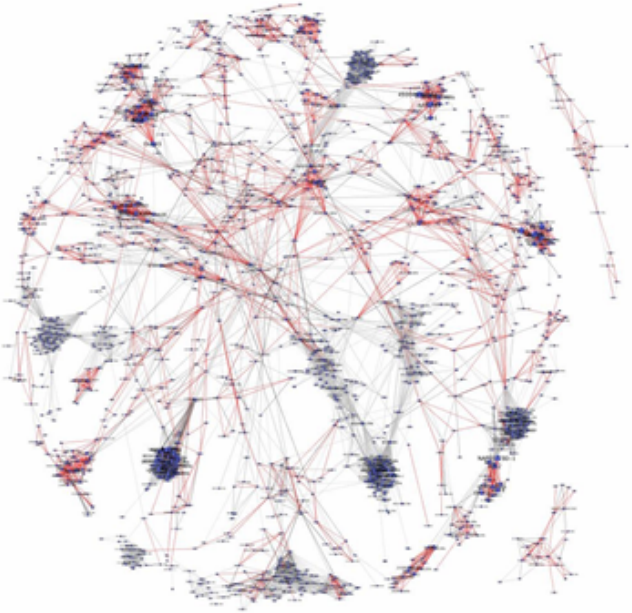
A **Fast and Scalable** Joint Estimator for Integrating **Additional Knowledge** in Learning Multiple Related Sparse Gaussian Graphical Models

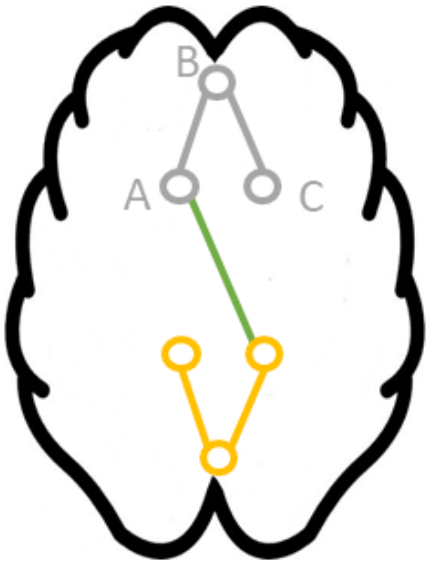
Beilun Wang[1] Arshdeep Sekhon[1] Yanjun Qi[1]

[1]Department of Computer Science, University of Virginia
http://jointggm.org/

Published @ ICML18;
July 2018

Motivation: Entity Graph

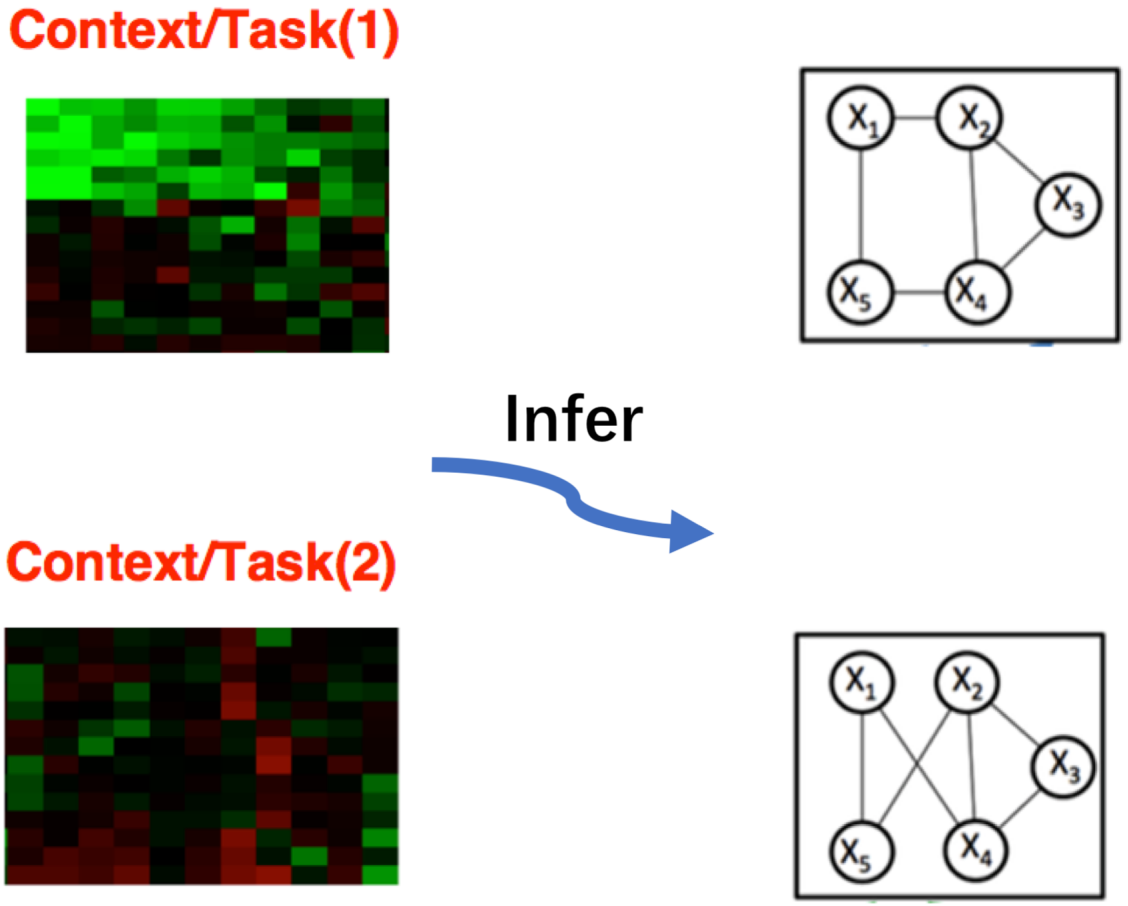




- Many applications need to know interactions among entities:
  - Gene Interactions
  - Brain connectivity
- Why to study the entity graph
  - Understanding
  - Diagnosis, e.g., marker
  - Treatment, e.g., drug development.

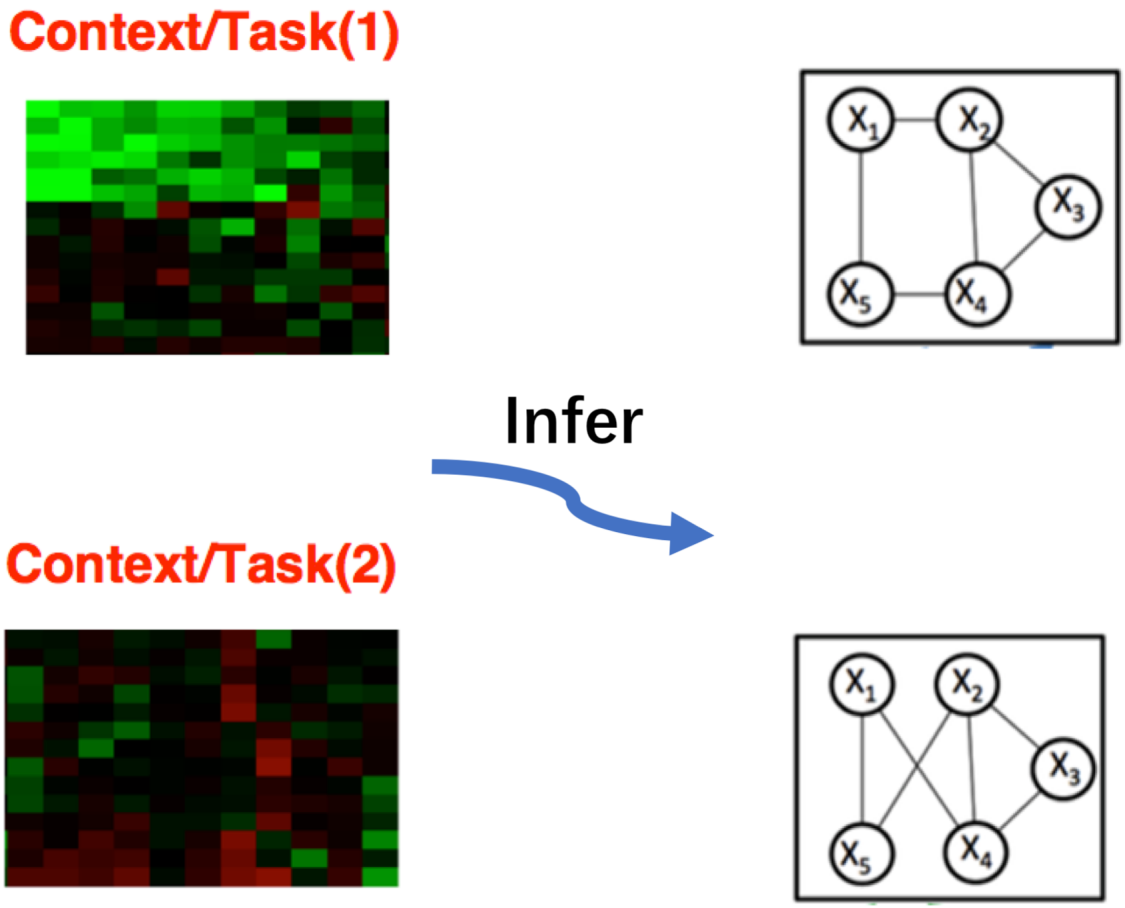
Motivation: Learning Multiple Related Graphs from Heterogeneous Samples about Multiple Contexts

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



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

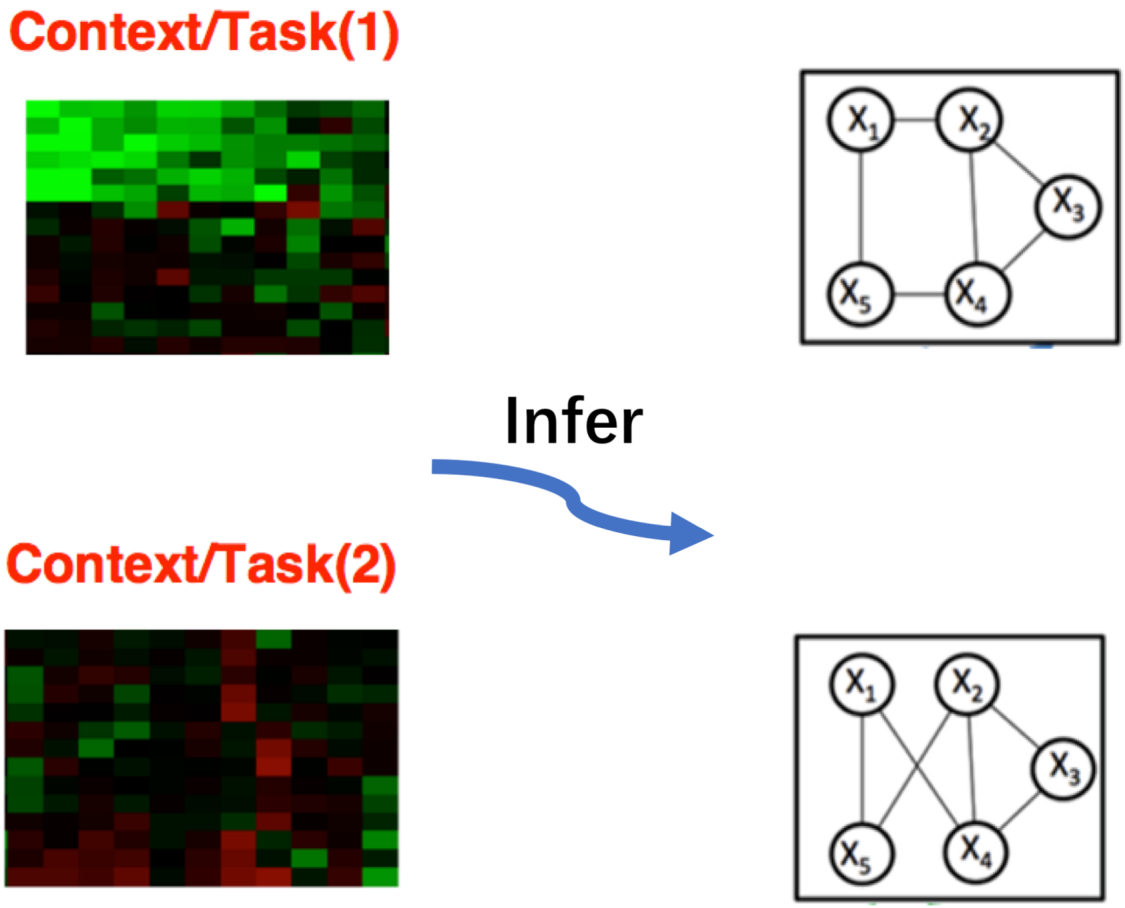
- Multiple Datasets $\mathbf{X}^{(1)}, \ldots \mathbf{X}^{(K)} \rightarrow$ Multiple Graphs $G^{(1)}, \ldots, G^{(K)}$.



- 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

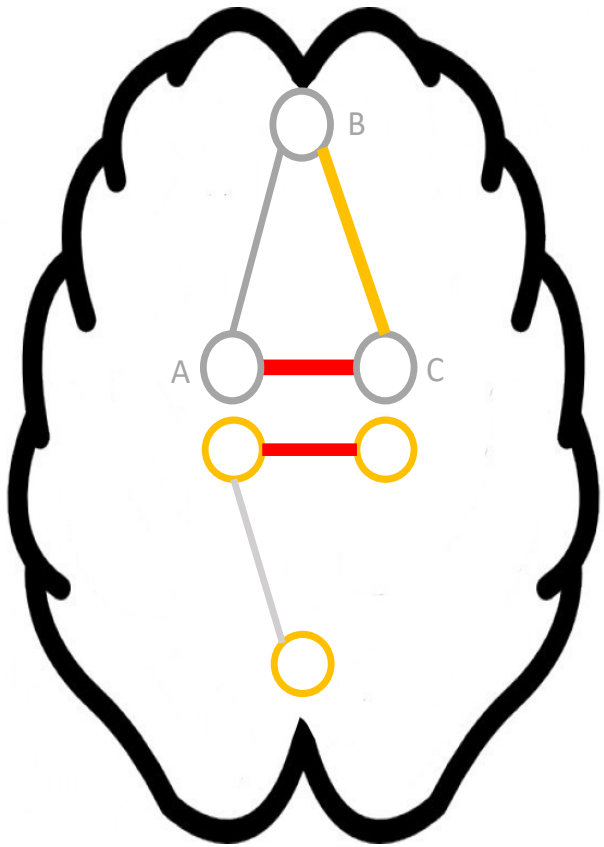
- Multiple Datasets $\mathbf{X}^{(1)}, \ldots \mathbf{X}^{(K)} \rightarrow$ Multiple Graphs $G^{(1)}, \ldots, G^{(K)}$.



- Current Approch: Multi-sGGMs $\Rightarrow$
  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

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

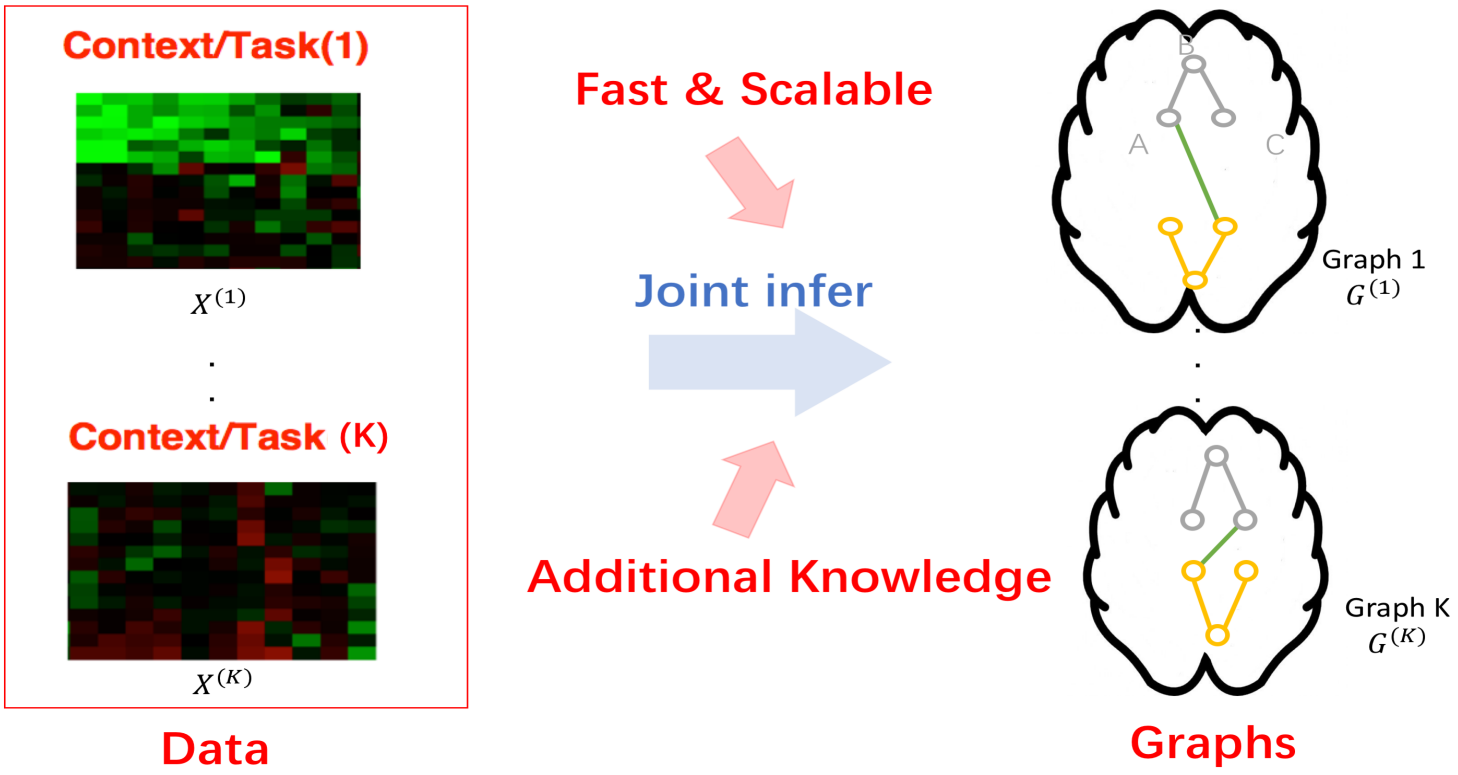
e.g., $K = 91$ and $p = 30$K

- W-SIMULE (Constrained ℓ_1 minimization based): 6 trillion years
- JGL (MLE): 3.5 days / iter

[1] [Singh et al.(2017)Singh, Wang, and Qi]
[2] [Danaher et al.(2013)Danaher, Wang, and Witten]

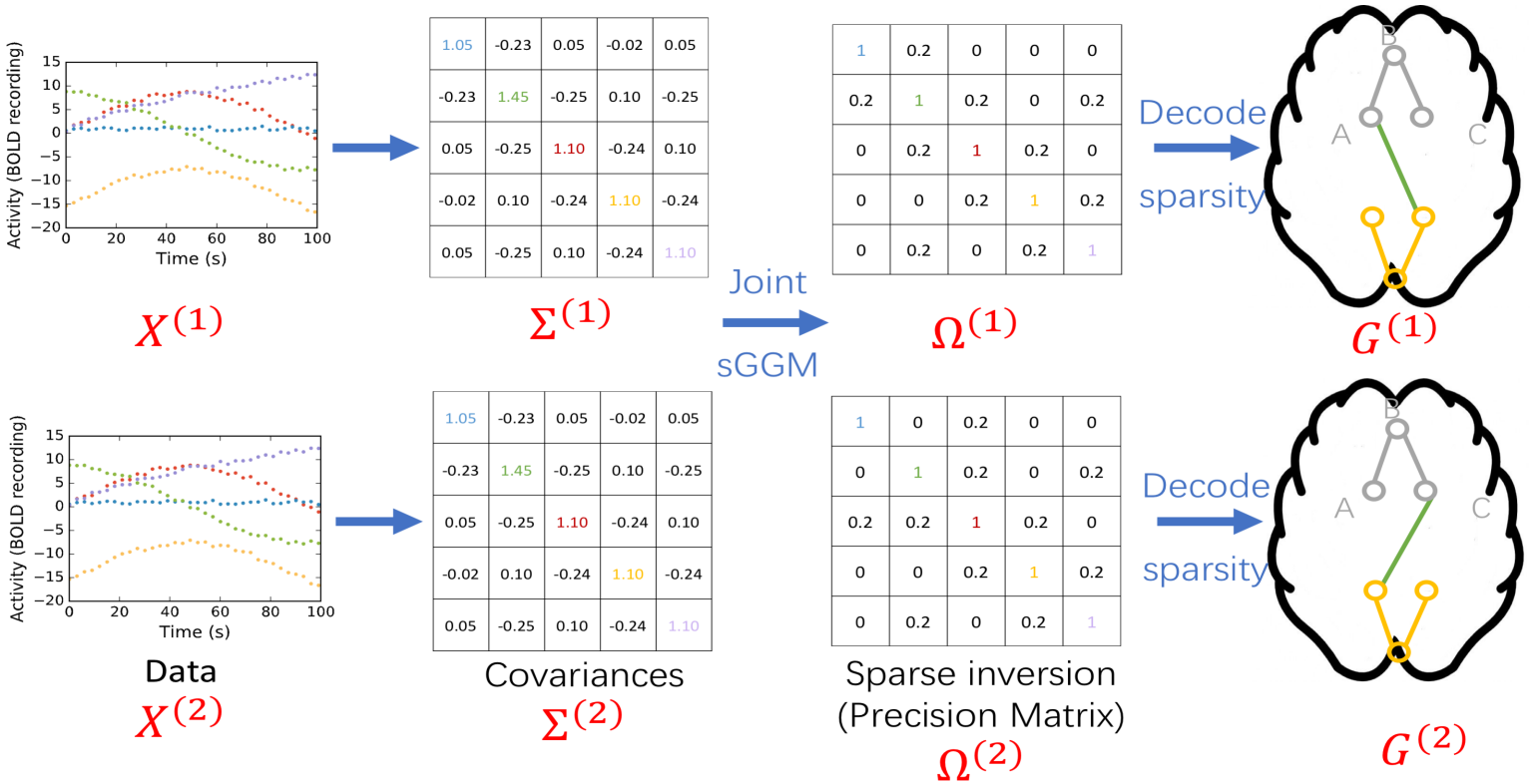
Summary of contributions: Integrating Additional Knowledge in Scalable Learning of multi-sGGMs

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



Notations: Multi-sGGMs

- A pipeline to infer Graph from heterogeneous datasets $\mathbf{X}^{(1)}, \ldots \mathbf{X}^{(K)}$[3].



[3] X^{tot}: the concatenation of $(X^{(1)}, X^{(2)}, \ldots, X^{(K)})$.
Σ^{tot}: the concatenation of $(\Sigma^{(1)}, \Sigma^{(2)}, \ldots, \Sigma^{(K)})$.
Ω^{tot}: the concatenation of $(\Omega^{(1)}, \Omega^{(2)}, \ldots, \Omega^{(K)})$.

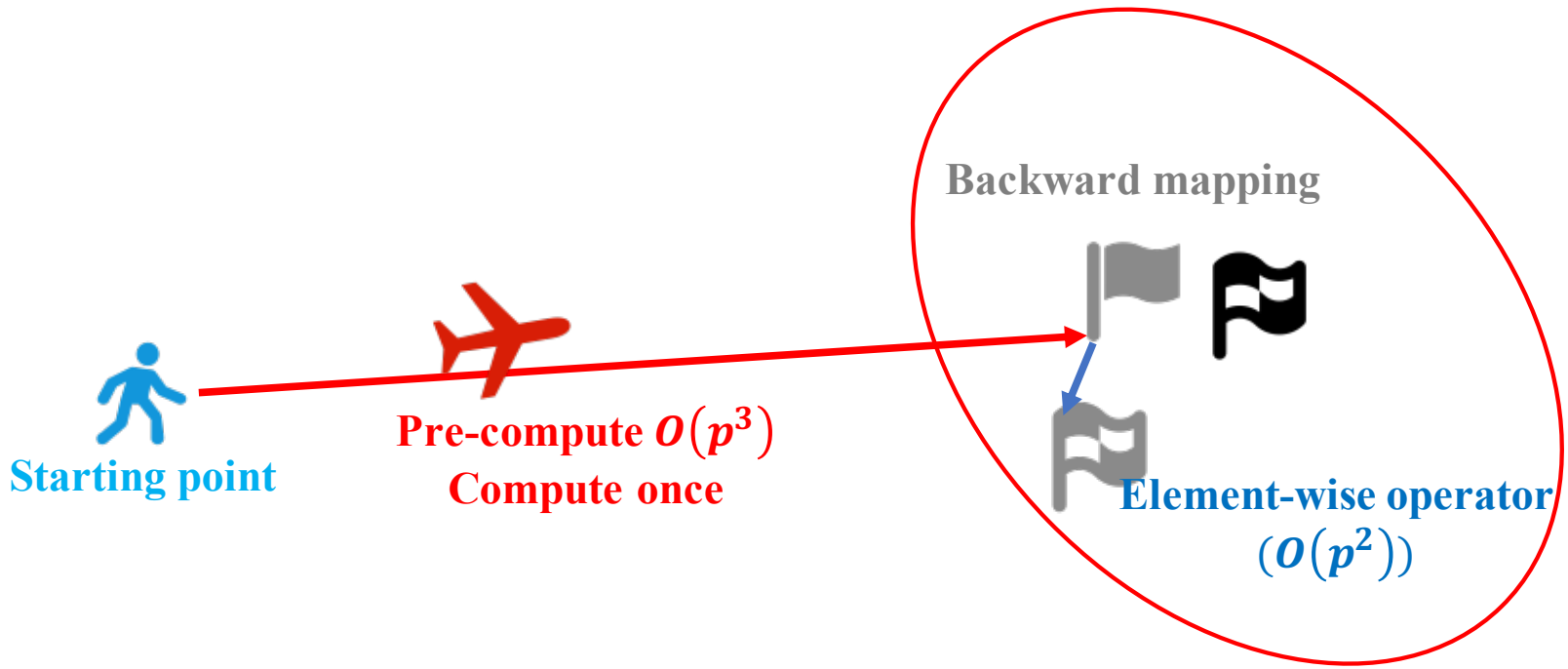
Notations

$X^{(i)}$	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}	$(W_I^{(1)}, W_I^{(2)}, \ldots, W_I^{(K)})$
W_S^{tot}	$(W_S, W_S, \ldots, W_S)$

Solution to Limitation II: Elementary Estimator for sGGM

Elementary Estimator

$$\begin{aligned} &\underset{\theta}{\operatorname{argmin}} \mathcal{R}(\theta) \\ &\text{Subject to: } \mathcal{R}^*(\theta - \mathcal{B}^*(\widehat{\phi})) \leq \lambda_n \end{aligned} \tag{1.1}$$



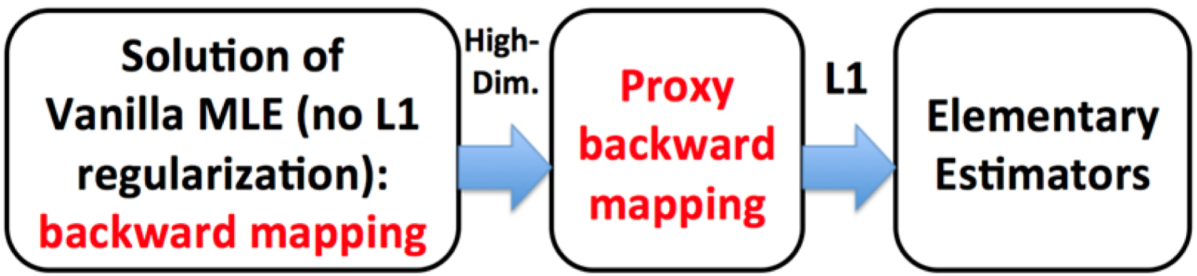
Background: Elementary Estimator

- The canonical parameter θ of an exponential family distribution can be learned through EE:

Elementary Estimator

$$\begin{aligned} &\underset{\theta}{\operatorname{argmin}} \mathcal{R}(\theta) \\ &\text{Subject to: } \mathcal{R}^*(\theta - \mathcal{B}^*(\widehat{\phi})) \leq \lambda_n \end{aligned} \tag{1.2}$$

- $\mathcal{B}(\widehat{\phi})$ denotes the backward mapping (i.e., vanilla MLE solution);
- For high-dimensional situations, vanilla MLE solutions are mostly not available. Therefore, we choose Proxy backward mapping $\mathcal{B}^*(\widehat{\phi})$.



Solution to Limitation II: Elementary Estimator for sGGM

Elementary Estimator

$$\begin{aligned} &\underset{\theta}{\operatorname{argmin}} \mathcal{R}(\theta) \\ &\text{Subject to: } \mathcal{R}^*(\theta - \mathcal{B}^*(\widehat{\phi})) \leq \lambda_n \end{aligned} \tag{1.3}$$

- For example, for sGGM:

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

Elementary Estimator for sGGM

$$\begin{aligned} &\underset{\Omega}{\operatorname{argmin}} ||\Omega||_1 \\ &\text{Subject to: } ||\Omega - [T_v(\widehat{\Sigma})]^{-1}||_\infty \leq \lambda_n \end{aligned} \tag{1.4}$$

single-sGGM Method	Time Complexity	Note
GLasso[4]	$O(p^3)$	Not a closed-form solution
EE-sGGM[5]	$O(p^2)$	entry-wise closed-form sharp convergence rate

multi-sGGM Method	Time Complexity	Notes
W-SIMULE[6]	$O(K^4 p^5)$	LP with Kp variables
JGL[7]	$O(T \times Kp^3)$	SVD
Proposed method	$O(K^4 p^2)$	entry-wise fast sharp convergence rate

[4] [Friedman et al.(2008)Friedman, Hastie, and Tibshirani]
[5] [Yang et al.(2014)Yang, Lozano, and Ravikumar]
[6] [Singh et al.(2017)Singh, Wang, and Qi]
[7] [Danaher et al.(2013)Danaher, Wang, and Witten]

Solution to Limitation I: Using Knowledge as Weight in Regularization (KW-norm)

EE	$\mathcal{R}(\cdot)$	θ	$\mathcal{B}^*$	$\mathcal{R}^*$
JEEK		Ω^{tot}	$inv[T_v(\widehat{\Sigma}^{tot})]$	

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

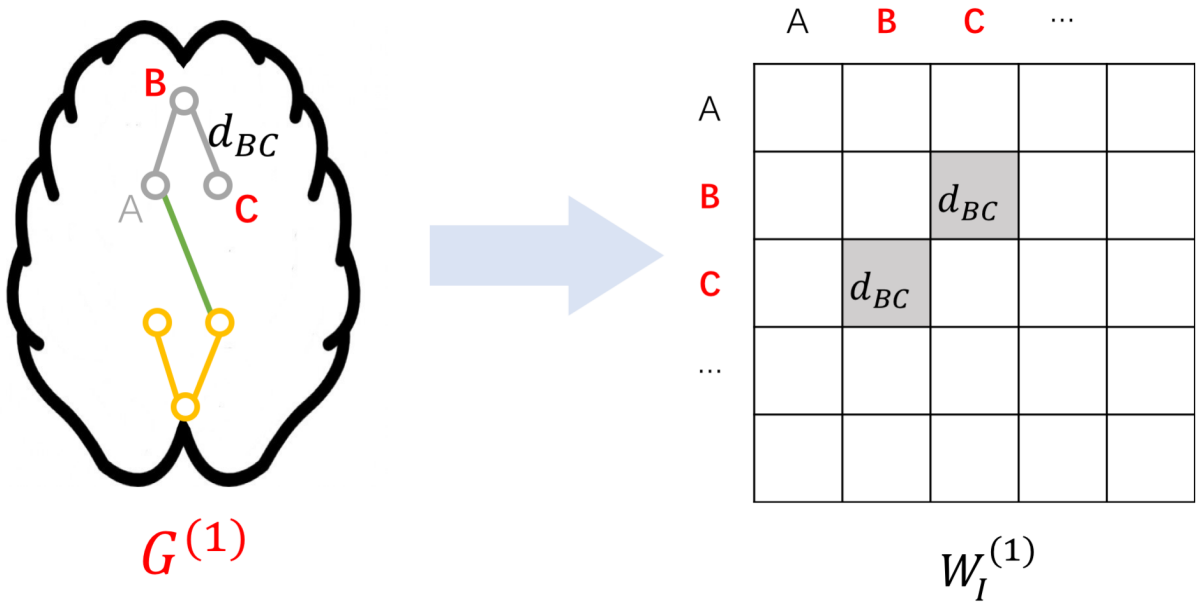
KW-norm

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

- W_I^{tot}: weights describing knowledge of each individual graph.
- W_S^{tot}: weights describing knowledge of the shared graph.
- No need to design knowledge-specific optimization
- KW-norm is flexible.

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

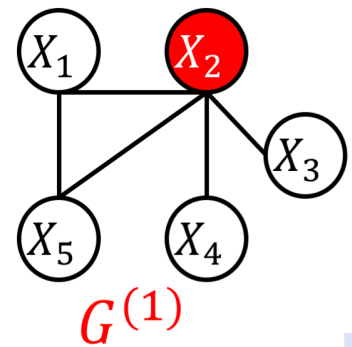
- e.g., Spatial distance among brain regions;



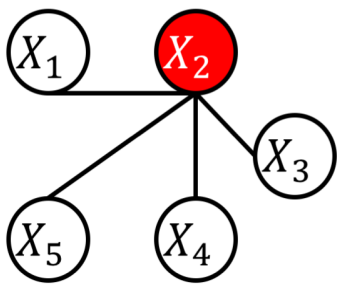
- Examples of adding more types of konwledge in Section S:5.

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

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



$G^{(1)}$



$G^{(2)}$

	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

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

$$\begin{aligned} \text{Subject to: } & ||(1 \oslash W_I^{tot}) \circ (\Omega^{tot} - inv(T_v(\widehat{\Sigma}^{tot})))||_\infty \leq \lambda_n \\ & ||(1 \oslash W_S^{tot}) \circ (\Omega^{tot} - inv(T_v(\widehat{\Sigma}^{tot})))||_\infty \leq \lambda_n \\ & \Omega^{tot} = \Omega_S^{tot} + \Omega_I^{tot} \end{aligned} \tag{2.2}$$

[8] Where $\oslash$ represents element wise division $[A \oslash B]_{ij} = A_{ij}/B_{ij}$.

Proposed method: JEEK – Solution

- Fast and Scalable solution[9] – p^2 small linear programming subproblems:

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

[9] $a_i := \Omega_I^{(i)}{}_{j,k}$ (the $\{j,k\}$-th entry of $\Omega^{(i)}$)
$b := \Omega_{S\,j,k}$
$c_i = [T_v(\widehat{\Sigma}^{(i)})]^{-1}_{j,k}$.
$W^{(i)}_{j,k} = w_i$ and $W^{S}_{j,k} = w_s$.

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
JEEK	$O(K^4p^2)$ ($\Rightarrow$ $O(K^4)$ if parallelizing completely)	YES
W-SIMULE	$O(K^4p^5)$	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
JEEK	$O(K^4p^2)$ ($\Rightarrow$ $O(K^4)$ if parallelizing completely)	YES
W-SIMULE	$O(K^4p^5)$	YES
JGL	$O(T \times Kp^3)$	NO

- Theoretical guaranteed

Theoretical Results

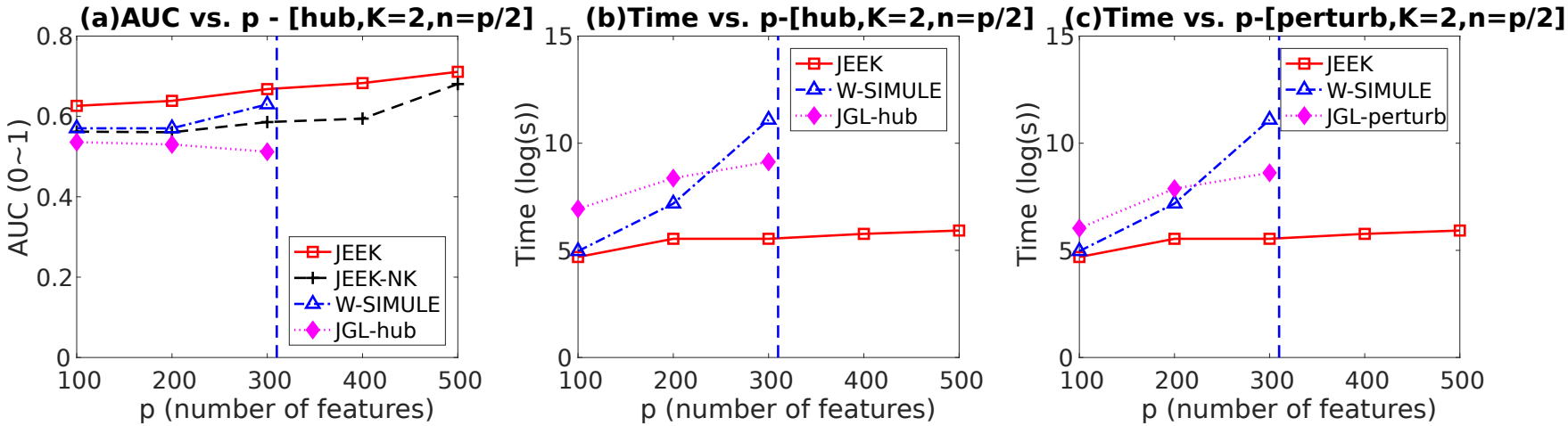
- Error bound: $||\widehat{\Omega}^{tot} - \Omega^{tot*}||$
- Sharp convergence rate as the state-of-art

$$
\begin{aligned}
&||\widehat{\Omega}^{tot} - \Omega^{tot*}||_F \leq 4\sqrt{k_i + k_s}\lambda_n \\
&\max(||(1 \oslash W_I^{tot}) \circ (\widehat{\Omega}^{tot} - \Omega^{tot*})||_\infty, ||(1 \oslash W_S^{tot}) \circ (\widehat{\Omega}^{tot} - \Omega^{tot*}||_\infty) \leq 2\lambda_n \\
&||(1 \oslash W_I^{tot}) \circ (\widehat{\Omega}_I^{tot} - \Omega_I^{tot*})||_1 + ||(1 \oslash W_S^{tot}) \circ (\widehat{\Omega}_S^{tot} - \Omega_S^{tot*})||_1 \leq 8(k_i + k_s)\lambda_n
\end{aligned} \tag{3.1}
$$

Where a, c, κ_1 and κ_2 are constants

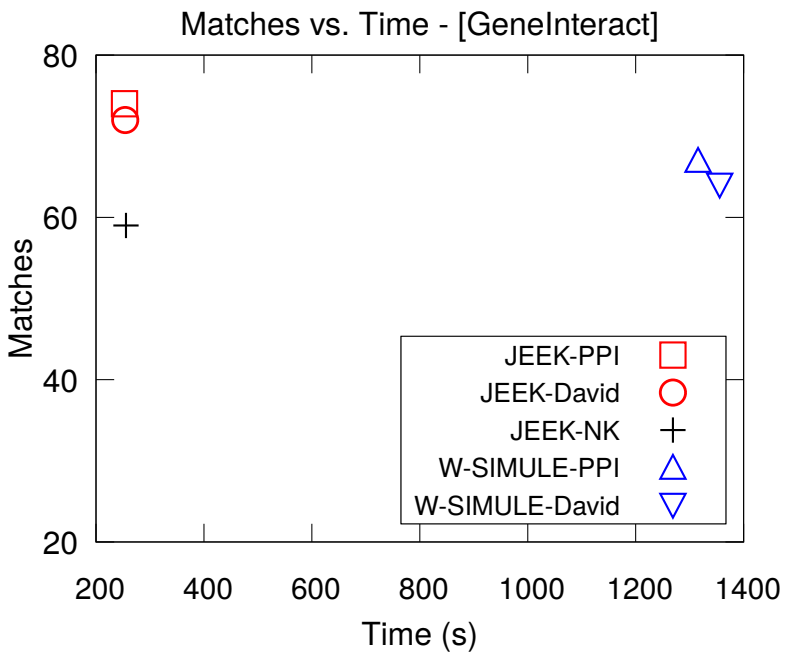
$$
\begin{aligned}
&||\widehat{\Omega}^{tot} - \Omega^{tot*}||_F \\
&\qquad \leq \frac{16\kappa_1 a \max_{j,k}(W_I^{tot}{}_{j,k}, W_S^{tot}{}_{j,k})}{\kappa_2} \sqrt{\frac{(k_i + k_s)\log(Kp)}{n_{tot}}}
\end{aligned} \tag{3.2}
$$

Empirical Results on Multiple Synthetic Datasets

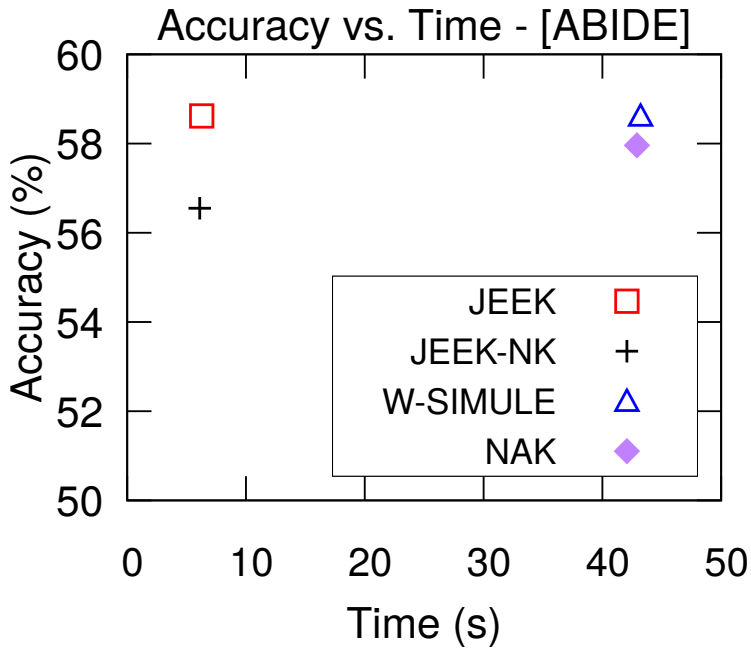


- JEEK outperforms the speed of the state-of arts significantly faster ($\sim 5000\times$ improvement);
- JEEK obtains better AUC as the state-of-the-art;
- JEEK obtains better AUC than JEEK-NK (no additional knowledge).

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)`
- See Poster 133
- Acknowledgement: NSF CAREER award No. 1453580

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.

J. Friedman, T. Hastie, and R. Tibshirani.

Sparse inverse covariance estimation with the graphical lasso.

Biostatistics, 9(3):432–441, 2008.

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.