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]

Abstract

We consider the problem of including additional knowledge in estimating sparse Gaussian graphical models (sGGMs) from aggregated samples, arising often in bioinformatics and neuroimaging applications. Previous joint sGGM estimators either fail to use existing knowledge or cannot scale-up to many tasks (large K) under a high-dimensional (large p) situation. In this paper, we propose a novel Joint Elementary Estimator incorporating additional Knowledge (JEEK) to infer multiple related sparse Gaussian Graphical models from large-scale heterogeneous data. Using domain knowledge as weights, we design a novel hybrid norm as the minimization objective to enforce the superposition of two weighted sparsity constraints, one on the shared interactions and the other on the task-specific structural patterns. This enables JEEK to elegantly consider various forms of existing knowledge based on the domain at hand and avoid the need to design knowledge-specific optimization. JEEK is solved through a fast and entry-wise parallelizable solution that largely improves the computational efficiency of the state-of-the-art $O(p^5K^4)$ to $O(p^2K^4)$. We conduct a rigorous statistical analysis showing that JEEK achieves the same convergence rate $O(\log(Kp)/n_{tot})$ as the state-of-the-art estimators that are much harder to compute. Empirically, on multiple synthetic datasets and two real-world data, JEEK outperforms the speed of the state-of-arts significantly while achieving the same level of prediction accuracy.

[1]Department of Computer Science, University of Virginia, *http://www.jointnets.org/*. Correspondence to: Beilun Wang <bw4mw@virginia.edu>, Yanjun Qi <yanjun@virginia.edu>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

1 Introduction

Technology revolutions in the past decade have collected large-scale heterogeneous samples from many scientific domains. For instance, genomic technologies have delivered petabytes of molecular measurements across more than hundreds of types of cells and tissues from national projects like ENCODE (Consortium et al., 2012) and TCGA (Network et al., 2011). Neuroimaging technologies have generated petabytes of functional magnetic resonance imaging (fMRI) datasets across thousands of human subjects (shared publicly through projects like openfMRI (Poldrack et al., 2013). Given such data, understanding and quantifying variable graphs from heterogeneous samples about multiple contexts is a fundamental analysis task.

Such variable graphs can significantly simplify network-driven studies about diseases (Ideker & Krogan, 2012), can help understand the neural characteristics underlying clinical disorders (Uddin et al., 2013) and can allow for understanding genetic or neural pathways and systems. The number of contexts (denoted as K) that those applications need to consider grows extremely fast, ranging from tens (e.g., cancer types in TCGA (Network et al., 2011)) to thousands (e.g., number of subjects in openfMRI (Poldrack et al., 2013)). The number of variables (denoted as p) ranges from hundreds (e.g., number of brain regions) to tens of thousands (e.g., number of human genes).

The above data analysis problem can be formulated as jointly estimating K conditional dependency graphs $G^{(1)}, G^{(2)}, \ldots, G^{(K)}$ on a single set of p variables based on heterogeneous samples accumulated from K distinct contexts. For homogeneous data samples from a given i-th context, one typical approach is the sparse Gaussian Graphical Model (sGGM) (Lauritzen, 1996; Yuan & Lin, 2007). sGGM assumes samples are independently and identically drawn from $N_p(\mu^{(i)}, \Sigma^{(i)})$, a multivariate Gaussian distribution with mean vector $\mu^{(i)}$ and covariance matrix $\Sigma^{(i)}$. The graph structure $G^{(i)}$ is encoded by the sparsity pattern of the inverse covariance matrix, also named precision matrix, $\Omega^{(i)}$. $\Omega^{(i)} := (\Sigma^{(i)})^{-1}$. $\Omega^{(i)}_{jk} = 0$ if and only if in $G^{(i)}$ an edge does not connect j-th node and k-th node (i.e., conditional independent). sGGM imposes an ℓ_1 penalty on the parameter $\Omega^{(i)}$ to achieve a consistent estimation

under high-dimensional situations. When handling heterogeneous data samples, rather than estimating sGGM of each condition separately, a multi-task formulation that jointly estimates K different but related sGGMs can lead to a better generalization(Caruana, 1997).

Previous studies for joint estimation of multiple sGGMs roughly fall into four categories: (Danaher et al., 2013; Mohan et al., 2013; Chiquet et al., 2011; Honorio & Samaras, 2010; Guo et al., 2011; Zhang & Wang, 2012; Zhang & Schneider, 2010; Zhu et al., 2014): (1) The first group seeks to optimize a sparsity regularized data likelihood function plus an extra penalty function $\mathcal{R}'$ to enforce structural similarity among multiple estimated networks. Joint graphical lasso (JGL) (Danaher et al., 2013) proposed an alternating direction method of multipliers (ADMM) based optimization algorithm to work with two regularization functions ($\ell_1 + \mathcal{R}'$). (2) The second category tries to recover the support of $\Omega^{(i)}$ using sparsity penalized regressions in a column by column fashion. Recently (Monti et al., 2015) proposed to learn population and subject-specific brain connectivity networks via a so-called "Mixed Neighborhood Selection" (MSN) method in this category. (3) The third type of methods seeks to minimize the joint sparsity of the target precision matrices under matrix inversion constraints. One recent study, named SIMULE (Shared and Individual parts of MULtiple graphs Explicitly) (Wang et al., 2017b), automatically infers both specific edge patterns that are unique to each context and shared interactions preserved among all the contexts (i.e. by modeling each precision matrix as $\Omega^{(i)} = \Omega_I^{(i)} + \Omega_S$) via the constrained ℓ_1 minimization. Following the CLIME estimator (Pang et al., 2014), the constrained ℓ_1 convex formulation can also be solved column by column via linear programming. However, all three categories of aforementioned estimators are difficult to scale up when the dimension p or the number of tasks K are large because they cannot avoid expensive steps like SVD (Danaher et al., 2013) for JGL, linear programming for SIMULE or running multiple iterations of p expensive penalized regressions in MNS. (4) The last category extends the so-called "Elementary Estimator" graphical model (EE-GM) formulation (Yang et al., 2014b) to revise JGL's penalized likelihood into a constrained convex program that minimizes ($\ell_1 + \mathcal{R}'$). One proposed estimator FASJEM (Wang et al., 2017a) is solved in an entry-wise manner and group-entry-wise manner that largely outperforms the speed of its JGL counterparts. More details of the related works are in Section (5).

One significant caveat of state-of-the-art joint sGGM estimators is the fact that little attention has been paid to incorporating existing knowledge of the nodes or knowledge of the relationships among nodes in the models. In addition to the samples themselves, additional information is widely available in real-world applications. In fact, incorporating the knowledge is of great scientific interest. A prime example is when estimating the functional brain connectivity networks among brain regions based on fMRI samples, the spatial position of the regions are readily available. Neuroscientists have gathered considerable knowledge regarding the spatial and anatomical evidence underlying brain connectivity (e.g., short edges and certain anatomical regions are more likely to be connected (Watts & Strogatz, 1998)). Another important example is the problem of identifying gene-gene interactions from patients' gene expression profiles across multiple cancer types. Learning the statistical dependencies among genes from such heterogeneous datasets can help to understand how such dependencies vary from normal to abnormal and help to discover contributing markers that influence or cause the diseases. Besides the patient samples, state-of-the-art bio-databases like HPRD (Prasad et al., 2009) have collected a significant amount of information about direct physical interactions among corresponding proteins, regulatory gene pairs or signaling relationships collected from high-qualify bio-experiments.

Although being strong evidence of structural patterns we aim to discover, this type of information has rarely been considered in the joint sGGM formulation of such samples. To the authors' best knowledge, only one study named as W-SIMULE tried to extend the constrained ℓ_1 minimization in SIMULE into weighted ℓ_1 for considering spatial information of brain regions in the joint discovery of heterogeneous neural connectivity graphs (Singh et al., 2017). This method was designed just for the neuroimaging samples and has $O(p^5K^4)$ time cost, making it not scalable for large-scale settings (more details in Section 3).

This paper aims to fill this gap by adding additional knowledge most effectively into scalable and fast joint sGGM estimations. We propose a novel model, namely Joint Elementary Estimator incorporating additional Knowledge (JEEK), that presents a principled and scalable strategy to include additional knowledge when estimating multiple related sGGMs jointly. Briefly speaking, this paper makes the following contributions:

- **Novel approach:** JEEK presents a new way of integrating additional knowledge in learning multi-task sGGMs in a scalable way. (Section 3)
- **Fast optimization:** We optimize JEEK through an entry-wise and group-entry-wise manner that can dramatically improve the time complexity to $O(p^2K^4)$. (Section 3.4)
- **Convergence rate:** We theoretically prove the convergence rate of JEEK as $O(\log(Kp)/n_{tot})$. This rate shows the benefit of joint estimation and achieves the same convergence rate as the state-of-the-art that are much harder to compute. (Section 4)
- **Evaluation:** We evaluate JEEK using several synthetic datasets and two real-world data, one from neuroscience and one from genomics. It outperforms state-of-the-art

baselines significantly regarding the speed. (Section 6)

JEEK provides the flexibility of using $(K+1)$ different weight matrices representing the extra knowledge. We try to showcase a few possible designs of the weight matrices in Section S:5, including (but not limited to):

- Spatial or anatomy knowledge about brain regions;
- Knowledge of known co-hub nodes or perturbed nodes;
- Known group information about nodes, such as genes belonging to the same biological pathway or cellular location;
- Using existing known edges as the knowledge, like the known protein interaction databases for discovering gene networks (a semi-supervised setting for such estimations).

We sincerely believe the scalability and flexibility provided by JEEK can make structure learning of joint sGGM feasible in many real-world tasks.

Att: Due to space limitations, we have put details of certain contents (e.g., proofs) in the appendix. Notations with "S:" as the prefix in the numbering mean the corresponding contents are in the appendix. For example, full proofs are in Section (S:3).

Notations: math notations we use are described in Section (S:1). $n_{tot} = \sum_{i=1}^{K} n_i$ is the total number of data samples.

2 Background

Sparse Gaussian graphical model (sGGM): The classic formulation of estimating sparse Gaussian Graphical model (Yuan & Lin, 2007) from a single given condition (single sGGM) is the "graphical lasso" estimator (GLasso) (Yuan & Lin, 2007; Banerjee et al., 2008). It solves the following ℓ_1 penalized maximum likelihood estimation (MLE) problem:

$$\underset{\Omega>0}{\operatorname{argmin}} -\log\det(\Omega) + <\Omega, \widehat{\Sigma}> + \lambda_n ||\Omega||_1 \qquad (2.1)$$

M-Estimator with Decomposable Regularizer in High--Dimensional Situations: Recently the seminal study (Negahban et al., 2009) proposed a unified framework for high-dimensional analysis of the following general formulation: M-estimators with decomposable regularizers:

$$\underset{\theta}{\operatorname{argmin}}\, \mathcal{L}(\theta) + \lambda_n \mathcal{R}(\theta) \qquad (2.2)$$

where $\mathcal{R}(\cdot)$ represents a decomposable regularization function and $\mathcal{L}(\cdot)$ represents a loss function (e.g., the negative log-likelihood function in sGGM $\mathcal{L}(\Omega) = -\log\det(\Omega) + <\Omega, \widehat{\Sigma}>$). Here $\lambda_n > 0$ is the tuning parameter.

Elementary Estimators (EE): Using the analysis framework from (Negahban et al., 2009), recent studies (Yang et al., 2014a;b;c) propose a new category of estimators named "Elementary estimator" (EE) with the following general formulation:

$$\begin{aligned} &\underset{\theta}{\operatorname{argmin}}\, \mathcal{R}(\theta) \\ &\text{subject to:} \mathcal{R}^*(\theta - \widehat{\theta}_n) \le \lambda_n \end{aligned} \qquad (2.3)$$

Where $\mathcal{R}^*(\cdot)$ is the dual norm of $\mathcal{R}(\cdot)$,

$$\mathcal{R}^*(v) := \sup_{u \neq 0} \frac{<u, v>}{\mathcal{R}(u)} = \sup_{\mathcal{R}(u) \le 1} <u, v> . \qquad (2.4)$$

The solution of Eq. (2.3) achieves the near optimal convergence rate as Eq. (2.2) when satisfying certain conditions. $\mathcal{R}(\cdot)$ represents a decomposable regularization function (e.g., ℓ_1-norm) and $\mathcal{R}^*(\cdot)$ is the dual norm of $\mathcal{R}(\cdot)$ (e.g., ℓ_∞-norm is the dual norm of ℓ_1-norm). λ_n is a regularization parameter.

The basic motivation of Eq. (2.3) is to build simpler and possibly fast estimators, that yet come with statistical guarantees that are nonetheless comparable to regularized MLE. $\widehat{\theta}_n$ needs to be carefully constructed, well-defined and closed-form for the purpose of simpler computations. The formulation defined by Eq. (2.3) is to ensure its solution having the desired structure defined by $\mathcal{R}(\cdot)$. For cases of high-dimensional estimation of linear regression models, $\widehat{\theta}_n$ can be the classical ridge estimator that itself is closed-form and with strong statistical convergence guarantees in high-dimensional situations.

EE-sGGM:(Yang et al., 2014b) proposed elementary estimators for graphical models (GM) of exponential families, in which $\widehat{\theta}_n$ represents so-called proxy of backward mapping for the target GM (more details in Section S:4). The key idea (summarized in the upper row of Figure 1) is to investigate the vanilla MLE and where it breaks down for estimating a graphical model of exponential families in the case of high-dimensions (Yang et al., 2014b). Essentially the vanilla graphical model MLE can be expressed as a backward mapping that computes the model parameters from some given moments in an exponential family distribution. For instance, in the case of learning Gaussian GM (GGM) with vanilla MLE, the backward mapping is $\widehat{\Sigma}^{-1}$ that estimates Ω from the sample covariance matrix (moment) $\widehat{\Sigma}$. We introduce the details of backward mapping in Section S:4.

However, even though this backward mapping has a simple closed form for GGM, the backward mapping is normally not well-defined in high-dimensional settings. When given the sample covariance $\widehat{\Sigma}$, we cannot just compute the vanilla MLE solution as $[\widehat{\Sigma}]^{-1}$ for GGM since $\widehat{\Sigma}$ is rank-deficient when $p > n$. Therefore Yang et al. (Yang et al., 2014b) used carefully constructed proxy backward maps as $\widehat{\theta}_n = [T_v(\widehat{\Sigma})]^{-1}$ that is both available in closed-form, and

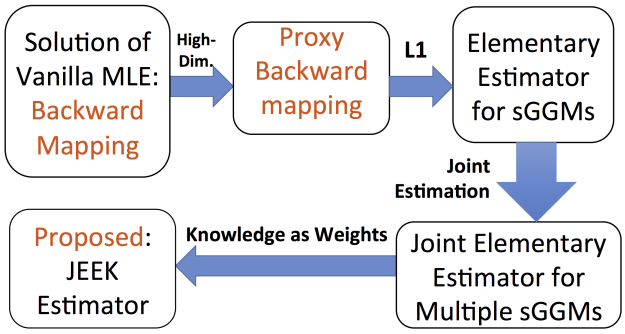

Figure 1. Basic idea of JEEK.

well-defined in high-dimensional settings for GGMs. We introduce the details of $[T_v(\widehat{\Sigma})]^{-1}$ and its statistical property in Section S:4. Now Eq. (2.3) becomes the following closed-form estimator for learning sparse Gaussian graphical models (Yang et al., 2014b):

$$\begin{aligned} &\underset{\Omega}{\operatorname{argmin}} ||\Omega||_{1,,\text{off}} \\ &\text{subject to:} ||\Omega - [T_v(\widehat{\Sigma})]^{-1}||_{\infty,\text{off}} \leq \lambda_n \end{aligned} \tag{2.5}$$

Eq. (2.5) is a special case of Eq. (2.3), in which $\mathcal{R}(\cdot)$ is the off-diagonal ℓ_1-norm and the precision matrix Ω is the θ we search for. When $\mathcal{R}(\cdot)$ is the ℓ_1-norm, the solution of Eq. (2.3) (and Eq. (2.5)) just needs to perform entry-wise thresholding operations on $\widehat{\theta}_n$ to ensure the desired sparsity structure of its final solution.

3 Proposed Method: JEEK

In applications of Gaussian graphical models, we typically have more information than just the data samples themselves. This paper aims to propose a simple, scalable and theoretically-guaranteed joint estimator for estimating multiple sGGMs with additional knowledge in large-scale situations.

3.1 A Joint EE (JEE) Formulation

We first propose to jointly estimate multiple related sGGMs from K data blocks using the following formulation:

$$\underset{\Omega^{(1)},\Omega^{(2)},\ldots,\Omega^{(K)}}{\operatorname{argmin}} \sum_{i=1}^{K} \mathcal{L}(\Omega^{(i)}) + \lambda_n \mathcal{R}(\Omega^{(1)}, \Omega^{(2)}, \ldots, \Omega^{(K)}) \tag{3.1}$$

where $\Omega^{(i)}$ denotes the precision matrix for i-th task. $\mathcal{L}(\Omega) = -\log\det(\Omega) + < \Omega, \widehat{\Sigma} >$ describes the negative log-likelihood function in sGGM. $\Omega^{(i)} \succ 0$ means that $\Omega^{(i)}$ needs to be a positive definite matrix. $\mathcal{R}(\cdot)$ represents a decomposable regularization function enforcing sparsity and structure assumptions (details in Section (3.2)).

For ease of notation, we denote that $\Omega^{tot} = (\Omega^{(1)}, \Omega^{(2)}, \ldots, \Omega^{(K)})$ and $\Sigma^{tot} = (\Sigma^{(1)}, \Sigma^{(2)}, \ldots, \Sigma^{(K)})$. Ω^{tot} and Σ^{tot} are both $p \times Kp$ matrices (i.e., Kp^2 parameters to estimate). Now define an inverse function as $\text{inv}(A^{tot}) := ({A^{(1)}}^{-1}, {A^{(2)}}^{-1}, \ldots, {A^{(K)}}^{-1})$, where A_{tot} is a given $p \times Kp$ matrix with the same structure as Σ_{tot}. Then we rewrite Eq. (3.1) into the following form:

$$\underset{\Omega^{tot}}{\operatorname{argmin}} \mathcal{L}(\Omega^{tot}) + \lambda_n \mathcal{R}(\Omega^{tot}) \tag{3.2}$$

Now connecting Eq. (3.2) to Eq. (2.2) and Eq. (2.3), we propose the following joint elementary estimator (JEE) for learning multiple sGGMs:

$$\begin{aligned} &\underset{\Omega^{tot}}{\operatorname{argmin}} \mathcal{R}(\Omega^{tot}) \\ &\text{subject to: } \mathcal{R}^*(\Omega^{tot} - \widehat{\Omega}^{tot}_{n_{tot}}) \leq \lambda_n \end{aligned} \tag{3.3}$$

The fundamental component in Eq. (2.3) for the single context sGGM was to use a well-defined proxy function to approximate the vanilla MLE solution (named as the backward mapping for exponential family distributions) (Yang et al., 2014b). The proposed proxy $\widehat{\theta}_n = [T_v(\widehat{\Sigma})]^{-1}$ is both well-defined under high-dimensional situations and also has a simple closed-form. Following a similar idea, when learning multiple sGGMs, we propose to use $inv(T_v(\widehat{\Sigma}^{tot}))$ for $\widehat{\Omega}^{tot}_{n_{tot}}$ and get the following joint elementary estimator:

$$\begin{aligned} &\underset{\Omega^{tot}}{\operatorname{argmin}} \mathcal{R}(\Omega^{tot}) \\ &\text{Subject to: } \mathcal{R}^*(\Omega^{tot} - inv(T_v(\widehat{\Sigma}^{tot}))) \leq \lambda_n \end{aligned} \tag{3.4}$$

3.2 Knowledge as Weight (KW-Norm)

The main goal of this paper is to design a principled strategy to incorporate existing knowledge (other than samples or structured assumptions) into the multi-sGGM formulation. We consider two factors in such a design:

(1) When learning multiple sGGMs jointly from real-world applications, it is often of great scientific interests to model and learn context-specific graph variations explicitly, because such variations can "fingerprint" important markers in domains like cognition (Ideker & Krogan, 2012) or pathology (Kelly et al., 2012). Therefore we design to share parameters between different contexts. Mathematically, we model $\Omega^{(i)}$ as two parts:

$$\Omega^{(i)} = \Omega_I^{(i)} + \Omega_S \tag{3.5}$$

where $\Omega_I^{(i)}$ is the individual precision matrix for context i and Ω_S is the shared precision matrix between contexts. Again, for ease of notation we denote $\Omega_I^{tot} = (\Omega_I^{(1)}, \Omega_I^{(2)}, \ldots, \Omega_I^{(K)})$ and $\Omega_S^{tot} = (\Omega_S, \Omega_S, \ldots, \Omega_S)$.

(2) We represent additional knowledge as positive weight matrices from $\mathbb{R}^{p \times p}$. More specifically, we represent

the knowledge of the task-specific graph as weight matrix $\{W^{(i)}\}$ and W_S representing existing knowledge of the shared network. The positive matrix-based representation is a powerful and flexible strategy that can describe many possible forms of existing knowledge. In Section (S:5), we provide a few different designs of $\{W^{(i)}\}$ and W_S for real-world applications. In total, we have weight matrices $\{W_I^{(1)}, W_I^{(2)}, \ldots, W_I^{(K)}, W_S\}$ to represent additional knowledge. To simplify notations, we denote $W_I^{tot} = (W_I^{(1)}, W^{(2)}, \ldots, W_I^{(K)})$ and $W_S^{tot} = (W_S, W_S, \ldots, W_S)$.

Now we propose the following knowledge as weight norm (kw-norm) combining the above two:

$$\mathcal{R}(\Omega^{tot}) = ||W_I^{tot} \circ \Omega_I^{tot}||_1 + ||W_S^{tot} \circ \Omega_S^{tot}||_1 \quad (3.6)$$

Here the Hadamard product $\circ$ is the element-wise product between two matrices i.e. $[A \circ B]_{ij} = A_{ij}B_{ij}$.

The kw-norm(Eq. (3.6)) has the following three properties:

- (i) kw-norm is a norm function if and only if any entries in W_I^{tot} and W_S^{tot} do not equal to 0.
- (ii) If the condition in (i) holds, kw-norm is a decomposable norm.
- (iii) If the condition in (i) holds, the dual norm of kw-norm is $\mathcal{R}^*(u) = \max(||W_I^{tot} \circ u||_\infty, ||W_S^{tot} \circ u||_\infty)$.

Section S:3.1 provides proofs of the above claims.

3.3 JEE with Knowledge (JEEK)

Plugging Eq. (3.6) to Eq. (3.4), we obtain the following formulation of JEEK for learning multiple related sGGMs from heterogereous samples:

$$\begin{aligned} &\underset{\Omega_I^{tot}, \Omega_S^{tot}}{\operatorname{argmin}} ||W_I^{tot} \circ \Omega_I^{tot}||_1 + ||W_S^{tot} \circ \Omega_S^{tot}|| \\ \text{Subject to: } &||W_I^{tot} \circ (\Omega^{tot} - inv(T_v(\widehat{\Sigma}^{tot})))||_\infty \le \lambda_n \\ &||W_S^{tot} \circ (\Omega^{tot} - inv(T_v(\widehat{\Sigma}^{tot})))||_\infty \le \lambda_n \\ &\Omega^{tot} = \Omega_S^{tot} + \Omega_I^{tot} \end{aligned} \quad (3.7)$$

In Section 4, we theoretically prove that the statistical convergence rate of JEEK achieves the same sharp convergence rate as the state-of-the-art estimators for multi-task sGGMs. Our proofs are inspired by the unified framework of the high-dimensional statistics (Negahban et al., 2009).

3.4 Solution of JEEK:

A huge computational advantage of JEEK (Eq. (3.7)) is that it can be decomposed into $p \times p$ independent small linear programming problems. To simplify notations, we denote ${\Omega_I^{(i)}}_{j,k}$ (the $\{j,k\}$-th entry of $\Omega^{(i)}$) as a_i. Similarly we denote $\Omega_{Sj,k}$ as b and $[T_v(\widehat{\Sigma}^{(i)})]_{j,k}^{-1}$ be c_i. Similarly we denote $W_{j,k}^{(i)} = w_i$ and $W_{j,k}^{S} = w_s$. "A group of entries" means a set of parameters $\{a_1, \ldots, a_K, b\}$ for certain j, k.

Algorithm 1. Joint Elementary Estimator with additional knowledge (JEEK) for Multi-task sGGMs

Input: Data sample matrix $\mathbf{X}^{(i)}$ ($i = 1$ **to** K), regularization hyperparameter λ_n, Knowledge weight matrices $\{W_I^{(i)}, W_S\}$ and **LP(.)** (a linear programming solver)
Output: $\{\Omega^{(i)}\}$ ($i = 1$ **to** K)

for $i = 1$ **to** K **do**
  Initialize $\widehat{\Sigma}^{(i)} = \frac{1}{n_i - 1}\sum_{s=1}^{n_i}(\mathbf{X}_{s,}^{(i)} - \widehat{\mu}^{(i)})(\mathbf{X}_{s,}^{(i)} - \widehat{\mu}^{(i)})^T$ (the sample covariance matrix of $\mathbf{X}^{(i)}$)
  Initialize $\Omega^{(i)} = \mathbf{0}_{p \times p}$
  Calculate the proxy backward mapping $[T_v(\widehat{\Sigma}^{(i)})]^{-1}$
end for
for $j = 1$ **to** p **do**
  for $k = 1$ **to** j **do**
    $c_i = [T_v(\widehat{\Sigma}^{(i)})]_{j,k}^{-1}$
    $w_i = W_{j,k}^{(i)}$
    $w_s = W_{Sj,k}$
    $a_i, b =$ **LP**$(w_i, w_s, c_i, \lambda_n)$ where $i = 1, \ldots, K$ and **LP(.)** solves Eq. (3.8)
    for $i = 1$ **to** K **do**
      ${\Omega^{(i)}}_{j,k} = {\Omega^{(i)}}_{k,j} = a_i + b$
      ${\Omega_I^{(i)}}_{j,k} = a_i$
      $\Omega_{Sj,k} = b$
    end for
  end for
end for

In order to estimate $\{a_1, \ldots, a_K, b\}$, JEEK (Eq. (3.7)) can be decomposed into the following formulation for a certain j, k :

$$\begin{aligned} &\underset{a_i, b}{\operatorname{argmin}} \sum_i |w_i a_i| + K|w_s b| \\ \text{Subject to: } &|a_i + b - c_i| \le \frac{\lambda_n}{\min(w_i, w_s)}, \\ &i = 1, \ldots, K \end{aligned} \quad (3.8)$$

Eq. (3.8) can be easily converted into a linear programming form of Eq. (S:1–1) with only $K + 1$ variables. The time complexity of Eq. (3.8) is $O(K^4)$. Considering JEEK has a total $p(p-1)/2$ of such subproblems to solve, the computational complexity of JEEK (Eq. (3.7)) is therefore $O(p^2K^4)$. We summarize the optimization algorithm of JEEK in Algorithm 1 (details in Section (S:1.2)).

4 Theoretical Analysis

KW-Norm: We presented the three properties of kw-norm in Section 3.2. The proofs of these three properties are included in Section (S:3.1).

Theoretical error bounds of Proxy Backward Mapping: (Yang et al., 2014b) proved that when ($p \ge n$), the proxy backward mapping $[T_v(\widehat{\Sigma})]^{-1}$ used by EE-sGGM achieves the sharp convergence rate to its truth (i.e., by proving $||T_v(\widehat{\Sigma}))^{-1} - \Sigma^{*-1}||_\infty = O(\sqrt{\frac{\log p}{n}})$). The proof was extended from the previous study (Rothman et al., 2009) that

devised $T_v(\widehat{\Sigma})$ for estimating covariance matrix consistently in high-dimensional situations. See detailed proofs in Section S:4.3. To derive the statistical error bound of JEEK, we need to assume that $inv(T_v(\widehat{\Sigma}^{tot}))$ are well-defined. This is ensured by assuming that the true ${\Omega^{(i)}}^*$ satisfy the conditions defined in Section (S:3.1).

Theoretical error bounds of JEEK: We now use the high-dimensional analysis framework from (Negahban et al., 2009), three properties of kw-norm, and error bounds of backward mapping from (Rothman et al., 2009; Yang et al., 2014b) to derive the statistical convergence rates of JEEK. Detailed proofs of the following theorems are in Section 4 .

Before providing the theorem, we need to define the structural assumption, the IS-Sparsity, we assume for the parameter truth.

(IS-Sparsity): The 'true' parameter of Ω^{tot^*} can be decomposed into two clear structures–$\{\Omega_I^{tot^*}$ and $\Omega_S^{tot^*}\}$. $\Omega_I^{tot^*}$ is exactly sparse with k_i non-zero entries indexed by a support set S_I and $\Omega_S^{tot^*}$ is exactly sparse with k_s non-zero entries indexed by a support set S_S. $S_I \bigcap S_S = \emptyset$. All other elements equal to 0 (in $(S_I \bigcup S_S)^c$).

Theorem 4.1. *Consider Ω^{tot} whose true parameter Ω^{tot^*} satisfies the **(IS-Sparsity)** assumption. Suppose we compute the solution of Eq. (3.7) with a bounded λ_n such that $\lambda_n \geq \max(||W_I^{tot} \circ (\Omega^{tot^*} - inv(T_v(\widehat{\Sigma}^{tot})))||_\infty, ||W_S^{tot} \circ (\Omega^{tot^*} - inv(T_v(\widehat{\Sigma}^{tot})))||_\infty)$, then the estimated solution $\widehat{\Omega}^{tot}$ satisfies the following error bounds:*

$$
\begin{aligned}
&||\widehat{\Omega}^{tot} - \Omega^{tot^*}||_F \leq 4\sqrt{k_i + k_s}\lambda_n \\
&\max(||W_I^{tot} \circ (\widehat{\Omega}^{tot} - \Omega^{tot^*})||_\infty, ||W_S^{tot} \circ (\widehat{\Omega}^{tot} - \Omega^{tot^*}||_\infty) \\
&\qquad \leq 2\lambda_n \\
&||W_I^{tot} \circ (\widehat{\Omega}_I^{tot} - \Omega_I^{tot^*})||_1 + ||W_S^{tot} \circ (\widehat{\Omega}_S^{tot} - \Omega_S^{tot^*})||_1 \\
&\qquad \leq 8(k_i + k_s)\lambda_n
\end{aligned}
\tag{4.1}
$$

Proof. See detailed proof in Section S:3.2 □

Theorem (4.1) provides a general bound for any selection of λ_n. The bound of λ_n is controlled by the distance between Ω^{tot^*} and $inv(T_v(\widehat{\Sigma}^{tot}))$. We then extend Theorem (4.1) to derive the statistical convergence rate of JEEK. This gives us the following corollary:

Corollary 4.2. *Suppose the high-dimensional setting, i.e., $p > \max(n_i)$. Let $v := a\sqrt{\frac{\log(Kp)}{n_{tot}}}$. Then for $\lambda_n := \frac{8\kappa_1 a}{\kappa_2}\sqrt{\frac{\log(Kp)}{n_{tot}}}$ and $n_{tot} > c \log Kp$, with a probability of at least $1 - 2C_1 \exp(-C_2 Kp \log(Kp))$, the estimated optimal solution $\widehat{\Omega}^{tot}$ has the following error bound:*

$$
\begin{aligned}
&||\widehat{\Omega}^{tot} - \Omega^{tot^*}||_F \\
&\qquad \leq \frac{16\kappa_1 a \max_{j,k}({W_I^{tot}}_{j,k}, {W_S^{tot}}_{j,k})}{\kappa_2}\sqrt{\frac{(k_i + k_s)\log(Kp)}{n_{tot}}}
\end{aligned}
\tag{4.2}
$$

where a, c, κ_1 and κ_2 are constants.

Proof. See detailed proof in Section S:3.2.2 (especially from Eq. (S:3–11) to Eq. (S:3–19)). □

Bayesian View of JEEK: In Section (S:2) we provide a direct Bayesian interpretation of JEEK through the perspective of hierarchical Bayesian modeling. Our hierarchical Bayesian interpretation nicely explains the assumptions we make in JEEK.

5 Connecting to Relevant Studies

JEEK is closely related to a few state-of-the-art studies summarized in Table 1. We compare the time complexity and functional properties of JEEK versus these studies.

NAK: (Bu & Lederer, 2017) For the single task sGGM, one recent study (Bu & Lederer, 2017) (following ideas from (Shimamura et al., 2007)) proposed to integrating Additional Knowledge (NAK)into estimation of graphical models through a weighted Neighbourhood selection formulation (NAK) as: $\widehat{\beta}^j = \underset{\beta,\beta_j=0}{\operatorname{argmin}} \frac{1}{2}||X^j - X\beta||_2^2 + ||\mathbf{r}_j \circ \beta||_1$. NAK is designed for estimating brain connectivity networks from homogeneous samples and incorporate distance knowledge as weight vectors. [1] In experiments, we compare JEEK to NAK (by running NAK R package K times) on multiple synthetic datasets of simulated samples about brain regions. The data simulation strategy was suggested by (Bu & Lederer, 2017). Same as the NAK (Bu & Lederer, 2017), we use the spatial distance among brain regions as additional knowledge in JEEK.

W-SIMULE: (Singh et al., 2017) Like JEEK, one recent study (Singh et al., 2017) of multi-sGGMs (following ideas from (Wang et al., 2017b)) also assumed that $\Omega^{(i)} = \Omega_I^{(i)} + \Omega_S$ and incorporated spatial distance knowledge in their convex formulation for joint discovery of heterogeneous neural connectivity graphs. This study, with name W-SIMULE (Weighted model for Shared and Individual parts of MULtiple graphs Explicitly) uses a weighted constrained ℓ_1 minimization:

$$
\begin{aligned}
&\underset{\Omega_I^{(i)},\Omega_S}{\operatorname{argmin}} \sum_i ||W \circ \Omega_I^{(i)}||_1 + \epsilon K||W \circ \Omega_S||_1 \\
&\text{Subject to: } ||\Sigma^{(i)}(\Omega_I^{(i)} + \Omega_S) - I||_\infty \leq \lambda_n,\ i = 1, \ldots, K
\end{aligned}
\tag{5.1}
$$

[1] Here $\widehat{\beta}^j$ indicates the sparsity of j-th column of a single $\widehat{\Omega}$. Namely, $\widehat{\beta}_k^j = 0$ if and only if $\widehat{\Omega}_{k,j} = 0$. $\mathbf{r}_j$ is a weight vector as the additional knowledge The NAK formulation can be solved by a classic Lasso solver like glmnet.

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

Table 1. Compare JEEK versus baselines. Here T is the number of iterations.

W-SIMULE simply includes the additional knowledge as a weight matrix W. [2]

Different from W-SIMULE, JEEK separates the knowledge of individual context and the shared using different weight matrices. While W-SIMULE also minimizes a weighted $\ell 1$ norm, its constraint optimization term is entirely different from JEEK. The formulation difference makes the optimization of JEEK much faster and more scalable than W-SIMULE (Section (6)). We have provided a complete theoretical analysis of error bounds of JEEK, while W-SIMULE provided no theoretical results. Empirically, we compare JEEK with W-SIMULE R package from (Singh et al., 2017) in the experiments.

JGL: (Danaher et al., 2013): Regularized MLE based multi-sGGMs Studies mostly follow the so called joint graphical lasso (JGL) formulation as Eq. (5.2):

$$\underset{\Omega^{(i)} \succ 0}{\operatorname{argmin}} \sum_{i=1}^{K}(-L(\Omega^{(i)}) + \lambda_n \sum_{i=1}^{K} ||\Omega^{(i)}||_1 + \lambda_n' \mathcal{R}'(\Omega^{(1)}, \Omega^{(2)}, \ldots, \Omega^{(K)}) \tag{5.2}$$

$\mathcal{R}'(\cdot)$ is the second penalty function for enforcing some structural assumption of group property among the multiple graphs. One caveat of JGL is that $\mathcal{R}'(\cdot)$ cannot model explicit additional knowledge. For instance,it can not incorporate the information of a few known hub nodes shared by the contexts. In experiments, we compare JEEK to JGL-co-hub and JGL-perturb-hub toolbox provided by (Mohan et al., 2013).

FASJEM: (Wang et al., 2017a) One very recent study extended JGL using so-called Elementary superposition-structured moment estimator formulation as Eq. (5.3):

$$\begin{aligned} &\underset{\Omega_{tot}}{\operatorname{argmin}} ||\Omega_{tot}||_1 + \epsilon \mathcal{R}'(\Omega_{tot}) \\ &s.t. ||\Omega_{tot} - \text{inv}(T_v(\widehat{\Sigma}_{tot}))||_\infty \leq \lambda_n \\ &\mathcal{R}'^*(\Omega_{tot} - \text{inv}(T_v(\widehat{\Sigma}_{tot}))) \leq \epsilon\lambda_n \end{aligned} \tag{5.3}$$

FASJEM is much faster and more scalable than the JGL estimators. However like JGL estimators it can not model additional knowledge and its optimization needs to be carefully re-designed for different $\mathcal{R}'(\cdot)$. [3]

[2] It can be solved by any linear programming solver and can be column-wise paralleled. However, it is very slow when $p > 200$ due to the expensive computation cost $O(K^4p^5)$.

[3] FASJEM extends JGL into multiple independent group-entry wise optimization just like JEEK. Here $\mathcal{R}'^*(\cdot)$ is the dual norm of $\mathcal{R}'(\cdot)$. Because (Wang et al., 2017a) only designs the optimization of two cases (group,2 and group,inf), we can not use it as a baseline.

Both NAK and W-SIMULE only explored the formulation for estimating neural connectivity graphs using spatial information as additional knowledge. Differently our experiments (Section (6)) extend the weight-as-knowledge formulation on weights as distance, as shared hub knowledge, as perturbed hub knowledge, and as nodes' grouping information (e.g., multiple genes are known to be in the same pathway). This has largely extends the previous studies in showing the real-world adaptivity of the proposed formulation. JEEK elegantly formulates existing knowledge based on the problem at hand and avoid the need to design knowledge-specific optimization.

6 Experiments

We empirically evaluate JEEK and baselines on four types of datasets, including two groups of synthetic data, one real-world fMRI dataset for brain connectivity estimation and one real-world genomics dataset for estimating interaction among regulatory genes (results in Section (6.2)). In order to incorporating various types of knowledge, we provide five different designs of the weight matrices in Section S:5. Details of experimental setup, metrics and hyper-parameter tuning are included in Section (S:6.1). Baselines used in our experiments have been explained in details by Section (5). We also use JEEK with no additional knowledge (JEEK-NK) as a baseline.

JEEK is available as the R package 'jeek' in CRAN.

6.1 Experiment: Simulated Samples with Known Hubs as Knowledge

Inspired the JGL-co-hub and JGL-perturb-hub toolbox (JGL-node) provided by (Mohan et al., 2013), we empirically show JEEK's ability to model known co-hub or perturbed-hub nodes as knowledge when estimating multiple sGGMs. We generate multiple simulated Gaussian datasets through the random graph model (Rothman et al., 2008) to simulate both the co-hub and perturbed-hub graph structures (details in S:7.1). We use JGL-node package, W-SIMULE and JEEK-NK as baselines for this set of experiments. The weights in $\{W_I^{tot}, W_S^{tot}\}$ are designed using the strategy proposed in Section (S:5).

We use AUC score (to reflect the consistency and variance of a method's performance when varying its important hyper-parameter) and computational time cost to compare JEEK with baselines. We compare all methods on many simulated cases by varying p from the set $\{100, 200, 300, 400, 500\}$

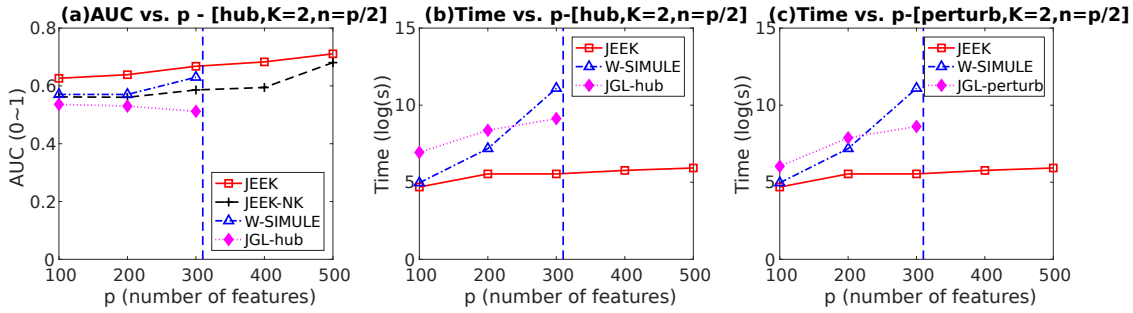

Figure 2. Performance comparison on simulation Datasets using co-Hub Knowledge: AUC vs. Time when varying number of nodes p.

and the number of tasks K from the set $\{2, 3, 4\}$. In Figure 2 and Figure S:1(a)(b), JEEK consistently achieves higher AUC-scores than the baselines JGL, JEEK-NK and W-SIMULE for all cases. JEEK is more than 10 times faster than the baselines on average. In Figure 2, for each $p > 300$ case (with $n = p/2$), W-SIMULE takes more than one month and JGL takes more than one day. Therefore we can not show them with $p > 300$.

6.2 Experiment: Gene Interaction Network from Real-World Genomics Data

Next, we apply JEEK and the baselines on one real-world biomedical data about gene expression profiles across two different cell types. We explored two different types of knowledge: (1) Known edges and (2) Known group about genes. Figure S:1(c) shows that JEEK has lower time cost and recovers more interactions than baselines (higher number of matched edges to the existing bio-databases.). More results are in Appendix Section (S:7.2) and the design of weight matrices for this case is in Section (S:5).

6.3 Experiment: Simulated Data about Brain Connectivity with Distance as Knowledge

Following (Bu & Lederer, 2017), we use one known Euclidean distance between human brain regions as additional knowledge W and use it to generate multiple simulated datasets (details in Section S:7.3). We compare JEEK with the baselines regarding (a) Scalability (computational time cost), and (b) effectiveness (F1-score, because NAK package does not allow AUC calculation). For each simulation case, the computation time for each estimator is the summation of a method's execution time over all values of λ_n. Figure S:2(a)(b) show clearly that JEEK outperforms its baselines. JEEK has a consistently higher F1-Score and is almost 6 times faster than W-SIMULE in the high dimensional case. JEEK performs better than JEEK-NK, confirming the advantage of integrating additional distance knowledge. While NAK is fast, its F1-Score is nearly 0 and hence, not useful for multi-sGGM structure learning.

6.4 Experiment: Functional Connectivity Estimation from Real-World Brain fMRI Data

We evaluate JEEK and relevant baselines for a classification task on one real-world publicly available resting-state fMRI dataset: ABIDE(Di Martino et al., 2014). The ABIDE data aims to understand human brain connectivity and how it reflects neural disorders (Van Essen et al., 2013). ABIDE includes two groups of human subjects: autism and control, and therefore we formulate it as $K = 2$ graph estimation. We utilize the spatial distance between human brain regions as additional knowledge for estimating functional connectivity edges among brain regions. We use Linear Discriminant Analysis (LDA) for a downstream classification task aiming to assess the ability of a graph estimator to learn the differential patterns of the connectome structures. (Details of the ABIDE dataset, baselines, design of the additional knowledge W matrix, cross-validation and LDA classification method are in Section (S:7.4).)

Figure S:2(c) compares JEEK and three baselines: JEEK-NK, W-SIMULE and W-SIMULE with no additional knowledge (W-SIMULE-NK). JEEK yields a classification accuracy of 58.62% for distinguishing the autism subjects versus the control subjects, clearly outperforming JEEK-NK and W-SIMULE-NK. JEEK is roughly 7 times faster than the W-SIMULE estimators, locating at the top left region in Figure S:2(c) (higher classification accuracy and lower time cost). We also experimented with variations of the W matrix and found the classification results are fairly robust to the variations of W (Section (S:7.4)).

7 Conclusions

We propose a novel method, JEEK, to incorporate additional knowledge in estimating multi-sGGMs. JEEK achieves the same asymptotic convergence rate as the state-of-the-art. Our experiments has showcased using weights for describing pairwise knowledge among brain regions, for shared hub knowledge, for perturbed hub knowledge, for describing group information among nodes (e.g., genes known to be in the same pathway), and for using known interaction edges as the knowledge.

References

Banerjee, O., El Ghaoui, L., and d'Aspremont, A. Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *The Journal of Machine Learning Research*, 9:485–516, 2008.

Bu, Y. and Lederer, J. Integrating additional knowledge into estimation of graphical models. *arXiv preprint arXiv:1704.02739*, 2017.

Caruana, R. Multitask learning. *Machine learning*, 28(1): 41–75, 1997.

Chiquet, J., Grandvalet, Y., and Ambroise, C. Inferring multiple graphical structures. *Statistics and Computing*, 21(4):537–553, 2011.

Consortium, E. P. et al. An integrated encyclopedia of dna elements in the human genome. *Nature*, 489(7414):57–74, 2012.

Danaher, P., Wang, P., and Witten, D. M. The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2013.

Di Martino, A., Yan, C.-G., Li, Q., Denio, E., Castellanos, F. X., Alaerts, K., Anderson, J. S., Assaf, M., Bookheimer, S. Y., Dapretto, M., et al. The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. *Molecular psychiatry*, 19(6): 659–667, 2014.

Guo, J., Levina, E., Michailidis, G., and Zhu, J. Joint estimation of multiple graphical models. *Biometrika*, pp. asq060, 2011.

Honorio, J. and Samaras, D. Multi-task learning of gaussian graphical models. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pp. 447–454, 2010.

Ideker, T. and Krogan, N. J. Differential network biology. *Molecular systems biology*, 8(1):565, 2012.

Kelly, C., Biswal, B. B., Craddock, R. C., Castellanos, F. X., and Milham, M. P. Characterizing variation in the functional connectome: promise and pitfalls. *Trends in cognitive sciences*, 16(3):181–188, 2012.

Lauritzen, S. L. *Graphical models*, volume 17. Clarendon Press, 1996.

Mohan, K., London, P., Fazel, M., Lee, S.-I., and Witten, D. Node-based learning of multiple gaussian graphical models. *arXiv preprint arXiv:1303.5145*, 2013.

Monti, R. P., Anagnostopoulos, C., and Montana, G. Learning population and subject-specific brain connectivity networks via mixed neighborhood selection. *arXiv preprint arXiv:1512.01947*, 2015.

Negahban, S., Yu, B., Wainwright, M. J., and Ravikumar, P. K. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. In *Advances in Neural Information Processing Systems*, pp. 1348–1356, 2009.

Network, C. G. A. R. et al. Integrated genomic analyses of ovarian carcinoma. *Nature*, 474(7353):609–615, 2011.

Pang, H., Liu, H., and Vanderbei, R. The fastclime package for linear programming and large-scale precision matrix estimation in r. *Journal of Machine Learning Research*, 15:489–493, 2014.

Poldrack, R. A., Barch, D. M., Mitchell, J., Wager, T., Wagner, A. D., Devlin, J. T., Cumba, C., Koyejo, O., and Milham, M. Toward open sharing of task-based fmri data: the openfmri project. *Frontiers in neuroinformatics*, 7:12, 2013.

Prasad, T. K., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., Telikicherla, D., Raju, R., Shafreen, B., Venugopal, A., et al. Human protein reference database?2009 update. *Nucleic acids research*, 37 (suppl 1):D767–D772, 2009.

Rothman, A. J., Bickel, P. J., Levina, E., Zhu, J., et al. Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2:494–515, 2008.

Rothman, A. J., Levina, E., and Zhu, J. Generalized thresholding of large covariance matrices. *Journal of the American Statistical Association*, 104(485):177–186, 2009.

Shimamura, T., Imoto, S., Yamaguchi, R., and Miyano, S. Weighted lasso in graphical gaussian modeling for large gene network estimation based on microarray data. In *Genome Informatics 2007: Genome Informatics Series Vol. 19*, pp. 142–153. World Scientific, 2007.

Singh, C., Wang, B., and Qi, Y. A constrained, weighted-l1 minimization approach for joint discovery of heterogeneous neural connectivity graphs. *arXiv preprint arXiv:1709.04090*, 2017.

Uddin, L. Q., Supekar, K., Lynch, C. J., Khouzam, A., Phillips, J., Feinstein, C., Ryali, S., and Menon, V. Salience network–based classification and prediction of symptom severity in children with autism. *JAMA psychiatry*, 70(8):869–879, 2013.

Van Essen, D. C., Smith, S. M., Barch, D. M., Behrens, T. E., Yacoub, E., Ugurbil, K., Consortium, W.-M. H., et al.

The wu-minn human connectome project: an overview. *Neuroimage*, 80:62–79, 2013.

Wang, B., Gao, J., and Qi, Y. A fast and scalable joint estimator for learning multiple related sparse gaussian graphical models. In *Artificial Intelligence and Statistics*, pp. 1168–1177, 2017a.

Wang, B., Singh, R., and Qi, Y. A constrained l1 minimization approach for estimating multiple sparse gaussian or nonparanormal graphical models. *Machine Learning*, 106 (9-10):1381–1417, 2017b.

Watts, D. J. and Strogatz, S. H. Collective dynamics of 'small-world'networks. *nature*, 393(6684):440–442, 1998.

Yang, E., Lozano, A., and Ravikumar, P. Elementary estimators for high-dimensional linear regression. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pp. 388–396, 2014a.

Yang, E., Lozano, A. C., and Ravikumar, P. Elementary estimators for graphical models. In *Advances in Neural Information Processing Systems*, pp. 2159–2167, 2014b.

Yang, E., Lozano, A. C., and Ravikumar, P. Elementary estimators for sparse covariance matrices and other structured moments. In *ICML*, pp. 397–405, 2014c.

Yuan, M. and Lin, Y. Model selection and estimation in the gaussian graphical model. *Biometrika*, 94(1):19–35, 2007.

Zhang, B. and Wang, Y. Learning structural changes of gaussian graphical models in controlled experiments. *arXiv preprint arXiv:1203.3532*, 2012.

Zhang, Y. and Schneider, J. G. Learning multiple tasks with a sparse matrix-normal penalty. In *Advances in Neural Information Processing Systems*, pp. 2550–2558, 2010.

Zhu, Y., Shen, X., and Pan, W. Structural pursuit over multiple undirected graphs. *Journal of the American Statistical Association*, 109(508):1683–1696, 2014.