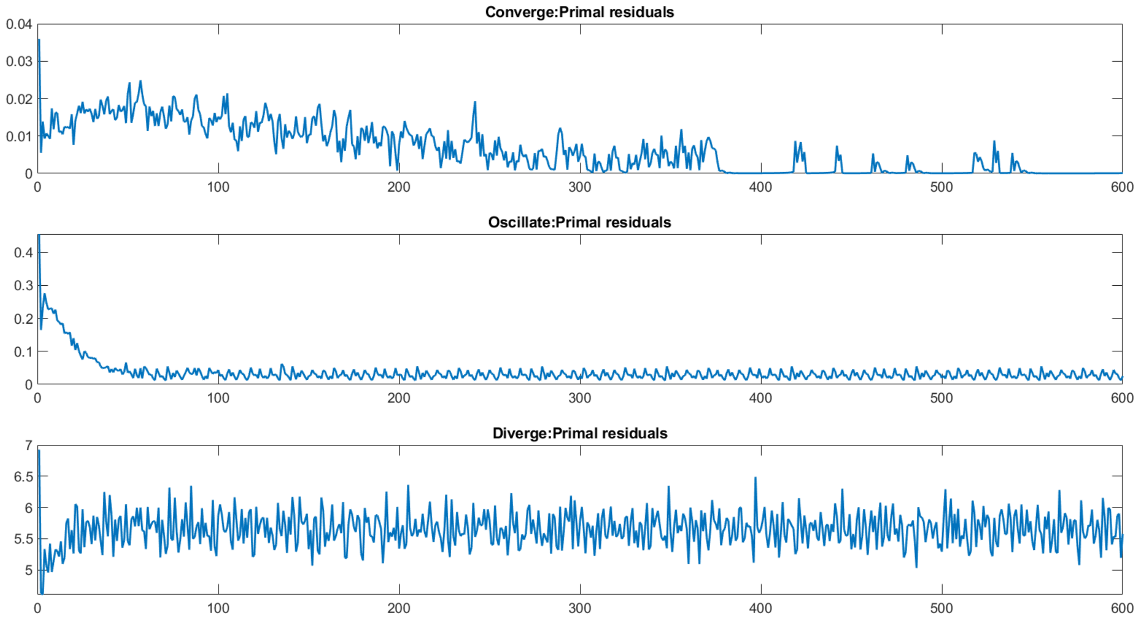
SUPPLEMENTARY

ADMM convergence issues

$$\min_{Ax+Bz=c} f(x) + g(z)$$



$f(x) + g(z)$



$\|Ax + Bz - c\|_2$

$$\mathcal{L}(x, z, y, \rho) = f(x) + g(z) + y^T r + \frac{\rho}{2}\|r\|_2^2$$

$$r = (Ax + Bz - c)$$