JOINT ESTIMATION OF MULTIPLE GRANGER GRAPHICAL MODELS USING NON-CONVEX PENALTY FUNCTIONS

Thesis presentation

Parinthorn Manomaisaowapak

Advisor: Assoc. Prof. Dr. Jitkomut Songsiri

Department of Electrical Engineering, Faculty of Engineering

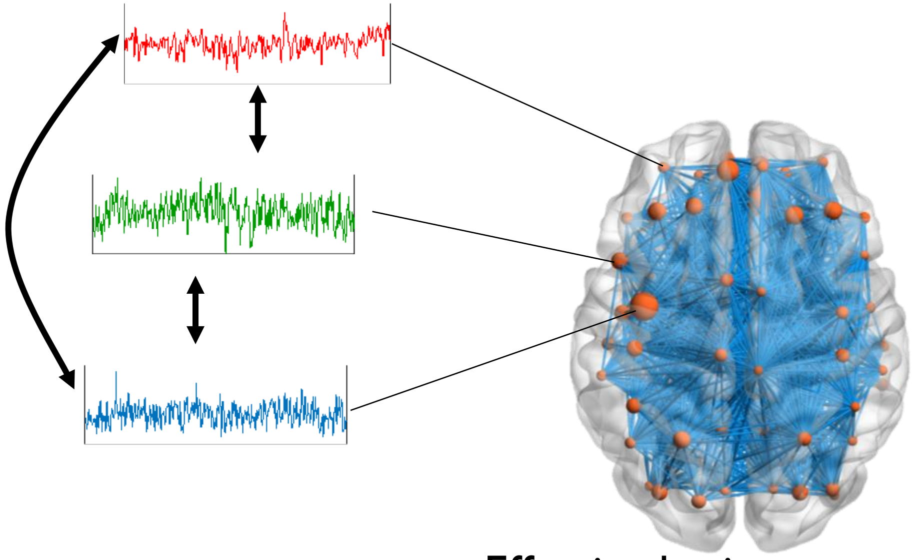
Chulalongkorn University

OUTLINE

- Introduction
- Background
- Methodology
- Algorithms
- Results
- Conclusion

INTRODUCTION

How to study relationship of time-series?



Effective brain connectivity

Causality analysis

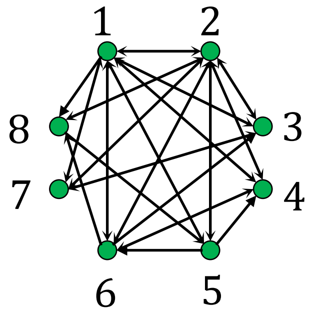
A strength of evidence

Granger causality(GC)

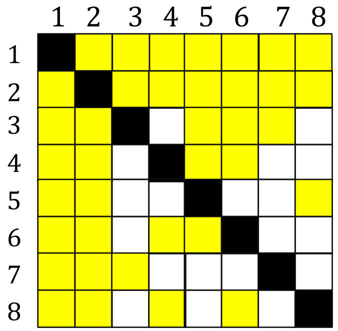
Based on dynamical models

Has direction

Causality network



Causality matrix



	1	2	3	4	5	6	7	8
1	⬛	🟨	🟨	🟨	🟨	🟨	🟨	🟨
2	🟨	⬛	🟨	🟨	🟨	🟨	🟨	🟨
3	🟨	🟨	⬛	⬜	🟨	🟨	🟨	⬜
4	🟨	🟨	⬜	⬛	🟨	🟨	⬜	⬜
5	🟨	🟨	⬜	⬜	⬛	⬜	⬜	🟨
6	🟨	🟨	⬜	🟨	🟨	⬛	⬜	⬜
7	🟨	🟨	🟨	⬜	⬜	⬜	⬛	⬜
8	🟨	🟨	⬜	🟨	⬜	🟨	⬜	⬛

Granger graphical model

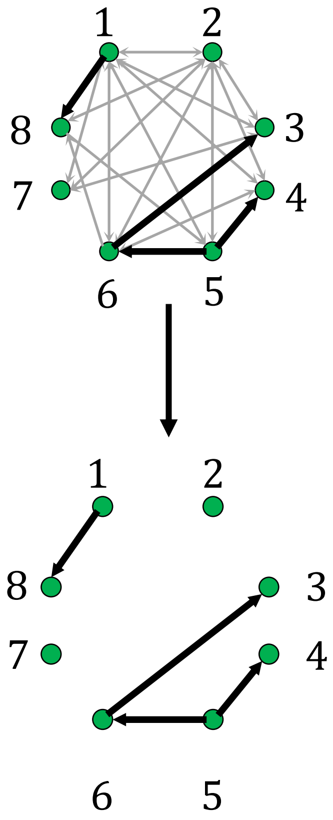
INTRODUCTION

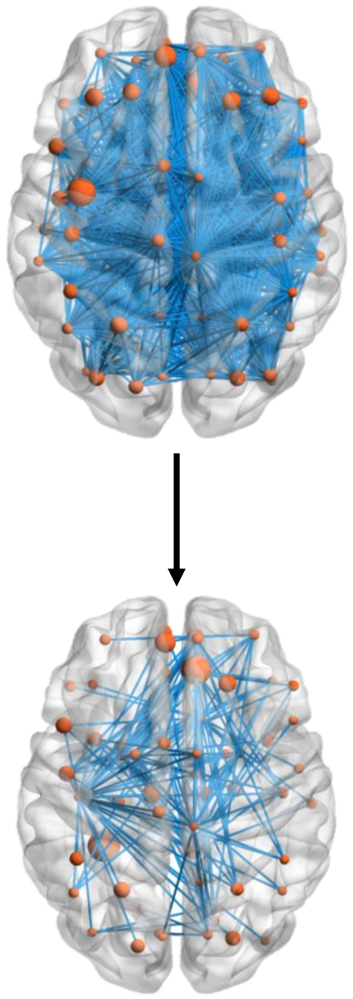
High dimensional GC network

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

Sparse estimation of GC network

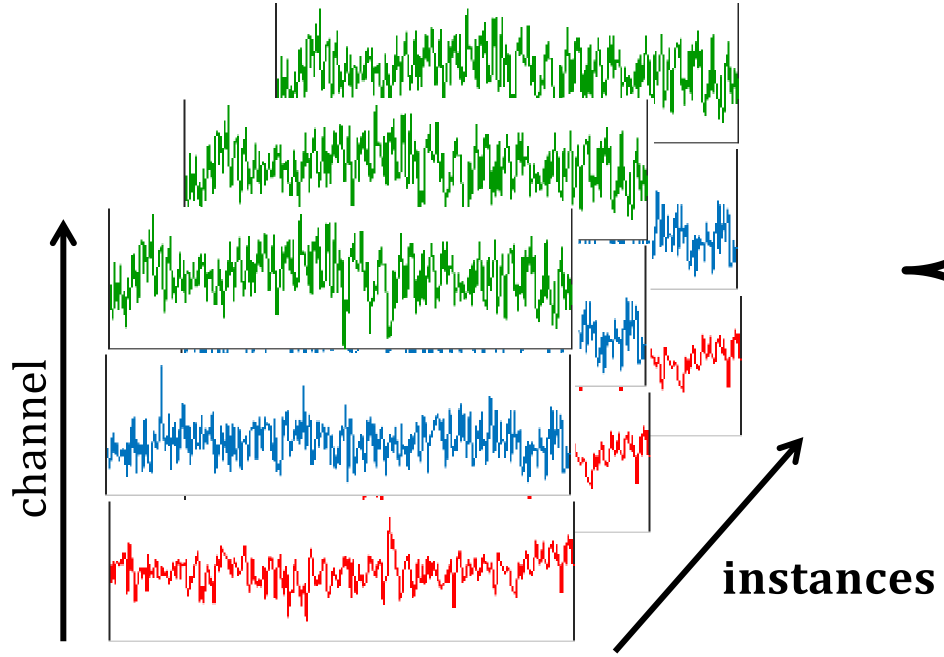
Causality network





INTRODUCTION

High dimensional multiple GC networks



analyze each **instance** separately → the relations will be revealed if the samples are large enough

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

Joint estimation of multiple models

BACKGROUND

Vector autoregressive model (VAR)

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

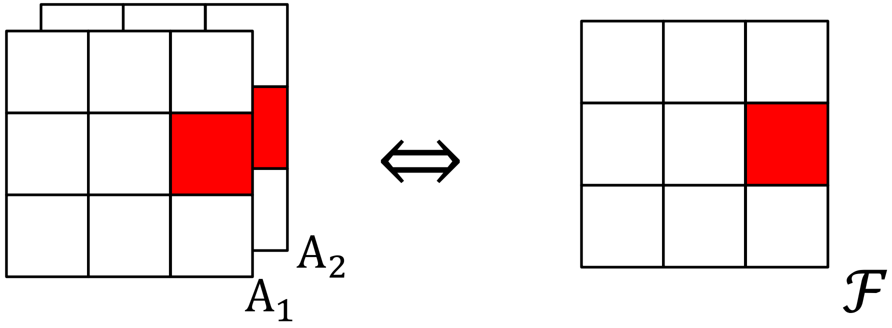
$A_r \in \mathbf{R}^{n\times n}$ → Least-square estimation

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

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

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

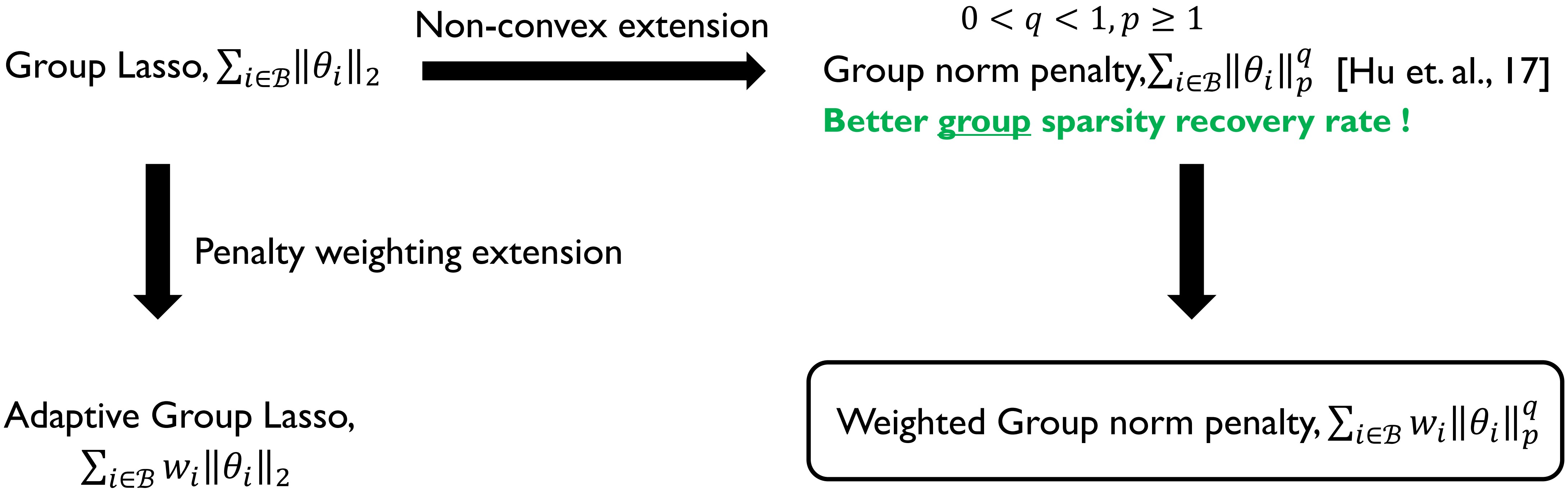


How can we force all VAR lags to be zero at once? → Regularized least-square estimation

penalty: Group lasso

BACKGROUND

Group norm penalty regularized regression



Group Lasso, $\sum_{i\in\mathcal{B}} ||\theta_i||_2$

Non-convex extension

$0 < q < 1, p \geq 1$

Group norm penalty, $\sum_{i\in\mathcal{B}} ||\theta_i||_p^q$ [Hu et. al., 17]

Better group sparsity recovery rate !

Penalty weighting extension

Adaptive Group Lasso, $\sum_{i\in\mathcal{B}} w_i ||\theta_i||_2$

Weighted Group norm penalty, $\sum_{i\in\mathcal{B}} w_i ||\theta_i||_p^q$

METHODOLOGY

High dimensional <u>multiple</u> GC networks

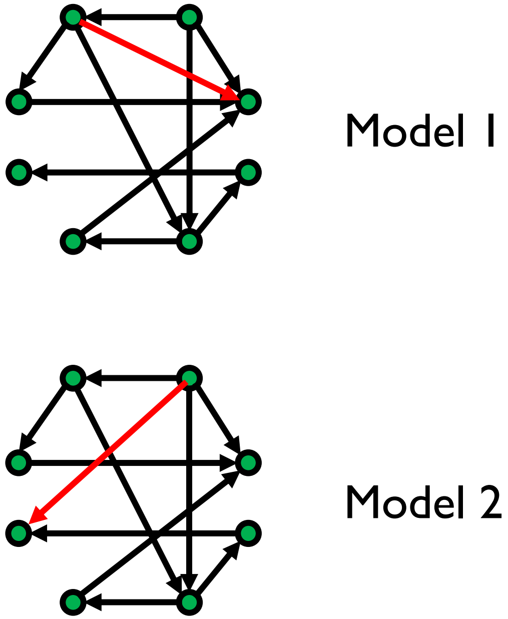
Joint estimation of multiple models

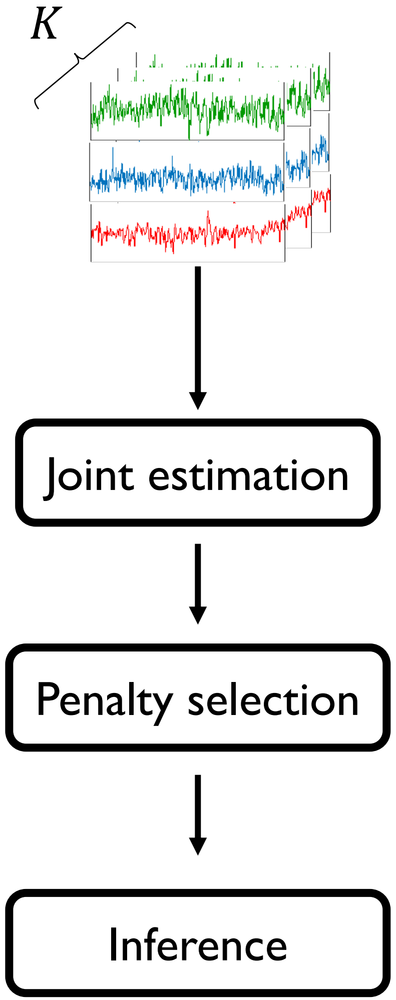
$$\min_{\theta_1,\ldots,\theta_K} \sum_i \left[f(\theta_i) + \boldsymbol{\lambda_1} h(\theta_i)\right] + \boldsymbol{\lambda_2} g(\theta_1, \ldots, \theta_K)$$

where h promotes differential sparsity in each model. 🔴

$\boxed{g}$ promotes common sparsity across all models. ●

Depends on the **assumption of model relations**





METHODOLOGY

We proposed three formulations, • Weighted non-convex Group norm penalty

CommonGrangerNet (CGN)

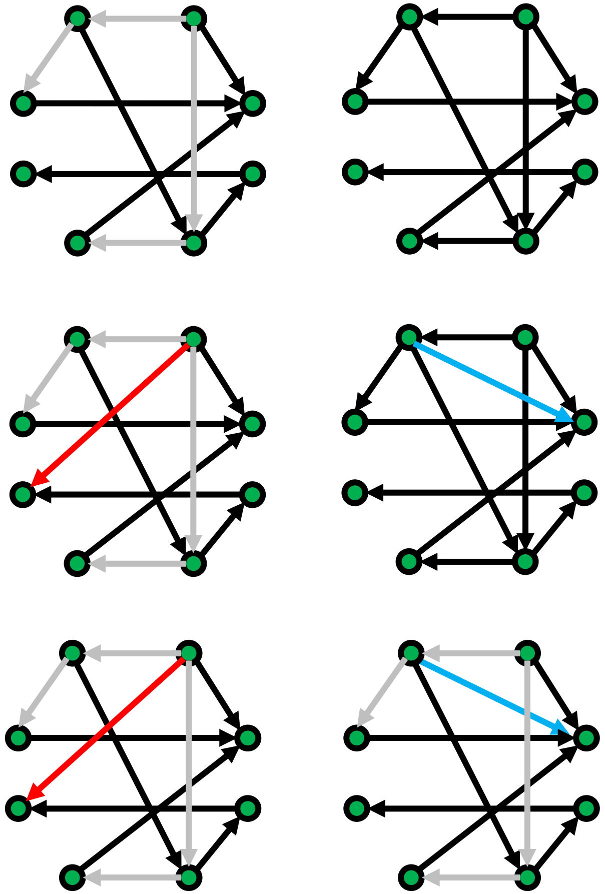
- **Common** network

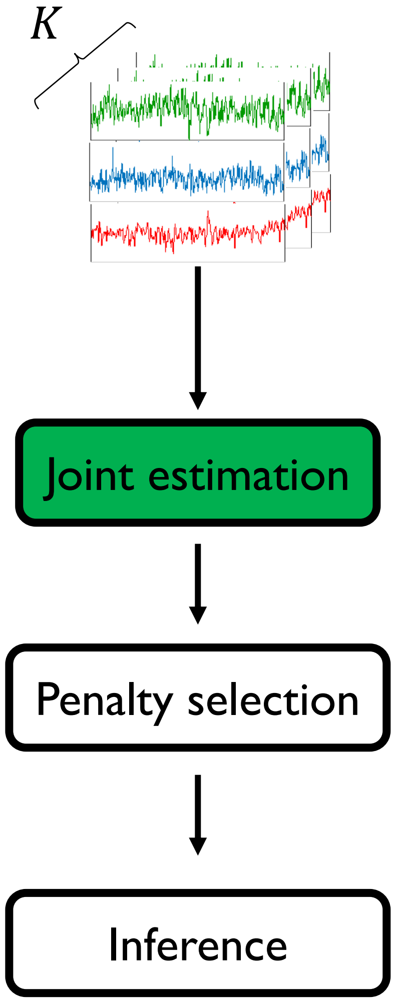
DifferentialGrangerNet (DGN)

- **Common** network
- **Differential** network

FusedGrangerNet (FGN)

- **Identical value common** network
- **Differential** network





Joint estimation

Penalty selection

Inference

METHODOLOGY

$$\min_{\theta_1,\dots,\theta_K} \sum_i \left[f_i(\theta_i) + \boldsymbol{\lambda_1} h(\theta_i)\right] + \boldsymbol{\lambda_2} g(\theta_1, \dots, \theta_K)$$

Unknown λ_1, λ_2 → vary λ_1, λ_2 for all combinations

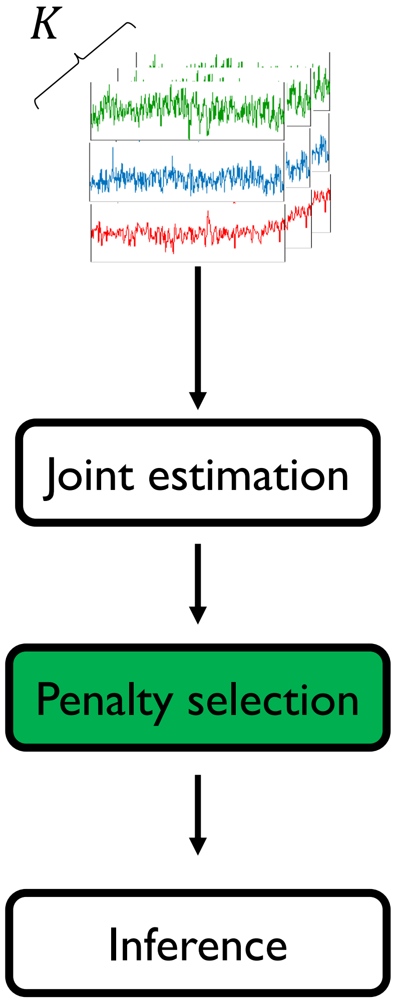
Find optimal λ_1, λ_2 by minimizing extended BIC

$$eBIC(\lambda_1, \lambda_2) = -2\,\mathcal{L}(\lambda_1, \lambda_2) + \boxed{\log(\mathrm{N}) \cdot \mathrm{df}(\lambda_1, \lambda_2) + 2\gamma \binom{n^2 pK}{\mathrm{df}(\lambda_1, \lambda_2)}}$$

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

Model complexity

Prior knowledge
on parameters space
with strength $0 \leq \gamma \leq 1$

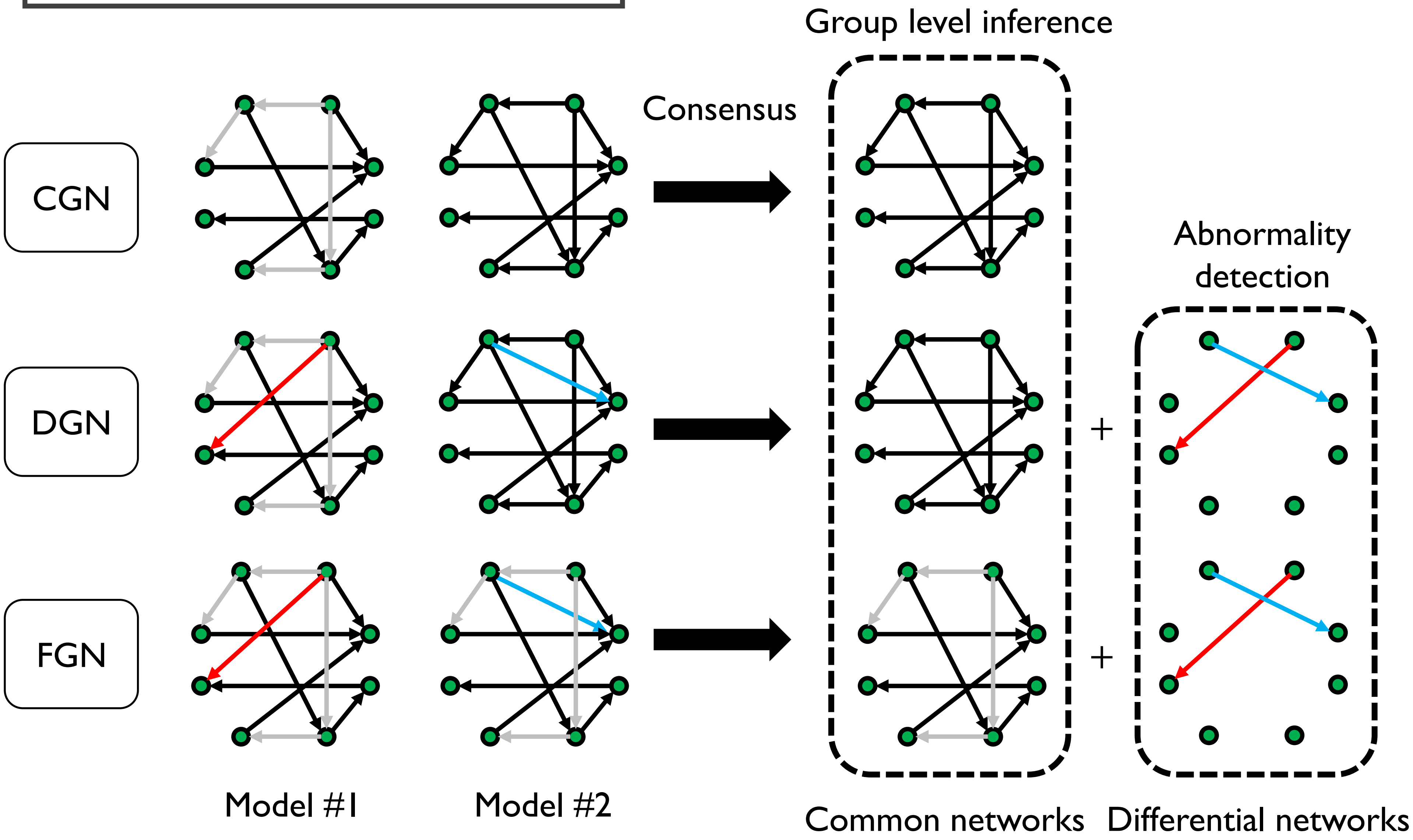


Joint estimation

Penalty selection

Inference

METHODOLOGY



CGN

DGN

FGN

Model #1

Model #2

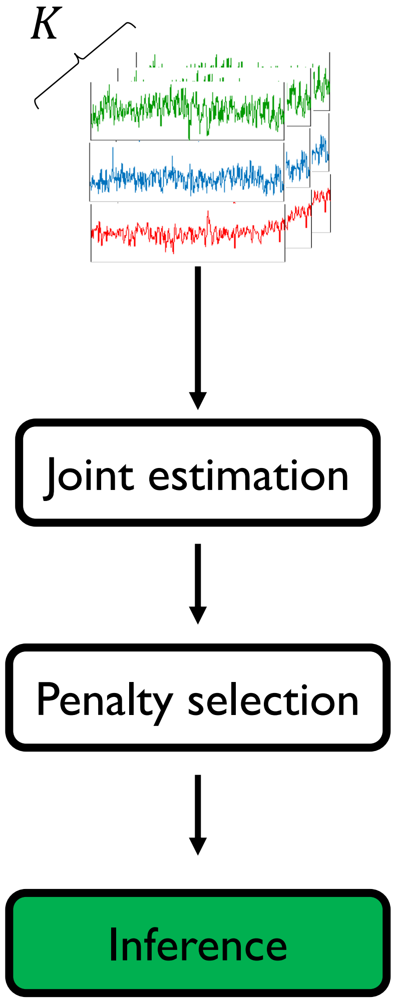
Consensus

Group level inference

Common networks

Abnormality detection

Differential networks



Joint estimation

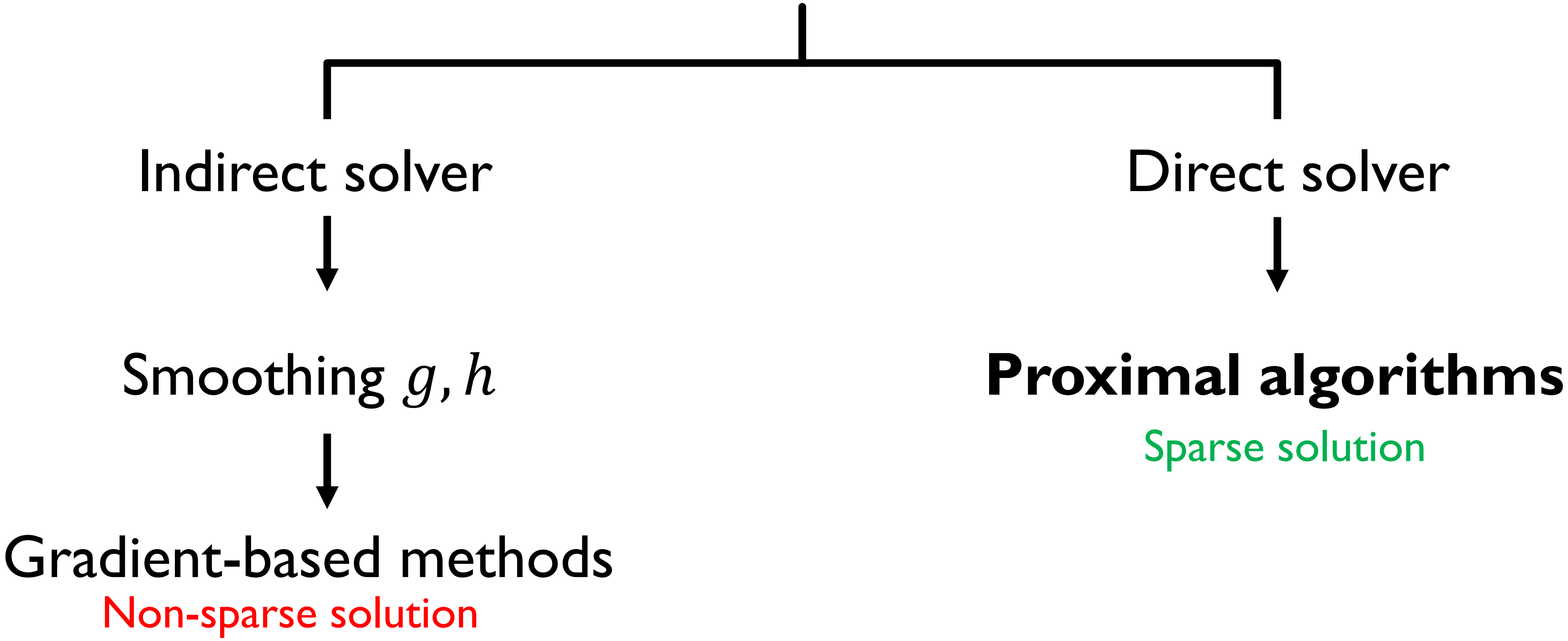
Penalty selection

Inference

ALGORITHMS

Proposed formulations in general form

- The problem is in the form of $\min_{x} f(x) + \overbrace{h_1(L_1 x) + h_2(L_2 x)}^{\overset{\text{VAR parameters}}{g(x)}}$
- ∇f is Lipschitz-continuous.
- Function g, h_i are possibly non-differentiable at the solution (zero)



Indirect solver

Smoothing g, h

Gradient-based methods

Non-sparse solution

Direct solver

Proximal algorithms

Sparse solution

ALGORITHMS

Available proximal algorithms to solve

$$\min_{x} f(x) + h_1(L_1 x) + h_2(L_2 x)$$

$$\min_{x} f(x) + g(x)$$

		Convex			Non-convex		
		Convergence guarantee					
Problem	Algorithm	CGN	DGN	FGN	CGN	DGN	FGN
$\min_{x} f(x) + g(x)$	Proximal gradient	✓	✓	✓	✓	✓	✓
	Accelerated proximal gradient (APG)	✓	✓	✓			
	Non-monotone APG [Li, 2015]	✓	✓	✓	✓	✓	✓
$\min_{x,z} f(x) + \tilde{g}(z)$ subjected to $_{Ax+Bz=c}$	ADMM with fixed penalty	✓	✓	✓			
	ADMM with spectral adaptive penalty [Xu, 2017]	✓	✓	✓			
	ADMM with heuristic adaptive penalty	✓	✓	✓	Converge in practice		

$$\min_{x,z} f(x) + \tilde{g}(z)$$

subjected to $_{Ax+Bz=c}$

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

$$\tilde{g}(z_1, z_2) = h_1(z_1) + h_2(z_2)$$

$$z = \begin{bmatrix} z_1 \\ z_2 \end{bmatrix}$$

ALGORITHMS

Available proximal algorithms to solve

$$\min_x f(x) + \boxed{h_1(L_1x) + h_2(L_2x)} \;—\; g$$

$\mathbf{prox}_{\alpha h_1}, \mathbf{prox}_{\alpha h_2}$ **have closed-form but not** $\mathbf{prox}_{\alpha g}$

Convergence guarantee

	Convex			Non-convex		
	CGN	DGN	FGN	CGN	DGN	FGN
Proximal gradient **(slow)**	✓	✓	✓	✓	✓	✓
Accelerated proximal gradient (APG)	✓	✓	✓			
Non-monotone APG	✓	✓	✓	✓	✓	✓

$$\min_x f(x) + g(x)$$

Proximal gradient $\quad x^+ = \mathrm{prox}_{\alpha g}(x - \alpha\nabla f(x))$

APG $\quad x^+ = \mathrm{prox}_{\alpha g}(\boldsymbol{y} - \boldsymbol{\alpha}\nabla f(\boldsymbol{y}))$

caching variables

Non-monotone APG

No closed form $\mathrm{prox}_{\alpha g}$

⬇

Numerical computation

$$\mathrm{prox}_{\alpha g}(v) = \operatorname*{argmin}_x g(x) + (1/\alpha)\|x - v\|_2^2$$

ALGORITHMS

Available proximal algorithms to solve

$$\min_{x} f(x) + \boxed{h_1(L_1 x) + h_2(L_2 x)} \text{ —— } g$$

$\mathbf{prox}_{\alpha h_1}, \mathbf{prox}_{\alpha h_2}$ **have closed-form but not** $\mathbf{prox}_{\alpha g}$

$$\min_{x,z} f(x) + \tilde{g}(z)$$

subjected to $_{Ax+Bz=c}$

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

$$\tilde{g}(z_1, z_2) = h_1(z_1) + h_2(z_2)$$

$$z = \begin{bmatrix} z_1 \\ z_2 \end{bmatrix}$$

	Convex			Non-convex		
	Convergence guarantee					
	CGN	DGN	FGN	CGN	DGN	FGN
ADMM with fixed penalty	✓	✓	✓			
ADMM with spectral adaptive penalty	✓	✓	✓			
ADMM with heuristic adaptive penalty	✓	✓	✓	Converge in practice		

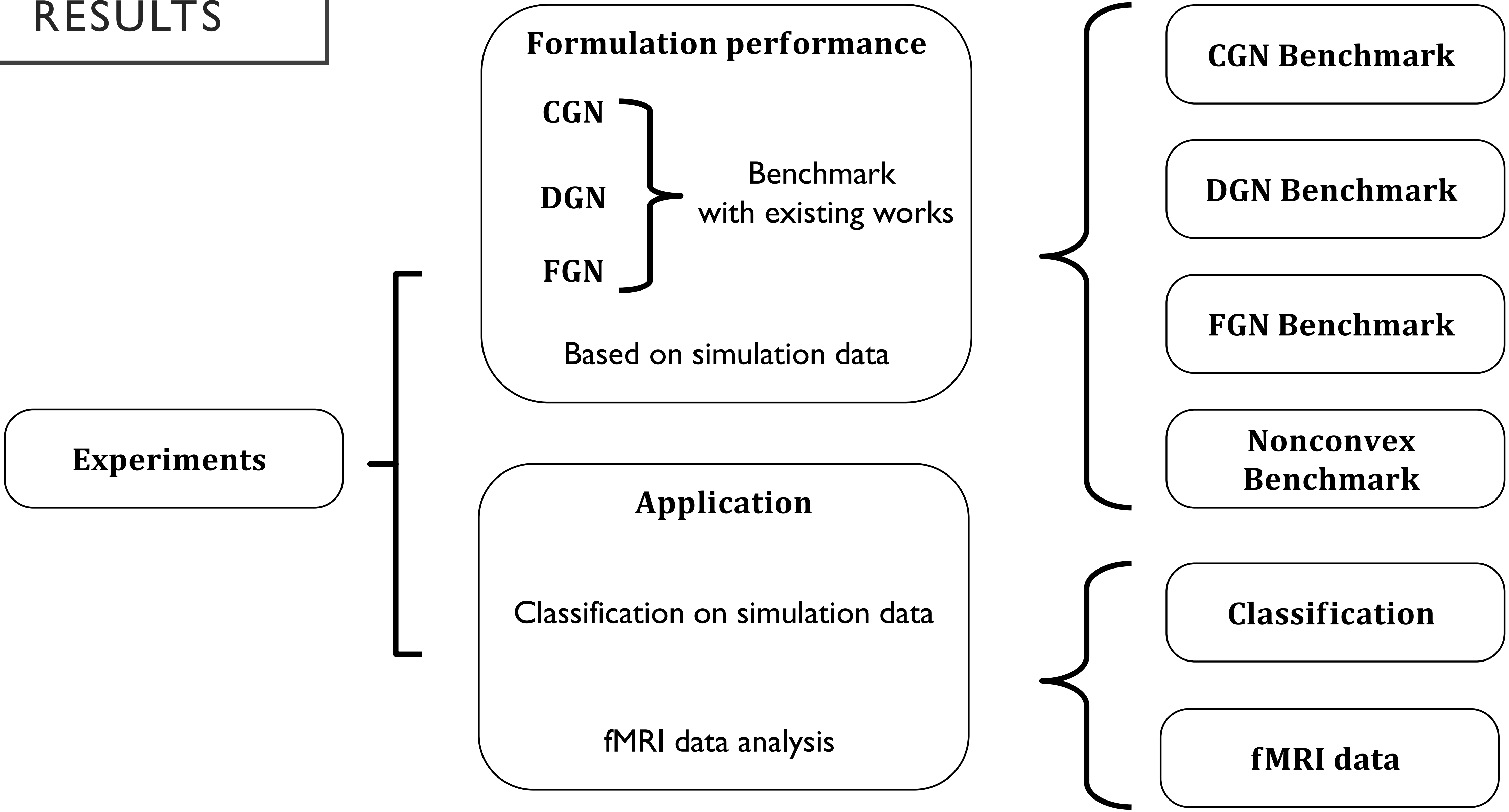
$$L_\rho(x, y, z) = f(x) + \tilde{g}(z) + y^T(c - Ax - Bz) + \frac{\rho}{2}\|c - Ax - Bz\|_2^2$$

[Xu, 2017]

$\rho^+ = \text{update}(\rho)$

- Spectral rule: Calculate from ADMM dual problem
- Heuristic rule: $\rho^+ = 2\rho$ Until primal residuals converged

RESULTS



Experiments

Formulation performance

CGN

DGN

FGN

Benchmark with existing works

Based on simulation data

Application

Classification on simulation data

fMRI data analysis

CGN Benchmark

DGN Benchmark

FGN Benchmark

Nonconvex Benchmark

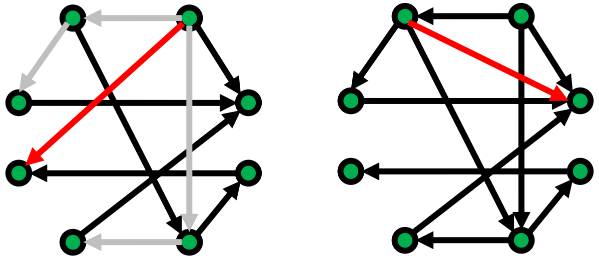
Classification

fMRI data

RESULTS

CGN BENCHMARK

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



Generated networks

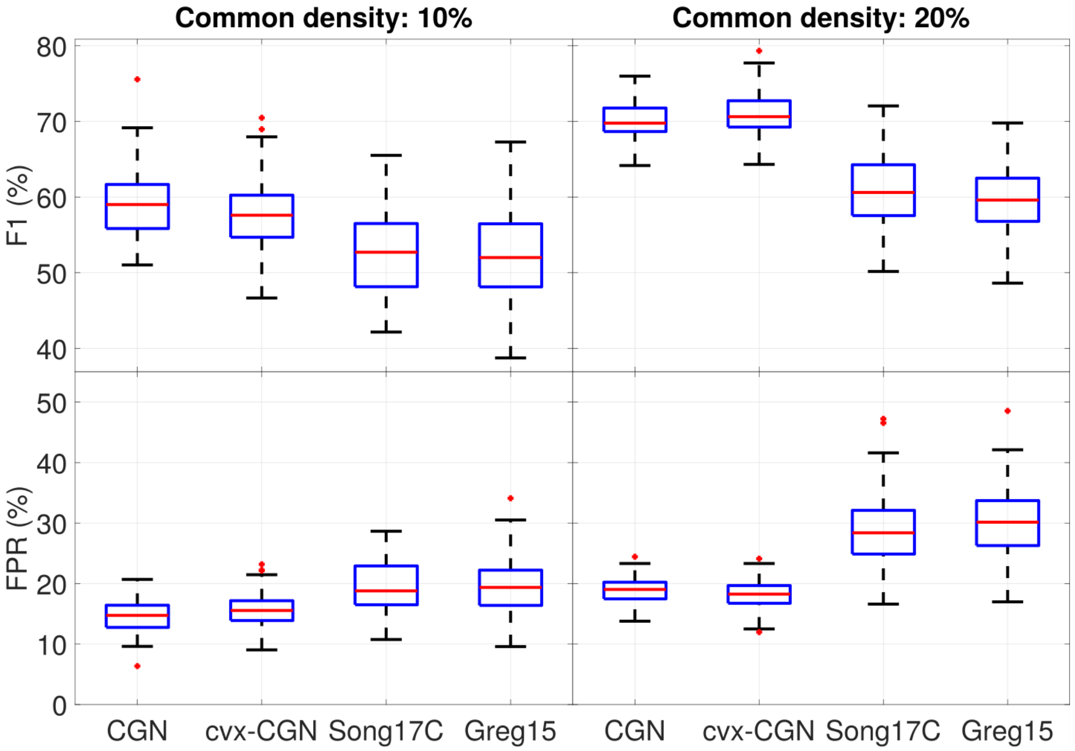
	all lags	non-convex penalty	weighted
[Gregorova, 2015]	✓		
[Songsiri, 2017]	✓		
cvx-CGN	✓		✓
CGN	✓	✓	✓

Problem parameters:

$n = 20, p = 1, K = 5$

Common density: **10%, 20%**

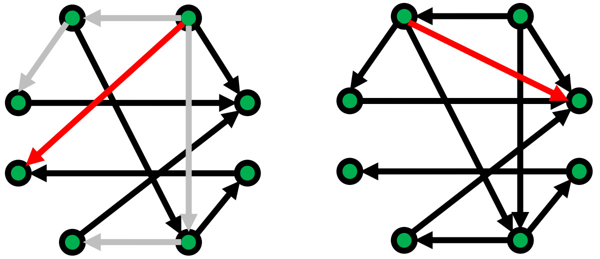
Differential density: 5%



- CGN and cvx-CGN had higher performance when density increased
- CGN and cvx-CGN had lowest FPR and highest F1 score median
- Song17C, Greg15 has similar performance

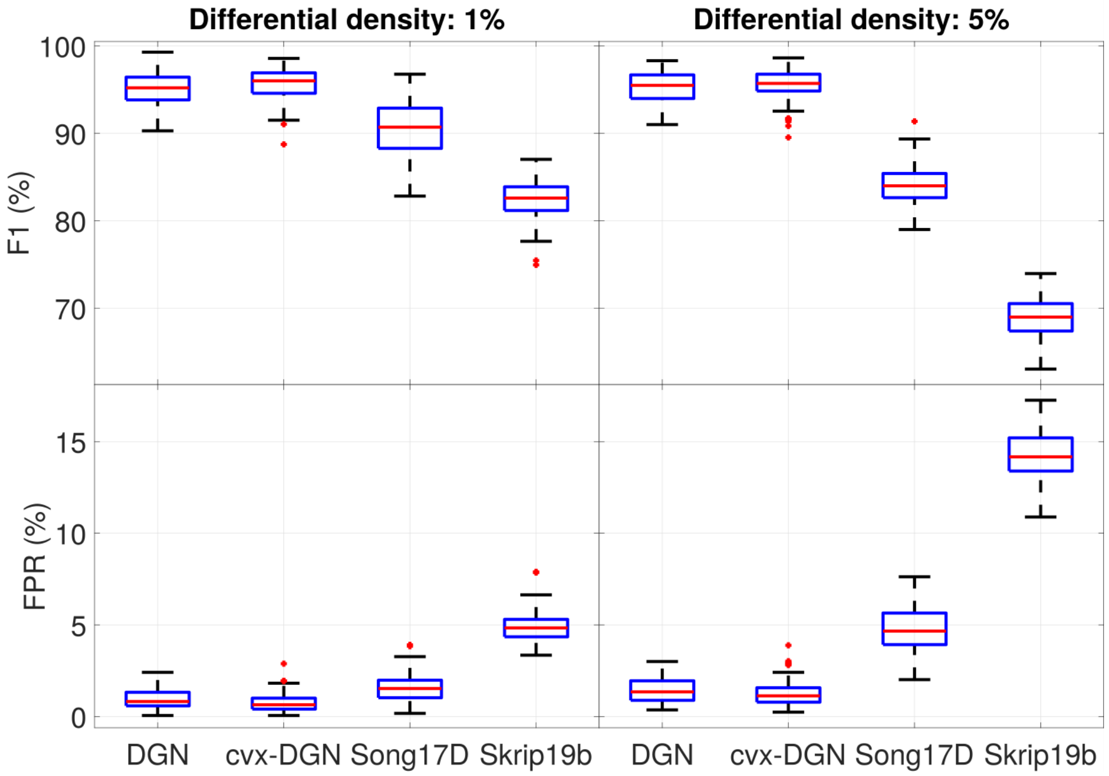
RESULTS | DGN BENCHMARK

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


Generated networks

	all lags	non-convex penalty	weighted
[Skripnikov, 2019b]	✓	(objective is non-convex)	
[Songsiri, 2017]	✓		
cvx-DGN	✓		✓
DGN	✓	✓	✓

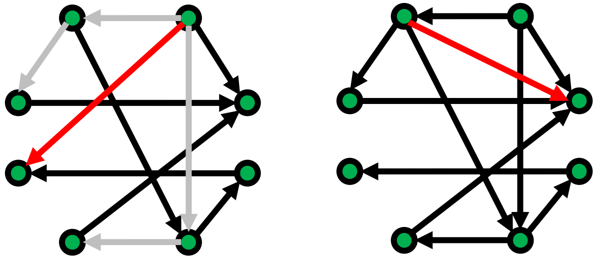
Problem parameters:
$n = 20, p = 1, K = \mathbf{5}, 50$
Common density: 10%
Differential density: **1%, 5%**



- Skrip19b is the most sensitive to the change in ground-truth density
- Almost all instances of proposed methods have higher F1 score than others in higher density setting
- Performance of the proposed methods did not degrade as differential density was increased

RESULTS | DGN BENCHMARK

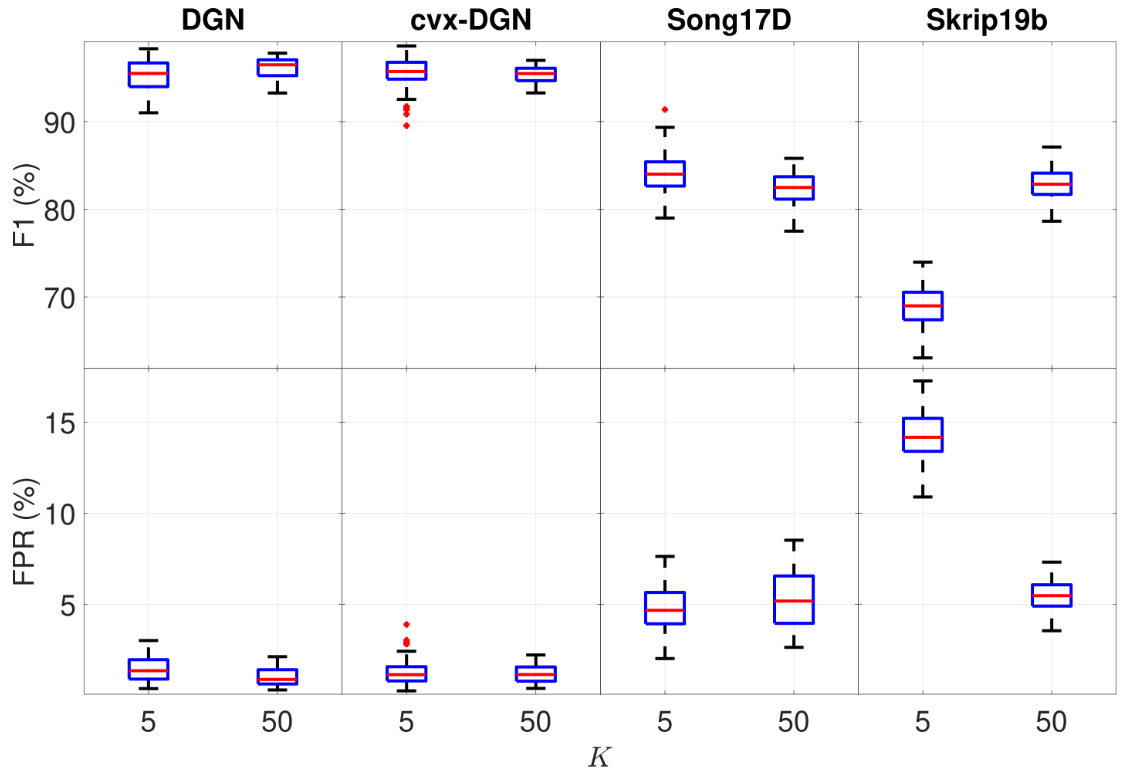
$$F_{ij} = 0 \Leftrightarrow (A_r)_{ij} = 0; r = 1,2,\dots p$$



Generated networks

	all lags	non-convex penalty	weighted
[Skripnikov, 2019b]	✓	(objective is non-convex)	
[Songsiri, 2017]	✓		
cvx-DGN	✓		✓
DGN	✓	✓	✓

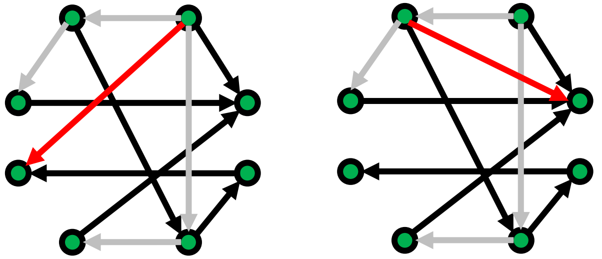
Problem parameters:
$n = 20, p = 1, K = \mathbf{5}, \mathbf{50}$
Common density: 10%
Differential density: 1%, **5%**



- Performance of DGN, cvx-DGN, Song17D was nearly the same as number of models increased
- Skrip19b has significant improvement as the number of models increased
- Almost all instances of DGN, cvx-DGN have higher F1 score than others

RESULTS | FGN BENCHMARK

$$F_{ij} = 0 \Leftrightarrow (A_r)_{ij} = 0; r = 1,2, \ldots p$$



Generated networks

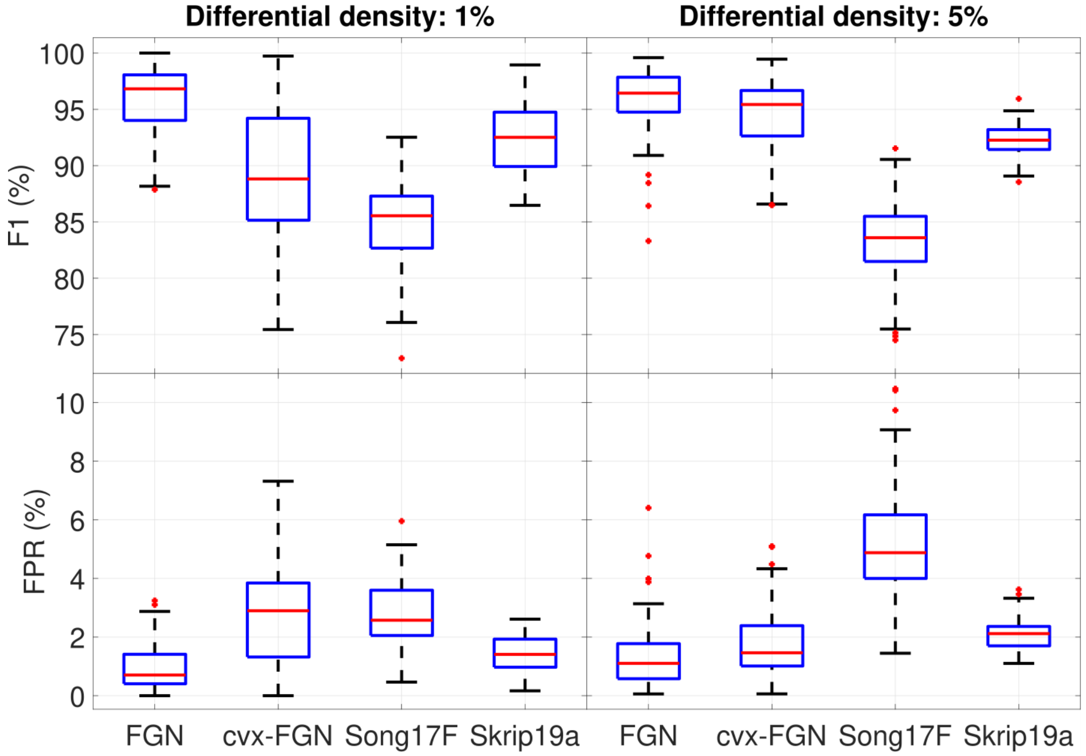
	all lags	non-convex penalty	weighted
[Skripnikov, 2019a]			
[Songsiri, 2015]	✓		
cvx-FGN	✓		✓
FGN	✓	✓	✓

Problem parameters:

$n = 20, p = 1, K = 5$

Common density: 10%

Differential density: **1%, 5%**



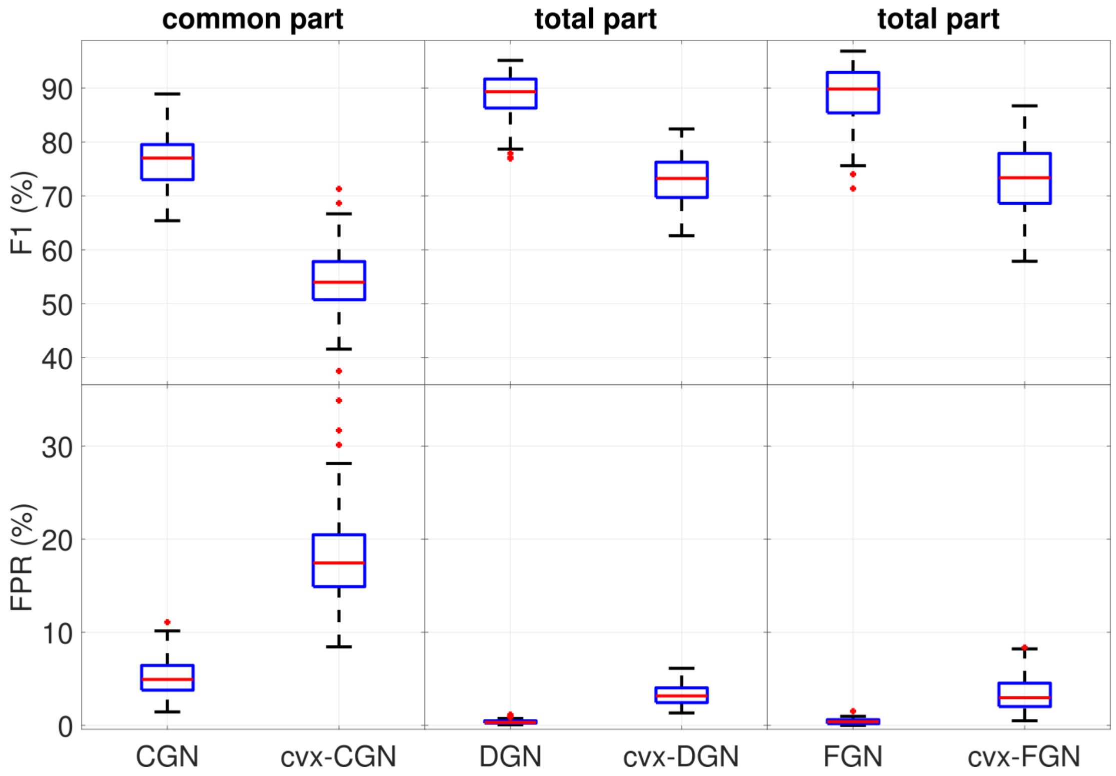
- Performance of FGN did not degrade as differential density was increased
- cvx-FGN has wide range of results in 1% setting
- Song17F has lower performance than Skrip19a

RESULTS | NON-CONVEX VS CONVEX

#model parameters : timepoints

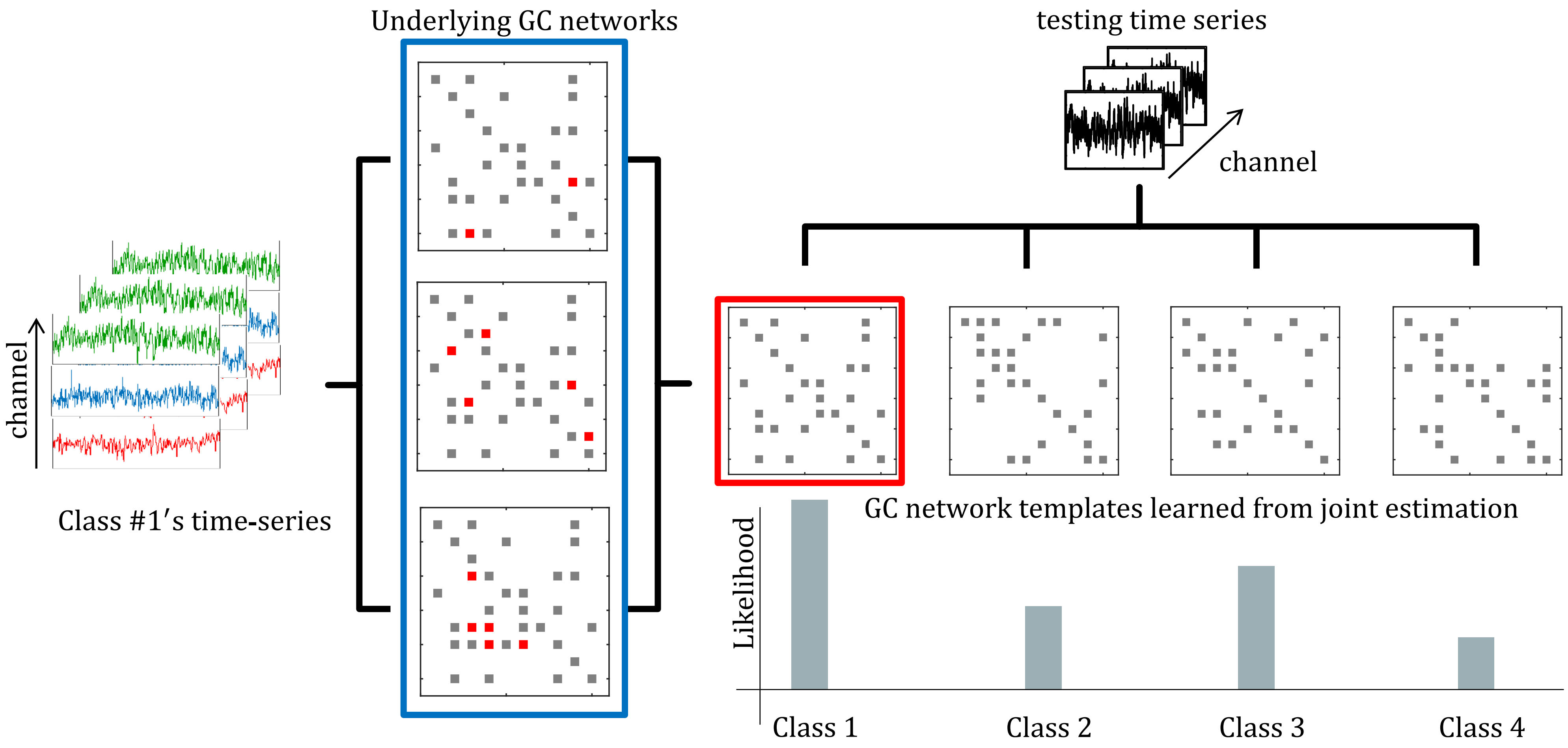
4 : 1 → **8 : 1**

Problem parameters:
$n = 20, p = 3, K = 5$
Common density: 10%
Differential density: 5%

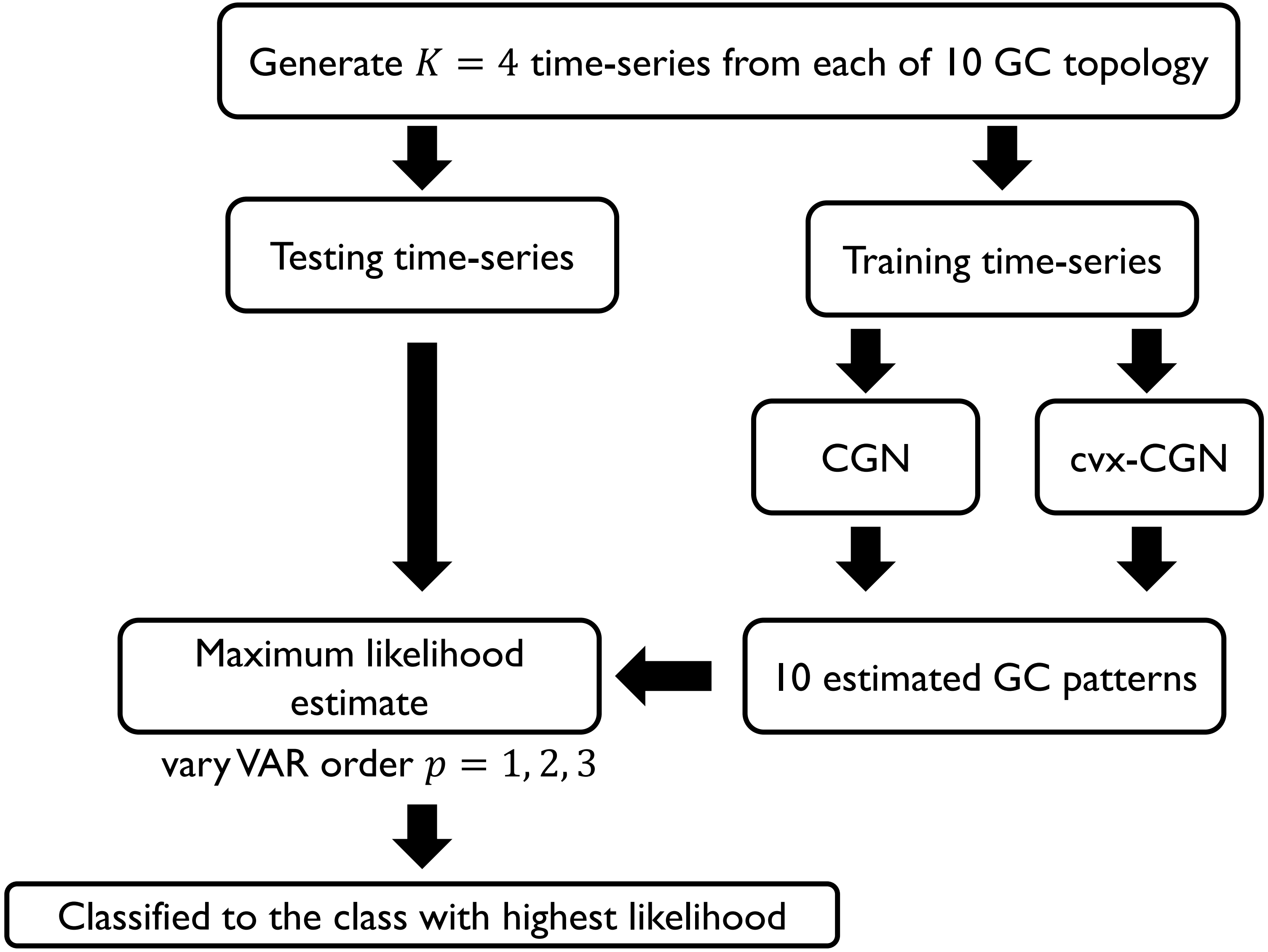


- All non-convex formulations significantly outperformed their convex relaxations
- Directly supported by theoretical sparsity recovery property
- Implication
  - Convex formulations can still be used
  if the number of time-points is sufficiently high.

Classification scheme: Likelihood ratio test



RESULTS | APPLICATION | CLASSIFICATION



Generate $K = 4$ time-series from each of 10 GC topology

Testing time-series

Training time-series

CGN

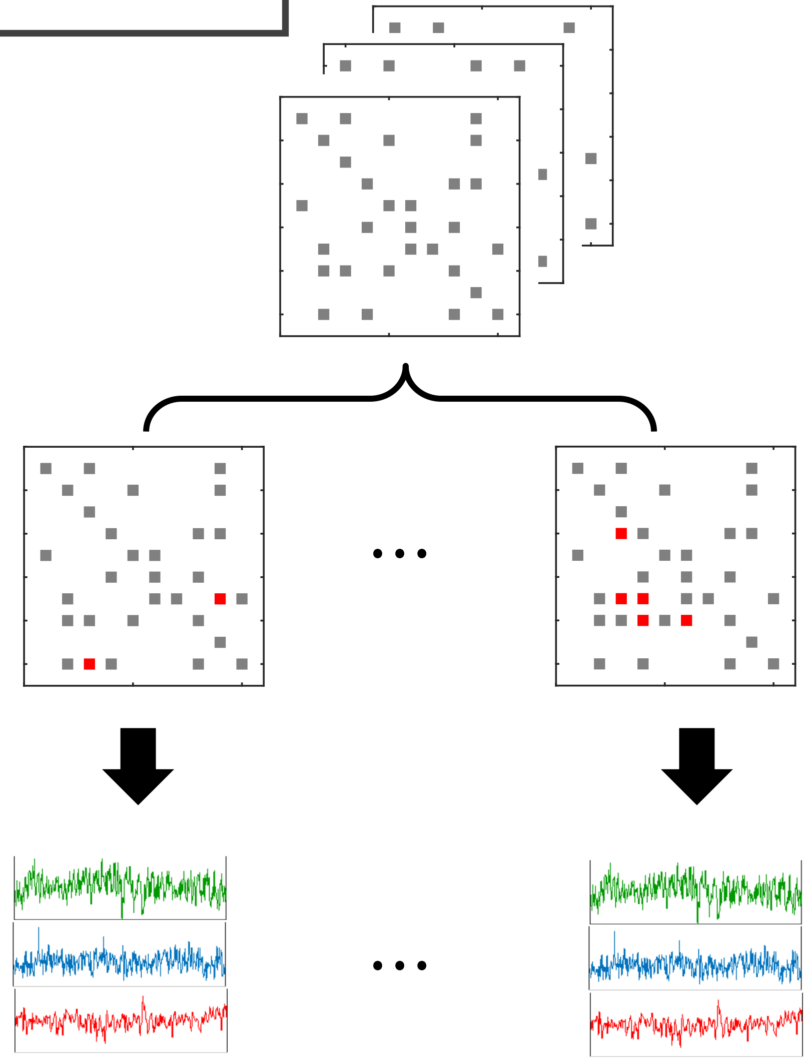
cvx-CGN

10 estimated GC patterns

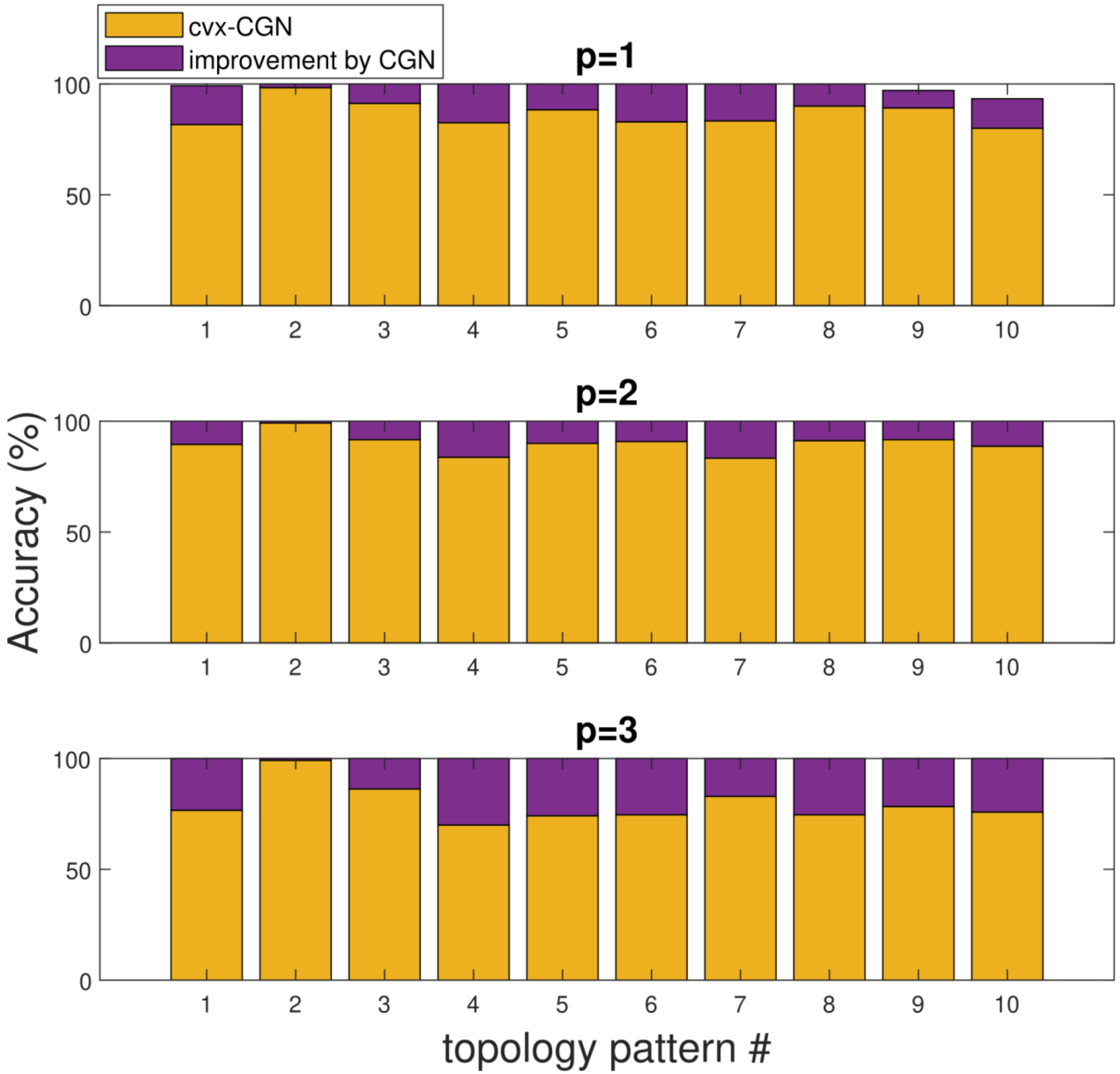
Maximum likelihood estimate

vary VAR order $p = 1, 2, 3$

Classified to the class with highest likelihood



RESULTS | APPLICATION | CLASSIFICATION



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

RESULTS | APPLICATION | ADHD-200

ADHD (Attention deficit hyperactivity disorder)

- ADHD is characterized by the inattention, hyperactivity, poor impulse control and emotion processing

- These characteristics can be explained by using a causality analysis tool to reveal the causal interconnections between brain regions or brain sub-networks

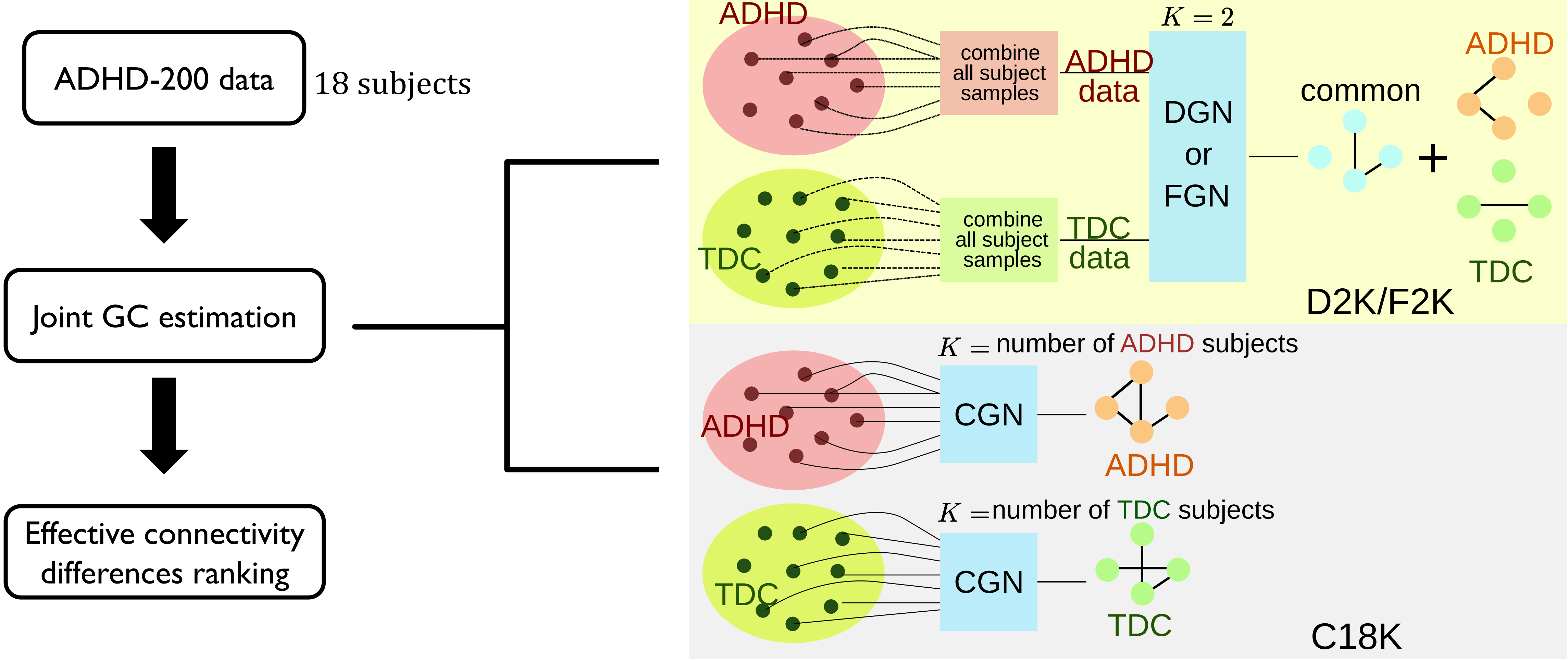
⬇

Necessary to find **group level** brain network differences between children with ADHD and the typically developed children (TDC) to make a better understanding of the disease

⬇

Joint estimation of effective brain connectivity

Brain network differences learning process


ADHD-200 data
18 subjects
Joint GC estimation
Effective connectivity differences ranking
ADHD
combine all subject samples
ADHD data
$K = 2$
DGN or FGN
common
+
ADHD
TDC
combine all subject samples
TDC data
TDC
D2K/F2K
$K =$ number of ADHD subjects
ADHD
CGN
ADHD
$K =$ number of TDC subjects
TDC
CGN
TDC
C18K

Learning paradigm

Results summary

- Most extra/missing links take place in the **orbitofrontal regions** and **limbic system**
- The functions of both orbitofrontal regions and limbic systems are known to be related with reward learning system, emotion processing and the process involved with the memory
- These results are consistent with the findings in ADHD literature from both functional connectivity studies, clinical studies

CONCLUSION

- We extended joint Granger graphical model estimation in three folds by using group penalty, non-convex penalty and weighted penalty
- We demonstrated the effectiveness of proposed methods by benchmarking with other works with intensive simulation experiments
- Our methods outperformed the other literature with the same prior information assumptions on the relations among all models
- We applied all formulations to reveal the effective brain connectivity differences between ADHD and TDC and the results were consistent with previously reported literature in both clinical studies and the studies with data-driven methods

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.

[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.

[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.

[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.

[Skrip19a] A. Skripnikov and G. Michailidis, "Joint estimation of multiple network Granger causal models," Econometrics and Statistics, vol. 10, pp. 120–133, 2019.

REFERENCES

[Skrip19b] A. Skripnikov and G. Michailidis, "Regularized joint estimation of related vector autoregressive models," Computational Statistics & Data Analysis, vol. 139, pp. 164–177, 2019.

[Songsiri 15] Songsiri, J. 2015. Learning multiple Granger graphical models via group fused lasso. In Proceedings of the IEEE 10th Asian Control Conference (ASCC).

[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.

[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.