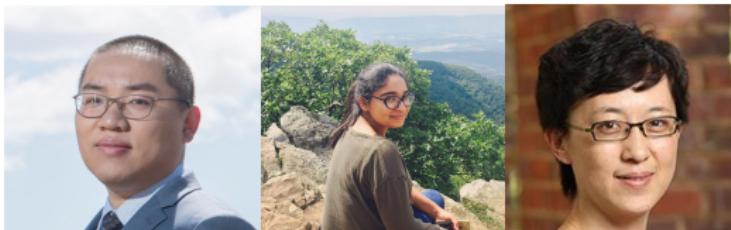


A Fast and Scalable Joint Estimator for Integrating Additional Knowledge in Learning Multiple Related Sparse Gaussian Graphical Models

Beilun Wang¹ Arshdeep Sekhon¹ Yanjun Qi¹

¹Department of Computer Science, University of Virginia
<http://jointggm.org/>

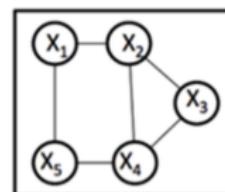
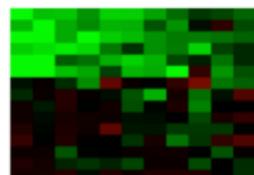
Published @ ICML18;
July 2018



Motivation: Learning Multiple Related Graphs from Heterogeneous Samples about Multiple Contexts

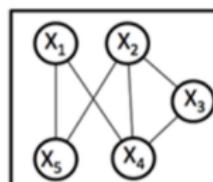
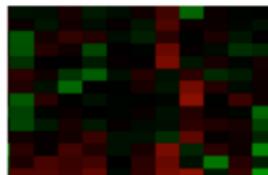
- Multiple Datasets $\mathbf{X}^{(1)}, \dots, \mathbf{X}^{(K)}$ → Multiple Graphs $G^{(1)}, \dots, G^{(K)}$.

Context/Task(1)



Infer

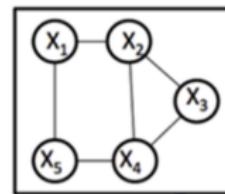
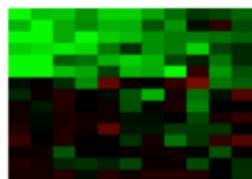
Context/Task(2)



Motivation: Learning Multiple Related Graphs from Heterogeneous Samples about Multiple Contexts

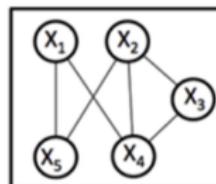
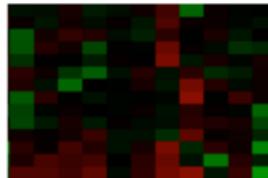
- Multiple Datasets $\mathbf{X}^{(1)}, \dots, \mathbf{X}^{(K)}$ → Multiple Graphs $G^{(1)}, \dots, G^{(K)}$.

Context/Task(1)



Infer

Context/Task(2)

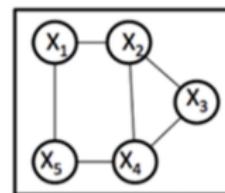
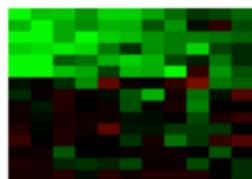


- e.g., Brain Connectomes from heterogeneous fMRI images
- e.g., Genetic Networks from heterogeneous RNA samples

Motivation: Learning Multiple Related Graphs from Heterogeneous Samples about Multiple Contexts

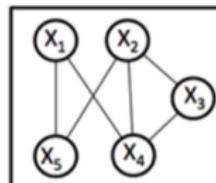
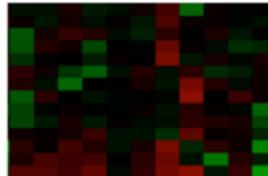
- Multiple Datasets $\mathbf{X}^{(1)}, \dots, \mathbf{X}^{(K)}$ → Multiple Graphs $G^{(1)}, \dots, G^{(K)}$.

Context/Task(1)



Infer

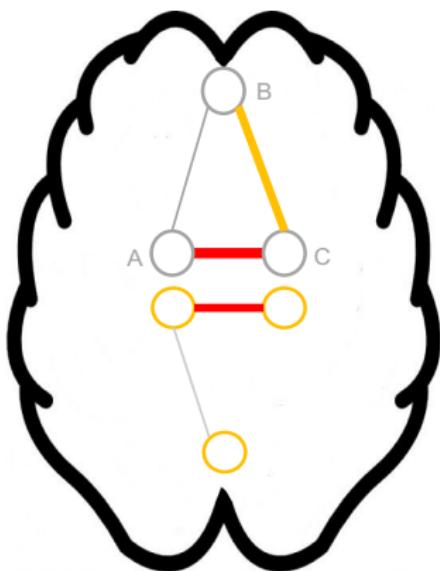
Context/Task(2)



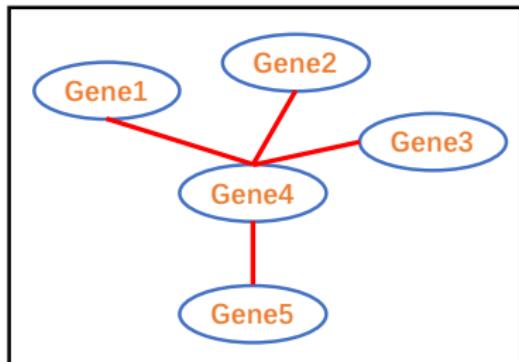
- Current Approach: Multi-sGGMs ⇒ Multi-task sparse Gaussian Graphical Models

Limitation I: Missing Known Knowledge

- No clear ways to consider **Known Additional Knowledge** in multi-sGGMs.
- However, in real-world applications, plenty of known information. (e.g., **red edges** in the figures.)



Spatial Knowledge



Genetic Pathway

Limitation II: Slow computation and Not scalable

- K : number of tasks
 p : number of features

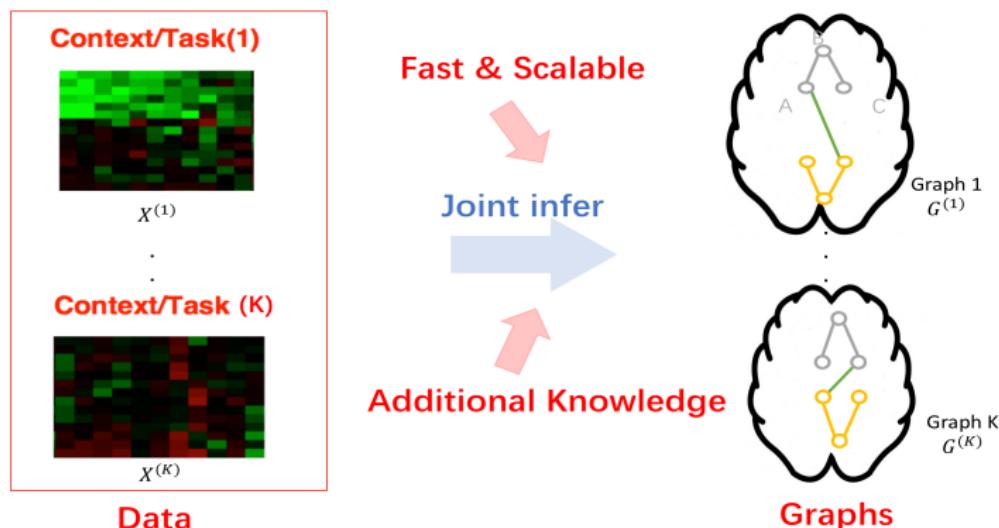
- | Method | Time Complexity | Bottleneck |
|-----------------------|--------------------|------------------------|
| W-SIMULE ¹ | $O(K^4 p^5)$ | LP with Kp variables |
| JGL ² | $O(T \times Kp^3)$ | SVD |

¹[Singh et al.(2017) Singh, Wang, and Qi]

²[Danaher et al.(2013) Danaher, Wang, and Witten]

Our Aim: Integrating Additional Knowledge in Scalable Learning of multi-sGGMs

- Our focus: How to estimate multiple graphs $G^{(1)}, \dots, G^{(K)}$ from heterogeneous data $\mathbf{X}^{(1)}, \dots, \mathbf{X}^{(K)}$ and integrate additional knowledge.



Notations

$X^{(i)}$ i -th Data matrix.

$\Sigma^{(i)}$ i -th Covariance matrix.

$\Omega^{(i)}$ i -th Inverse of covariance matrix (precision matrix).

p The total number of feature variables.

n_{tot} The total number of samples.

X^{tot} the concatenation of all Data matrices.

Σ^{tot} the concatenation of all Covariance matrices.

Ω^{tot} the concatenation of all Inverse of covariance matrices (precision matrices).

W_I^{tot} $(W_I^{(1)}, W_I^{(2)}, \dots, W_I^{(K)})$

W_S^{tot} (W_S, W_S, \dots, W_S)

Background: Elementary Estimator for sGGM

Elementary Estimator

$$\operatorname{argmin}_{\theta} \mathcal{R}(\theta) \quad (1.1)$$

$$\text{Subject to: } \mathcal{R}^*(\theta - \mathcal{B}^*(\hat{\phi})) \leq \lambda_n$$

- For example, for sGGM:

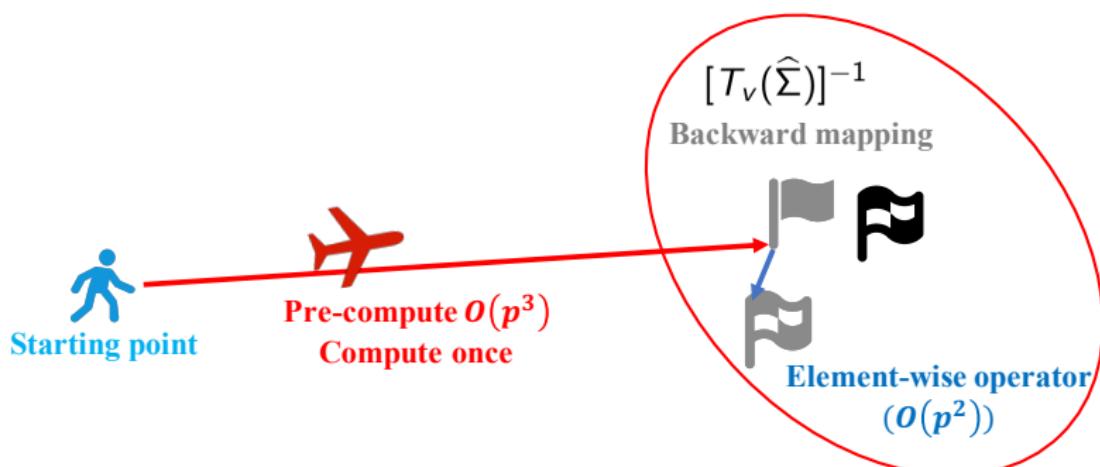
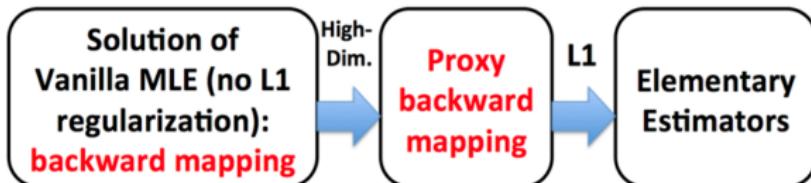
EE	$\mathcal{R}(\cdot)$	θ	\mathcal{B}^*	\mathcal{R}^*
EE-sGGM	$\ \cdot\ _1$	Ω	$[T_v(\hat{\Sigma})]^{-1}$	$\ \cdot\ _1$

Elementary Estimator for sGGM

$$\operatorname{argmin}_{\Omega} \|\Omega\|_1 \quad (1.2)$$

$$\text{Subject to: } \|\Omega - [T_v(\hat{\Sigma})]^{-1}\|_{\infty} \leq \lambda_n$$

Background: Elementary Estimator – visualization



- | single-sGGM Method | Time Complexity | Note |
|----------------------|-----------------|---|
| QUIC ³ | $O(p^3)$ | Not a closed-form solution |
| EE-sGGM ⁴ | $O(p^2)$ | entry-wise
closed-form
sharp convergence rate |

- | multi-sGGM Method | Time Complexity | Notes |
|-----------------------|--------------------|--|
| W-SIMULE ⁵ | $O(K^4 p^5)$ | LP with Kp variables |
| JGL ⁶ | $O(T \times Kp^3)$ | SVD |
| Proposed method | $O(K^4 p^2)$ | entry-wise
fast
sharp convergence rate |

³[Hsieh et al.(2011)Hsieh, Sustik, Dhillon, and Ravikumar]

⁴[Yang et al.(2014)Yang, Lozano, and Ravikumar]

⁵[Singh et al.(2017)Singh, Wang, and Qi]

⁶[Danaher et al.(2013)Danaher, Wang, and Witten]

Proposed: Using Knowledge as Weight in Regularization (KW-norm)

- Integrating additional knowledge through a novel regularization function $\mathcal{R}(\cdot)$

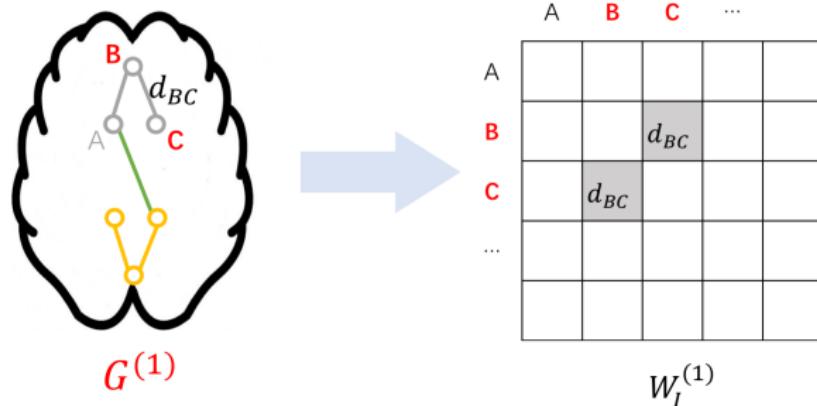
KW-norm

$$\mathcal{R}(\Omega^{tot}) = \|W_I^{tot} \circ \Omega_I^{tot}\|_1 + \|W_S^{tot} \circ \Omega_S^{tot}\|_1 \quad (2.1)$$

- W_I^{tot} : weights describing knowledge of each individual graph.
- W_S^{tot} : weights describing knowledge of the shared graph.
- KW-norm is **flexible**.

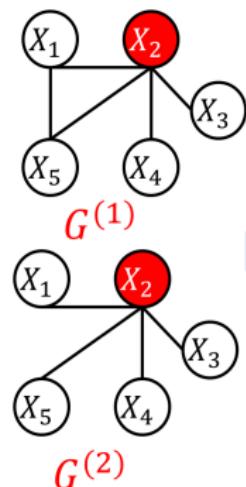
Example I: KW-norm representing the edge-level knowledge

- e.g., Spatial distance among brain regions;



Example II: KW-norm describing the node-level knowledge

- e.g., X_2 is a known hub node;



	1	2	3	4	5
1		$1/\gamma$	1	1	1
2	$1/\gamma$		$1/\gamma$	$1/\gamma$	$1/\gamma$
3	1	$1/\gamma$		1	1
4	1	$1/\gamma$	1		1
5	1	$1/\gamma$	1	1	

W_s

Proposed Method: Joint Elementary Estimator incorporating additional Knowledge (JEEK)

EE	$\mathcal{R}(\cdot)$	θ	\mathcal{B}^*	\mathcal{R}^*
EE-sGGM	$\ \cdot\ _1$	Ω	$[T_v(\widehat{\Sigma})]^{-1}$	$\ \cdot\ _\infty$
JEEK	kw-norm	Ω^{tot}	$inv[T_v(\widehat{\Sigma}^{tot})]$	kw-dual

JEEK

$$\underset{\Omega_I^{tot}, \Omega_S^{tot}}{\operatorname{argmin}} \|W_I^{tot} \circ \Omega_I^{tot}\|_1 + \|W_S^{tot} \circ \Omega_S^{tot}\|$$

Subject to: $\|W_I^{tot} \circ (\Omega^{tot} - inv(T_v(\widehat{\Sigma}^{tot})))\|_\infty \leq \lambda_n \quad (2.2)$

$$\|W_S^{tot} \circ (\Omega^{tot} - inv(T_v(\widehat{\Sigma}^{tot})))\|_\infty \leq \lambda_n$$

$$\Omega^{tot} = \Omega_S^{tot} + \Omega_I^{tot}$$

Proposed method: JEEK – Solution

- Fast and Scalable solution⁷ – p^2 small linear programming subproblems:

$$\operatorname{argmin}_{a_i, b} \sum_i |w_i a_i| + K |w_s b|$$

$$\begin{aligned} \text{Subject to: } & |a_i + b - c_i| \leq \frac{\lambda_n}{\min(w_i, w_s)}, \\ & i = 1, \dots, K \end{aligned} \tag{2.3}$$

⁷ $a_i := \Omega_{I,j,k}^{(i)}$ (the $\{j, k\}$ -th entry of $\Omega^{(i)}$)

$b := \Omega_{Sj,k}$

$c_i = [T_v(\widehat{\Sigma}^{(i)})]_{j,k}^{-1}$.

$W_{j,k}^{(i)} = w_i$ and $W_{j,k}^S = w_s$.

Why JEEK is better

- Rich and flexible for integrating additional knowledge
 - e.g., spatial, anatomy, hub, pathway, location, known edges;

Why JEEK is better

- Rich and flexible for integrating additional knowledge
 - e.g., spatial, anatomy, hub, pathway, location, known edges;
- Parallelizable optimization with small sub-problems. Faster than the previous studies:

Method	Time Complexity	Additional Knowledge
JECK	$O(K^4 p^2)$ ($\Rightarrow O(K^4)$ if parallelizing completely)	YES
W-SIMULE	$O(K^4 p^5)$	YES
JGL	$O(T \times Kp^3)$	NO

Why JEEK is better

- Rich and flexible for integrating additional knowledge
 - e.g., spatial, anatomy, hub, pathway, location, known edges;
- Parallelizable optimization with small sub-problems. Faster than the previous studies:

Method	Time Complexity	Additional Knowledge
JECK	$O(K^4 p^2)$ ($\Rightarrow O(K^4)$ if parallelizing completely)	YES
W-SIMULE	$O(K^4 p^5)$	YES
JGL	$O(T \times Kp^3)$	NO

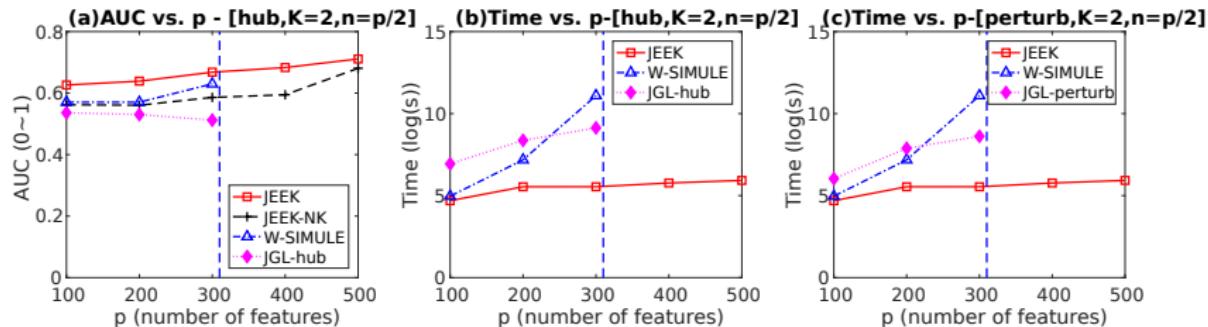
- Theoretical guaranteed

Theoretical Results

- Error bound: $\|\Delta^* - \widehat{\Delta}\|$
- Sharp convergence rate as the state-of-art

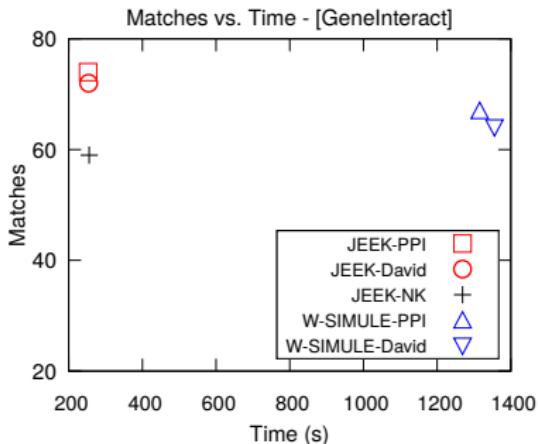
$$\begin{aligned} \|\widehat{\Omega}^{tot} - \Omega^{tot*}\|_F &\leq 4\sqrt{k_i + k_s}\lambda_n \\ \max(\|W_I^{tot} \circ (\widehat{\Omega}^{tot} - \Omega^{tot*})\|_\infty, \|W_S^{tot} \circ (\widehat{\Omega}^{tot} - \Omega^{tot*})\|_\infty) &\leq 2\lambda_n \quad (3.1) \\ \|W_I^{tot} \circ (\widehat{\Omega}_I^{tot} - \Omega_I^{tot*})\|_1 + \|W_S^{tot} \circ (\widehat{\Omega}_S^{tot} - \Omega_S^{tot*})\|_1 &\leq 8(k_i + k_s)\lambda_n \end{aligned}$$

Empirical Results on Multiple Synthetic Datasets

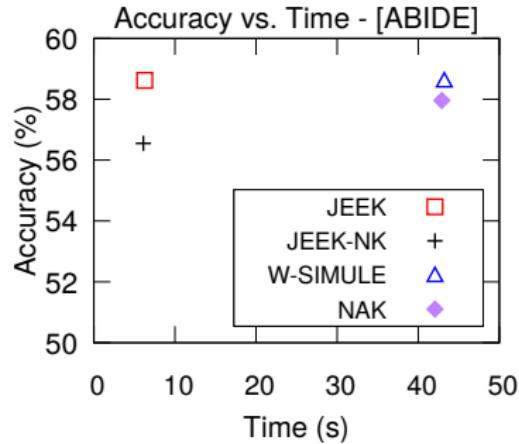


- **JEEK** outperforms the speed of the state-of arts significantly;
- **JEEK** obtains better or same AUC as the state-of-the-art;
- **JEEK** obtains better AUC than JEEK-NK (no additional knowledge).

Empirical Results on Two Real-world Datasets



(a)



(b)

- (a). On real-world gene expression data about leukemia cells vs. normal blood cells. Used multiple types of additional knowledge;
- (b). On real-world Brain fMRI dataset: ABIDE. Using LDA as a downstream classification for evaluating JEEK vs. baselines.

R Package Publicly Available !!!

- The project website: <http://jointggm.org/>
- R package "jeek":
 - `install.packages("jeek")`
 - `library("jeek")`
 - `demo(jeek)`

References

-  P. Danaher, P. Wang, and D. M. Witten.
The joint graphical lasso for inverse covariance estimation across multiple classes.
Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2013.
-  C.-J. Hsieh, M. A. Sustik, I. S. Dhillon, and P. D. Ravikumar.
Sparse inverse covariance matrix estimation using quadratic approximation.
In *NIPS*, pages 2330–2338, 2011.
-  C. Singh, B. Wang, and Y. Qi.
A constrained, weighted-l1 minimization approach for joint discovery of heterogeneous neural connectivity graphs.
arXiv preprint arXiv:1709.04090, 2017.
-  E. Yang, A. C. Lozano, and P. Ravikumar.
Elementary estimators for sparse covariance matrices and other structured moments.
In *ICML*, pages 397–405, 2014.