




Inter-Subject Analysis: A Partial Gaussian Graphical Model Approach

Cong Ma , Junwei Lu & Han Liu


To cite this article: Cong Ma , Junwei Lu & Han Liu (2020): Inter-Subject Analysis: A Partial Gaussian Graphical Model Approach, Journal of the American Statistical Association, DOI: [10.1080/01621459.2020.1841645](https://doi.org/10.1080/01621459.2020.1841645)

To link to this article: <https://doi.org/10.1080/01621459.2020.1841645>

 View supplementary material [↗](#)

 Published online: 17 Dec 2020.

 Submit your article to this journal [↗](#)

 Article views: 187

 View related articles [↗](#)

 View Crossmark data [↗](#)



Inter-Subject Analysis: A Partial Gaussian Graphical Model Approach

Cong Ma^a, Junwei Lu^b, and Han Liu^c

^aDepartment of Operations Research and Financial Engineering, Princeton University, Princeton, NJ; ^bDepartment of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA; ^cDepartment of Computer Science and Department of Statistics, Northwestern University, Evanston, IL

ABSTRACT

Different from traditional intra-subject analysis, the goal of inter-subject analysis (ISA) is to explore the dependency structure between different subjects with the intra-subject dependency as nuisance. ISA has important applications in neuroscience to study the functional connectivity between brain regions under natural stimuli. We propose a modeling framework for ISA that is based on Gaussian graphical models, under which ISA can be converted to the problem of estimation and inference of a partial Gaussian graphical model. The main statistical challenge is that we do not impose sparsity constraints on the whole precision matrix and we only assume the inter-subject part is sparse. For estimation, we propose to estimate an alternative parameter to get around the nonsparse issue and it can achieve asymptotic consistency even if the intra-subject dependency is dense. For inference, we propose an “untangle and chord” procedure to de-bias our estimator. It is valid without the sparsity assumption on the inverse Hessian of the log-likelihood function. This inferential method is general and can be applied to many other statistical problems, thus it is of independent theoretical interest. Numerical experiments on both simulated and brain imaging data validate our methods and theory. Supplementary materials for this article are available online.

ARTICLE HISTORY

Received September 2017
Accepted October 2020

KEYWORDS

fMRI data; Gaussian graphical models; Nuisance parameter; Sample splitting; Uncertainty assessment

1. Introduction

Inter-subject analysis (ISA) refers to the inference of the dependency structures between different subjects with intra-subject dependencies as nuisance. The subject may be a pathway consisting of an assembly of genes or a group of stocks from the same sector in financial markets. Often the dependency structure between different subjects is of scientific interest while the dependencies within each subject are complicated and hard to infer. The goal of ISA is to explore the inter-subject dependencies with intra-subject dependencies as nuisance.

1.1. Motivating Example

To motivate the use of ISA, we consider the functional magnetic resonance imaging (fMRI) data analysis. fMRI provides scientists a noninvasive way to observe the neural activity in the human brain (Lindquist 2008). Traditionally, fMRI measurements are obtained under highly controlled experimental settings where subjects are asked to perform identical and demanding attention tasks. Recent studies show that neuronal responses and brain activities are more reliable under naturalistic stimuli, for instance, watching a movie episode or listening to an audiobook (Mechler et al. 1998; Yao et al. 2007; Belitski et al. 2008). This motivates the use of fMRI under more naturalistic settings (Zacks et al. 2001; Hartley et al. 2003; Bartels and Zeki 2004). However, this brings substantial noise to the fMRI measurements since individual cognitive processes that are not related to the ongoing stimuli cannot be constrained or removed as in controlled research settings (Hasson et al. 2004;

Simony et al. 2016). Conventional intra-subject analysis which computes voxel-by-voxel correlations in the same subject can be influenced by such noise and fail to detect the stimulus-induced correlations.

Hasson et al. (2004) introduced inter-subject correlations (ISC) to partially resolve this problem. Instead of computing the intra-subject correlations, ISC calculates the correlation coefficients of corresponding voxels across different experimental subjects (see Figure 1). It is based on the assumption that individual variations are uncorrelated across subjects and high ISC indicate stimulus related activations. Although ISC can isolate the individual noise, as a measure of marginal dependence, it fails to eliminate the confounding effects of other factors (Horwitz and Rapoport 1988; Lee et al. 2011). Conditional dependence has long been studied to remedy this problem in both statistics and biology community (Marrelec et al. 2006; Huang et al. 2010; Varoquaux et al. 2012).

1.2. Modeling Framework

In this article, we propose a new modeling framework named ISA to infer the conditional dependency between different subjects. Formally, let $X = (X_1, \dots, X_d)^\top \sim N(0, \Sigma^*)$ be a d -dimensional Gaussian random vector with precision matrix $\Omega^* = (\omega_{jk}^*)$. Let \mathcal{G}_1 and \mathcal{G}_2 be two disjoint subsets of $\{1, 2, \dots, d\}$ with cardinality $|\mathcal{G}_1| = d_1$ and $|\mathcal{G}_2| = d_2 := d - d_1$. We use $X_{\mathcal{G}_1}$ and $X_{\mathcal{G}_2}$ to denote the corresponding subvectors of X and they represent features of two different subjects. We use Σ_1^* , Σ_2^* , and Σ_{12}^* to denote the covariance within $X_{\mathcal{G}_1}$, within

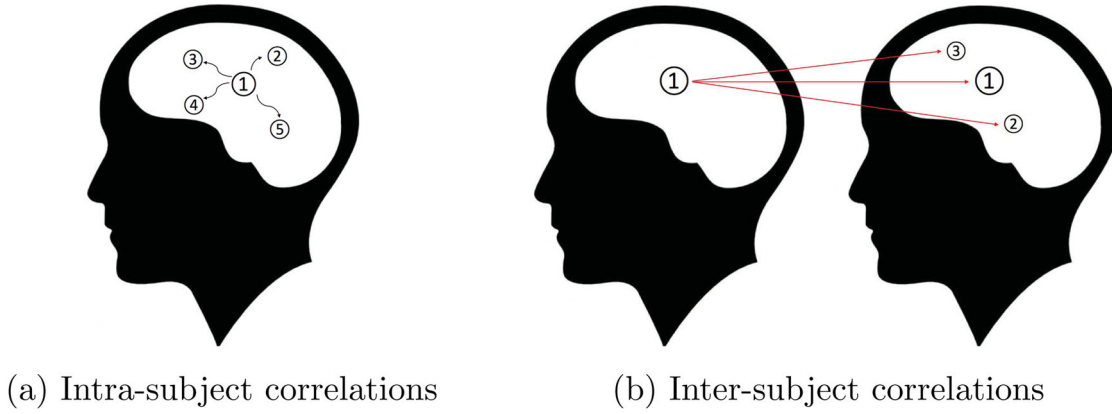


Figure 1. (a) The intra-subject correlations where the correlations among voxels in the same subject are calculated. (b) The inter-subject correlations where the correlations of voxels are computed across subjects.

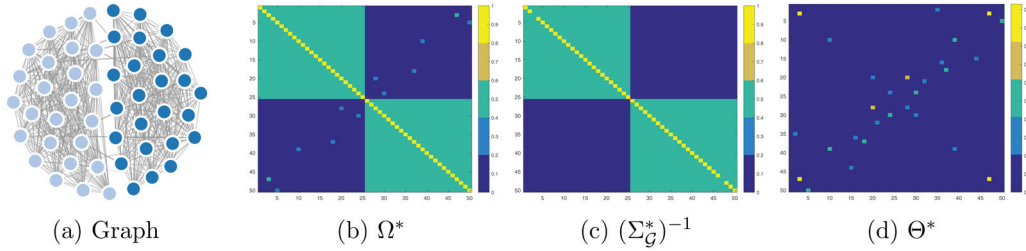


Figure 2. (a) A Gaussian graphical model with two subjects, where each color represents a subject. It can be seen that the inter-subject connections are sparse while the connections within each subject are dense. (b) The heatmap of the precision matrix of the Gaussian graphical model. (c) The heatmap of the corresponding $(\Sigma_{\mathcal{G}}^*)^{-1}$. (d) The heatmap of $\Theta^* = \Omega^* - (\Sigma_{\mathcal{G}}^*)^{-1}$. As can be seen, Θ^* is sparse even if both Ω_1^* and Ω_2^* are dense.

$X_{\mathcal{G}_2}$, and between $X_{\mathcal{G}_1}$ and $X_{\mathcal{G}_2}$, respectively. The submatrices Ω_1^* , Ω_2^* , and Ω_{12}^* are defined similarly. It is well known that X_j and X_k are conditionally independent given the remaining variables if and only if $\omega_{jk}^* = 0$. Our modeling assumption is that Ω_{12}^* purely represents the dependency driven by the common stimuli while Ω_1^* and Ω_2^* can be influenced by the individual cognitive process. Hence, we are willing to assume that Ω_{12}^* is sparse while reluctant to put any sparsity constraint on Ω_1^* or Ω_2^* . The main statistical challenge we address in this article is to obtain estimation consistency and valid inference of Ω_{12}^* under nonsparse nuisance parameters Ω_1^* and Ω_2^* .

1.3. Contributions

There are two major contributions of this article.

Our first contribution is a new estimator for Ω_{12}^* when the precision matrix Ω^* is not sparse. The idea is to find an alternative sparse matrix to estimate. The candidate we propose is $\Theta^* := \Omega^* - (\Sigma_{\mathcal{G}}^*)^{-1}$, where $\Sigma_{\mathcal{G}}^* = \text{diag}(\Sigma_1^*, \Sigma_2^*)$ and $\text{diag}(A, B)$ denotes the block diagonal matrix whose diagonals are A and B . The key observation is that if Ω_{12}^* is sparse, then Θ^* is also sparse. More precisely, we have $\|\Theta^*\|_0 \leq 2s^2 + 2s$ whenever $\|\Omega_{12}^*\|_0 \leq s$, where $\|A\|_0$ counts the number of nonzero entries in A .¹ This observation holds true even if both Ω_1^* and Ω_2^* are dense. We illustrate this phenomenon using a numerical example in Figure 2. Following this observation, we can reparameterize the precision matrix using $\Omega^* = \Theta^* +$

$(\Sigma_{\mathcal{G}}^*)^{-1}$, in which Θ^* contains the parameter of interest (as Θ^* 's off-diagonal block $\Theta_{12}^* = \Omega_{12}^*$) and $(\Sigma_{\mathcal{G}}^*)^{-1}$ is a nuisance parameter. We then propose to estimate Ω_{12}^* by minimizing an ℓ_1 regularized pseudo-likelihood with respect to Θ . This estimator, which is named sparse edge estimator for intense nuisance graphs (STRINGS) achieves consistency under mild conditions.

Our second contribution is to propose a general ‘‘untangle and chord’’ procedure to de-bias high-dimensional estimators when the inverse Hessian of the log-likelihood function is not sparse. In general, a de-biased estimator $\hat{\beta}^u$ takes the following form:

$$\hat{\beta}^u = \hat{\beta} - M \nabla \mathcal{L}_n(\hat{\beta}),$$

where $\hat{\beta}$ is the regularized estimator for the parameter of interest β^* , M is a bias correction matrix, and \mathcal{L}_n denotes the negative log-likelihood function. By the Taylor expansion of $\mathcal{L}_n(\hat{\beta})$ at the true parameter β^* , we have

$$\begin{aligned} \sqrt{n} \cdot (\hat{\beta}^u - \beta^*) &\approx -\sqrt{n} M \nabla \mathcal{L}_n(\beta^*) - \sqrt{n} [M \nabla^2 \mathcal{L}_n(\beta^*) - I] \\ &\quad \times (\hat{\beta} - \beta^*). \end{aligned}$$

Clearly, the leading term $\sqrt{n} M \nabla \mathcal{L}_n(\beta^*)$ is asymptotically normal under mild conditions. And if the remainder term $\sqrt{n} [M \nabla^2 \mathcal{L}_n(\beta^*) - I] (\hat{\beta} - \beta^*)$ converges to 0 in probability, we can conclude that $\hat{\beta}^u$ is asymptotically normal. One way to achieve this goal is to let M be a consistent estimator of the inverse of $\nabla^2 \mathcal{L}_n(\beta^*)$. This is why previous inferential methods require the sparsity assumption on the inverse Hessian.

¹See Appendix in the supplementary materials for a proof of this observation.

It will be shown that the Hessian in ISA is not necessarily sparse. To get around this issue, two crucial observations are needed. First, it suffices to use constrained optimization to control the statistical error of $M\nabla^2\mathcal{L}_n(\beta^*) - I$. Second, to prevent M from tangling with $\nabla\mathcal{L}_n(\beta^*)$ and sabotaging the asymptotic normality of the leading term, we can split the data into two parts: the untangle step estimates the parameter on the first split and the chord step constructs M using the second split. We show that the “untangle and chord” procedure debiases the STRINGS estimator and we can construct valid tests and confidence intervals based on the de-biased estimator. The “untangle and chord” strategy is general and can be applied to many other high-dimensional problems without the sparsity assumption on the inverse Hessian.

1.4. Related Work

Estimators for the precision matrix including the maximum likelihood estimator and the column-wise estimator have been considered in Yuan and Lin (2007), Banerjee, Ghaoui, and d’Aspremont (2008), Rothman et al. (2008), Friedman, Hastie, and Tibshirani (2008), Ravikumar et al. (2011), Meinshausen and Bühlmann (2006), Yuan (2010), and Cai, Liu, and Luo (2011). All of them require the sparsity of the whole precision matrix. Hence, they are not applicable to ISA.

Inferential methods based on inverting KKT conditions (van de Geer et al. 2014; Zhang and Zhang 2014) and decorrelated score functions (Ning and Liu 2017) have been extended to Gaussian graphical models (Gu et al. 2015; Jankova and van de Geer 2015). They all require the inverse Hessian of the log-likelihood function to be sparse and hence cannot be applied in our setting. One exception is the inference for the high-dimensional linear model proposed by Javanmard and Montanari (2014). Their result heavily depends on the special structure of the linear model. First, the design matrix is independent of the noise and second, the inverse Hessian matrix is simply the inverse covariance of the design, which is irrelevant to the regression parameters. Their method is difficult to extend to general estimators.

Efforts have also been made to relax the sparsity assumption on the precision matrix in Gaussian graphical model estimation. Yuan and Zhang (2014) proposed a decoupling approach to estimate Ω_{12}^* . However, their method requires at least one of the diagonal blocks Ω_1^* or Ω_2^* to be sparse. And it is no longer valid if both of them are dense. Liu et al. (2016) shared the similar goal with ours. They proposed a density ratio framework to estimate the dependency between two groups of variables. First, their estimation theory does not apply to Gaussian distributions due to the unboundedness of the density ratio. Second, their procedure relies on approximating the normalization function using two sample U -statistics which are complicated and computationally expensive. Compared with the above works, our work not only considers the estimation consistency but also proposes valid procedures to assess uncertainty in the high-dimensional setting.

1.5. Notation

The following notations are used throughout the article. For any $n \in \mathbb{N}$ we use the shorthand notation $[n] = \{1, \dots, n\}$. For a

vector $v = (v_1, \dots, v_d)^\top \in \mathbb{R}^d$, let $\|v\|_q = (\sum_{i=1}^d v_i^q)^{1/q}$, $1 \leq q < \infty$. Furthermore, let $\|v\|_\infty = \max_j |v_j|$. For a matrix $A = (A_{jk}) \in \mathbb{R}^{m \times n}$, we define $\text{supp}(A) = \{(j, k) | A_{jk} \neq 0\}$. We use A_{j*} and A_{*k} to denote the j th row and k th column of A , respectively. We use $\|A\|_q = \sup_{\|x\|_q=1} \|Ax\|_q$ to denote the induced ℓ_q -norm of a matrix. In particular, $\|A\|_1 = \max_{1 \leq k \leq n} \sum_{j=1}^m |A_{jk}|$, which is the maximum absolute column sum of the matrix A . $\|A\|_2$ is the largest singular value of A . $\|A\|_\infty = \max_{1 \leq j \leq m} \sum_{k=1}^n |A_{jk}|$, which is the maximum absolute row sum of the matrix A . We also use $\|A\|_{\max} = \max_{jk} |A_{jk}|$, $\|A\|_{1,1} = \sum_{jk} |A_{jk}|$, and $\|A\|_F = (\sum_{jk} A_{jk}^2)^{1/2}$ to denote the ℓ_{\max} -, $\ell_{1,1}$ -, and ℓ_F -norms of the matrix A . $\lambda_{\min}(A)$ is used to denote the minimum eigenvalue of the matrix A and $|A|$ is used to denote the determinant of A . We use $\Phi(x)$ to denote the cumulative distribution function of a standard normal random variable. For a sequence of random variables X_n , we write $X_n \rightsquigarrow X$, for some random variable X , if X_n converges in distribution to X .

2. The STRINGS Estimator

In this section, we present the STRINGS estimator for the inter-subject precision matrix Ω_{12}^* . The basic idea is to use the maximum likelihood principle. Given a data matrix $\mathbb{X} \in \mathbb{R}^{n \times d}$, where rows of \mathbb{X} represent iid samples from a Gaussian distribution with mean 0 and covariance Σ^* , the negative log-likelihood for the precision matrix is given by

$$\mathcal{L}(\Omega) = \text{Tr}(\Omega \hat{\Sigma}) - \log |\Omega|, \tag{1}$$

where $\hat{\Sigma} = (1/n) \cdot \mathbb{X}^\top \mathbb{X}$ is the sample covariance matrix.

Since our focus is on estimating the inter-subject dependency Ω_{12}^* , a naive reparameterization of the precision matrix is $\Omega^* = \Omega_D^* + \Omega_O^*$, where Ω_D^* is the block diagonal matrix corresponding to X_{G_1} and X_{G_2} , that is, $\Omega_D^* = \text{diag}(\Omega_{G_1}^*, \Omega_{G_2}^*)$. And Ω_O^* is the off-diagonal part involving Ω_{12}^* and $\Omega_{12}^{*\top}$. Under such a reparameterization, we can reformulate (1) as

$$\mathcal{L}(\Omega_O, \Omega_D) = \text{Tr}[(\Omega_O + \Omega_D) \hat{\Sigma}] - \log |\Omega_O + \Omega_D|. \tag{2}$$

Adopting the maximum likelihood principle, we want to minimize the negative log-likelihood (2) with respect to the parameter Ω_O . Hence, we can ignore the terms independent of Ω_O in (2). This gives us an equivalent minimization of $\text{Tr}(\Omega_O \hat{\Sigma}) - \log |\Omega_O + \Omega_D|$ with respect to Ω_O . However, the objective function still depends on the nuisance parameter Ω_D and it is difficult to obtain an estimator for Ω_D . Thus, this naive reparameterization will not work.

Recall that if Ω_{12}^* is s -sparse, then $\Theta^* = \Omega^* - (\Sigma_G^*)^{-1}$ is $(2s^2 + 2s)$ -sparse. Based on this key observation, we reparameterize the precision matrix using Θ^* and $(\Sigma_G^*)^{-1}$, in which Θ^* contains the parameter of interest (as $\Theta_{12}^* = \Omega_{12}^*$) and $(\Sigma_G^*)^{-1}$ is the nuisance parameter. Under the new reparameterization $\Omega^* = \Theta^* + (\Sigma_G^*)^{-1}$, we can rewrite (1) as

$$\mathcal{L}(\Theta, \Sigma_G^{-1}) = \text{Tr}[(\Theta + \Sigma_G^{-1}) \hat{\Sigma}] - \log |\Theta + \Sigma_G^{-1}|. \tag{3}$$

Using the fact that

$$\begin{aligned} \log |\Theta + \Sigma_G^{-1}| &= \log |\Sigma_G^{-1}(\Sigma_G \Theta \Sigma_G + \Sigma_G) \Sigma_G^{-1}| \\ &= \log |\Sigma_G^{-1}| + \log |\Sigma_G \Theta \Sigma_G + \Sigma_G| \\ &\quad + \log |\Sigma_G^{-1}|, \end{aligned}$$

we can further decompose (3) into the following form:

$$\begin{aligned} \mathcal{L}(\Theta, \Sigma_{\mathcal{G}}^{-1}) &= \text{Tr}(\Theta \hat{\Sigma}) + \text{Tr}(\Sigma_{\mathcal{G}}^{-1} \hat{\Sigma}) - \log |\Sigma_{\mathcal{G}}^{-1}| \\ &\quad - \log |\Sigma_{\mathcal{G}} \Theta \Sigma_{\mathcal{G}} + \Sigma_{\mathcal{G}}| - \log |\Sigma_{\mathcal{G}}^{-1}|. \end{aligned} \quad (4)$$

Ignoring the terms independent of Θ in (4), we have that minimizing (4) with respect to Θ is equivalent to minimizing $\text{Tr}(\Theta \hat{\Sigma}) - \log |\Sigma_{\mathcal{G}} \Theta \Sigma_{\mathcal{G}} + \Sigma_{\mathcal{G}}|$ with respect to Θ . Now we still have the nuisance parameter $\Sigma_{\mathcal{G}}$. However, in this case, we can use the naive plug-in estimator for $\Sigma_{\mathcal{G}}$, that is, $\hat{\Sigma}_{\mathcal{G}}$ which is the block diagonal matrix of $\hat{\Sigma}$ corresponding to $X_{\mathcal{G}_1}$ and $X_{\mathcal{G}_2}$. This gives us the following empirical loss function for Θ :

$$\mathcal{L}_n(\Theta) = \text{Tr}(\Theta \hat{\Sigma}) - \log |\hat{\Sigma}_{\mathcal{G}} \Theta \hat{\Sigma}_{\mathcal{G}} + \hat{\Sigma}_{\mathcal{G}}|. \quad (5)$$

Correspondingly, we will use $\mathcal{L}(\Theta) = \text{Tr}(\Theta \Sigma^*) - \log |\Sigma_{\mathcal{G}}^* \Theta \Sigma_{\mathcal{G}}^* + \Sigma_{\mathcal{G}}^*|$ to denote the population loss function. Since Θ^* is sparse, we further impose a sparsity penalty on the objective function. Here we choose the $\ell_{1,1}$ penalty, and the STRINGS estimator has the following form

$$\hat{\Theta} = \arg \min \mathcal{L}_n(\Theta) + \lambda \|\Theta\|_{1,1}. \quad (6)$$

Note that for the empirical loss function $\mathcal{L}_n(\Theta)$ in (5) to be convex, $\hat{\Sigma}_{\mathcal{G}}$ needs to be positive definite. Otherwise, the log-determinant term will always be $-\infty$. However, in the high-dimensional regime, the naive plug-in estimator $\hat{\Sigma}_{\mathcal{G}}$ will be rank deficient. To resolve this issue, we can perturb our plug-in estimator with $\sqrt{\log d/n} \cdot I$. This perturbation trick has also been used to solve the initialization problem in Cai, Liu, and Luo (2011). We choose the size of the perturbation to be $\sqrt{\log d/n}$ so that it will not affect the concentration property of the estimator $\hat{\Sigma}_{\mathcal{G}}$. Although (6) is a convex program, solving it in high dimensions using semidefinite programming is both time-consuming and memory-consuming. We propose a computationally efficient algorithm based on alternating direction method of multipliers (ADMM) to solve (6). The details are deferred to Appendix Section B in the supplementary materials.

3. “Untangle and Chord” the STRINGS

In this section, we introduce our proposed method to test the existence of certain inter-subject interaction and construct a confidence interval for entries in Ω_{12}^* . Formally, for $1 \leq j \leq d_1$ and $d_1 + 1 \leq k \leq d$, we are interested in the following two types of inferential problems:

- Confidence interval: For a particular parameter θ_{jk}^* , where θ_{jk}^* is the (j, k) th entry of Θ^* , how to construct a confidence interval for it?
- Hypothesis testing: Consider the null hypothesis $H_0 : \theta_{jk}^* = 0$, how to construct a valid test for H_0 ?

To address these two types of questions, we rely on obtaining an asymptotically normal estimator of θ_{jk}^* . After this, constructing confidence intervals and testing hypotheses follow naturally. Hence in the following we introduce our way to de-bias the STRINGS estimator.

As we mentioned in Section 1, KKT-inversion type of methods (van de Geer et al. 2014; Zhang and Zhang 2014) cannot be

applied here since they require the inverse Hessian to be sparse. Recall that the population loss function is given by $\mathcal{L}(\Theta) = \text{Tr}(\Theta \Sigma^*) - \log |\Sigma_{\mathcal{G}}^* \Theta \Sigma_{\mathcal{G}}^* + \Sigma_{\mathcal{G}}^*|$. Its Hessian can be calculated as following:

$$\nabla^2 \mathcal{L}(\Theta) = [(\Sigma_{\mathcal{G}}^* \Theta + I)^{-1} \Sigma_{\mathcal{G}}^*] \otimes [(\Sigma_{\mathcal{G}}^* \Theta + I)^{-1} \Sigma_{\mathcal{G}}^*].$$

Thus, we have $[\nabla^2 \mathcal{L}(\Theta^*)]^{-1} = \Omega^* \otimes \Omega^*$. We can see that the inverse Hessian can be dense since we do not impose any assumption on Ω^* . Getting around with this difficulty requires new sets of inferential tools. Rather than inverting the KKT conditions, we propose to de-bias the STRINGS estimator in (6) utilizing the estimating equation for Θ^* . Moreover, sample splitting is adopted to achieve the desired asymptotic normality. To see this, recall the definition of Θ^* that $\Theta^* = \Omega^* - (\Sigma_{\mathcal{G}}^*)^{-1}$, we can derive the following estimating equation:

$$\Sigma^* \Theta^* \Sigma_{\mathcal{G}}^* + \Sigma^* - \Sigma_{\mathcal{G}}^* = 0. \quad (7)$$

We first present a heuristic explanation on our debiasing procedure. Based on the sample version of (7), we construct a de-biased estimator as following

$$\hat{\Theta}^u = \hat{\Theta} - M(\hat{\Sigma} \hat{\Theta} \hat{\Sigma}_{\mathcal{G}} + \hat{\Sigma} - \hat{\Sigma}_{\mathcal{G}}) P^{\top}, \quad (8)$$

where M and P are two bias correction matrices to be specified later. To gain the intuition why $\hat{\Theta}^u$ defined in (8) is an asymptotically normal estimator, we calculate the difference between $\hat{\Theta}^u$ and Θ^* as follows.

$$\begin{aligned} \hat{\Theta}^u - \Theta^* &= \hat{\Theta} - \Theta^* - M[\hat{\Sigma}(\hat{\Theta} - \Theta^* + \Theta^*) \hat{\Sigma}_{\mathcal{G}} \\ &\quad + \hat{\Sigma} - \hat{\Sigma}_{\mathcal{G}}] P^{\top} \\ &= -M(\hat{\Sigma} \Theta^* \hat{\Sigma}_{\mathcal{G}} + \hat{\Sigma} - \hat{\Sigma}_{\mathcal{G}}) P^{\top} + \hat{\Theta} - \Theta^* \\ &\quad - M \hat{\Sigma}(\hat{\Theta} - \Theta^*) \hat{\Sigma}_{\mathcal{G}} P^{\top}. \end{aligned}$$

Through some algebra, we have $\hat{\Theta}^u - \Theta^* = \text{Leading} + \text{Remainder}$, where

$$\begin{aligned} \text{Leading} &= -M[(\hat{\Sigma} - \Sigma^*)(I + \Theta^* \Sigma_{\mathcal{G}}^*) \\ &\quad - (I - \Sigma^* \Theta^*)(\hat{\Sigma}_{\mathcal{G}} - \Sigma_{\mathcal{G}}^*)] P^{\top}, \end{aligned} \quad (9)$$

$$\begin{aligned} \text{Remainder} &= -M(\hat{\Sigma} - \Sigma^*) \Theta^* (\hat{\Sigma}_{\mathcal{G}} - \Sigma_{\mathcal{G}}^*) P^{\top} + \hat{\Theta} - \Theta^* \\ &\quad - M \hat{\Sigma}(\hat{\Theta} - \Theta^*) \hat{\Sigma}_{\mathcal{G}} P^{\top}. \end{aligned} \quad (10)$$

First, to make the Remainder term in (10) small, it requires $M \hat{\Sigma} \approx I$ and $P \hat{\Sigma}_{\mathcal{G}} \approx I$. In other words, M and P should function as the inverse of Σ^* and $\Sigma_{\mathcal{G}}^*$, respectively. Second, regarding the Leading term, we can see that it is an empirical process type quantity. It is asymptotically normal provided that M and P are independent of the remaining random quantities in (9). This motivates us to utilize sample splitting to obtain the two bias correction matrices M and P . In all, we have an “untangle and chord” procedure to de-bias the STRINGS estimator. Concretely, we split the data $\mathbb{X} \in \mathbb{R}^{2n \times d}$ into two parts \mathcal{D}_1 and \mathcal{D}_2 with equal number of samples. Note that here we inflate the sample size to $2n$. This is purely for simplifying the notations. The untangle step uses the first data \mathcal{D}_1 to get an initial STRINGS estimator $\hat{\Theta}$. The chord step utilizes the second data \mathcal{D}_2 to obtain the bias correction matrices M and P with desired properties, that is, $M \hat{\Sigma} \approx I$ and $P \hat{\Sigma}_{\mathcal{G}} \approx I$. Precisely we

use a CLIME-type procedure to get M and P . For $M \in \mathbb{R}^{d \times d}$, we solve the following convex program:

$$\begin{aligned} \min \quad & \|M\|_\infty \\ \text{subject to} \quad & \|M\hat{\Sigma}' - I\|_{\max} \leq \lambda', \end{aligned} \quad (11)$$

where $\hat{\Sigma}'$ is the sample covariance matrix of the second sample \mathcal{D}_2 and λ' is the approximation error we want to achieve. $\hat{\Sigma}'$ and λ' can be viewed as two inputs to this CLIME-type procedure. We solve a similar convex problem to obtain $P \in \mathbb{R}^{d \times d}$ with different inputs and an additional block constraint:

$$\begin{aligned} \min \quad & \|P\|_\infty \\ \text{subject to} \quad & \|P\hat{\Sigma}'_{\mathcal{G}} - I\|_{\max} \leq \lambda', \\ & P = \text{diag}(P_1, P_2), P_1 \in \mathbb{R}^{d_1 \times d_1}, P_2 \in \mathbb{R}^{d_2 \times d_2}, \end{aligned} \quad (12)$$

where $\hat{\Sigma}'_{\mathcal{G}}$ is the block diagonal sample covariance matrix corresponding to $X_{\mathcal{G}_1}$ and $X_{\mathcal{G}_2}$ on the second sample \mathcal{D}_2 . Notice here we add another constraint that P needs to be a block diagonal matrix. This is expectable since $\hat{\Sigma}'_{\mathcal{G}}$ is a block diagonal matrix.

Given the de-biased estimator $\hat{\Theta}^u = (\hat{\theta}_{jk}^u)$ in (8), we can obtain a confidence interval for θ_{jk}^* and conduct hypothesis testing on $H_0 : \theta_{jk}^* = 0$ under a valid estimation of the asymptotic variance of $\hat{\theta}_{jk}^u$. The asymptotic variance is involved, and we defer the details to Section 4.2.

4. Theoretical Results

4.1. Estimation Consistency

In the next theorem, we give the convergence rate for the STRINGS estimator in (6).

Theorem 4.1. Suppose the inter-subject dependencies $\|\Omega_{12}^*\|_0 \leq s$ and hence $\|\Theta^*\|_0 \leq s^* := 2s^2 + 2s$. Further we assume $\|\Sigma^*\|_{\max} \leq K$ and $\|\Theta^*\|_{1,1} \leq R$ for some absolute constants K and R . Under the sample size condition that $s^{*2} \sqrt{\log d/n} = o(1)$, there exists a constant $C > 0$ such that for sufficiently large n , if $\lambda = 2C\sqrt{\log d/n}$, with probability at least $1 - 4d^{-1}$, we have

$$\begin{aligned} \|\hat{\Omega}_{12} - \Omega_{12}^*\|_F &\leq \frac{14C}{\rho_{\min}^2} \sqrt{\frac{s^* \log d}{n}} \quad \text{and} \\ \|\hat{\Omega}_{12} - \Omega_{12}^*\|_{1,1} &\leq \frac{56C}{\rho_{\min}^2} s^* \sqrt{\frac{\log d}{n}}, \end{aligned}$$

provided that $\lambda_{\min}(\Sigma^*) \geq \rho_{\min} > 0$.

A few remarks on the assumptions are in order. First, $\|\Omega_{12}^*\|_0 \leq s$ is the main assumption for our theoretical results. It imposes sparsity on Ω_{12}^* , that is, the dependency structure between $X_{\mathcal{G}_1}$ and $X_{\mathcal{G}_2}$ is sparse. Note that we do not make any assumption about the sparsity of the overall precision matrix Ω^* . It can be rather dense. Second, $\|\Sigma^*\|_{\max} \leq K$ and $\|\Theta^*\|_{1,1} \leq R$ are two regularity conditions on the Gaussian graphical model. The first specifies that the covariance between any two variables cannot be too large. It is weaker than $\|\Sigma^*\|_2 \leq K$ since $\|\Sigma^*\|_{\max} \leq \|\Sigma^*\|_2$. This assumption can be commonly found

in literatures on covariance and precision matrix estimation (Bickel and Levina 2008; Rothman et al. 2008). An easy consequence is that $\|\Sigma_{\mathcal{G}}^*\|_{\max} \leq K$ since $\Sigma_{\mathcal{G}}^*$ is the block diagonal of Σ^* . $\|\Theta^*\|_{1,1} \leq R$ requires the inter-subject dependency has constant sparsity. Similar conditions can be found in literatures on differential networks (Zhao, Cai, and Li 2014).

Theorem 4.1 shares the same spirit with the convergence results for ℓ_1 regularized maximum likelihood estimator (Rothman et al. 2008), that is the rate for estimating an s^* -sparse parameter Θ^* is $\sqrt{s^* \log d/n}$ in Frobenius norm. However, there are two things worth to be noted here. The first is that in **Theorem 4.1**, s^* can be replaced with any upper bound of $\|\Theta^*\|_0$ and the result is still valid. We know s^* is an upper bound of $\|\Theta^*\|_0$. In the worst case, $\|\Theta^*\|_0$ can be as large as s^* and when $\|\Theta^*\|_0$ is smaller, the rate in **Theorem 4.1** can be improved. Second, recall that $s^* \asymp s^2$, where s is the sparsity of Ω_{12}^* . Considering our goal is to estimate Ω_{12}^* , the rate seems to be suboptimal. Especially in the case $d_1 = 1$, neighborhood selection (Meinshausen and Bühlmann 2006; Yuan 2010) and CLIME (Cai, Liu, and Luo 2011) can obtain the optimal rate $\sqrt{s \log d/n}$ for the Frobenius norm. However, as we pointed out in Section 1, these methods cannot be applied when $d_1 \asymp d_2$ due to the violation of the sparsity assumption on Ω^* .

4.2. Asymptotic Inference

In this section, we give the limiting distribution of the de-biased estimator in (8). The asymptotic normality result is presented in **Theorem 4.3**. Based on this, we propose valid asymptotic confidence intervals and test statistics for parameters in Ω_{12}^* .

We first state a version of asymptotic normality result which involves population quantities.

Theorem 4.2 (Asymptotic normality). Suppose the conditions in **Theorem 4.1** hold. Further assume $\|\Omega^*\|_1 \leq L$ for some absolute constant $L > 0$. Let $\hat{\Theta}^u$ be the de-biased estimator with $\lambda' = C\sqrt{\log d/n}$, where C is a sufficiently large constant. For any $1 \leq j \leq d_1$ and $d_1 + 1 \leq k \leq d$, define the asymptotic variance as

$$\begin{aligned} \xi_{jk}^2 &= (M_{j*} \Sigma^* M_{j*}^\top) [P_{k*} (I + \Sigma_{\mathcal{G}}^* \Theta^*) \Sigma_{\mathcal{G}}^* P_{k*}^\top] + (M_{j*} \Sigma_{\mathcal{G}}^* P_{k*}^\top)^2 \\ &\quad - (M_{j*} \Sigma^* P_{k*}^\top)^2 \\ &\quad - [M_{j*} (I - \Sigma^* \Theta^*) \Sigma_{\mathcal{G}_2}^* (I - \Theta^* \Sigma^*) M_{j*}^\top] (P_{k*} \Sigma_{\mathcal{G}}^* P_{k*}^\top), \end{aligned} \quad (13)$$

where $\Sigma_{\mathcal{G}_2}^* = \text{diag}(0, \Sigma_2^*)$. Under the scaling condition $s^* \log d/\sqrt{n} = o(1)$, we have

$$\sqrt{n} \cdot (\hat{\theta}_{jk}^u - \theta_{jk}^*) / \xi_{jk} \rightsquigarrow N(0, 1).$$

Remark 4.1. Note that the asymptotic variance ξ_{jk}^2 in (13) depends on the be-biasing matrices M and P , which are estimated from the second half of the data.

Remark 4.2. $\|\Omega^*\|_1 \leq L$ is a milder condition than the sparsity constraints on Ω^* in the sense that Ω^* can be rather dense. And this is the case for ISA. To further understand the essence of this assumption, we discuss connections between $\lambda_{\min}(\Sigma^*) \geq \rho_{\min}$ and $\|\Omega^*\|_1 \leq L$. Since $\Omega^* = (\Sigma^*)^{-1}$, it

is not hard to see that $\lambda_{\max}(\Omega^*) \leq 1/\rho_{\min}$. Hence, we have $\max_{j \in [d]} \|\Omega_{*j}^*\|_2 \leq 1/\rho_{\min}$. Here, instead of the column-wise ℓ_2 -norm boundedness, we assume that $\max_{j \in [d]} \|\Omega_{*j}^*\|_1 \leq L$. It is indeed stronger than the ℓ_2 one, but it is weaker than the sparsity assumption on Ω^* . Moreover, as shown by the lower bound in Cai, Liu, and Zhou (2016), imposing this assumption does not make it possible to consistently estimate the parameter. Based on Theorem 1.1 in Cai, Liu, and Zhou (2016), we have that the optimal rate for the matrix ℓ_1 -norm is $\mathbb{E}\|\hat{\Omega} - \Omega^*\|_1^2 \asymp d^2 \log d/n$, which means no consistent estimator for the whole precision matrix Ω^* exists when $d > n$.

To obtain the formula for the asymptotic variance ξ_{jk}^2 in (13), we use the Isserlis' theorem (Isserlis 1916) to calculate the fourth order moment of the Gaussian distribution. We can see that ξ_{jk}^2 still depends on population quantities Σ^* and Θ^* . Thus, ξ_{jk}^2 is unknown in practice, and we need to get a consistent estimator $\hat{\xi}_{jk}^2$ to construct confidence intervals for θ_{jk}^* .

Lemma 4.1 (Variance estimation). Define $\hat{\Sigma}_{\mathcal{G}_2} = \text{diag}(0, \hat{\Sigma}_2)$. For any $1 \leq j \leq d_1$ and $d_1 + 1 \leq k \leq d$, let $\hat{\xi}_{jk}^2$ be the empirical version of (13), that is,

$$\begin{aligned} \hat{\xi}_{jk}^2 &= (M_{j*} \hat{\Sigma}_{M_{j*}}^\top) [P_{k*} (I + \hat{\Sigma}_{\mathcal{G}} \hat{\Theta}) \hat{\Sigma}_{\mathcal{G}} P_{k*}^\top] + (M_{j*} \hat{\Sigma}_{\mathcal{G}} P_{k*}^\top)^2 \\ &\quad - (M_{j*} \hat{\Sigma}_{P_{k*}}^\top)^2 \\ &\quad - [M_{j*} (I - \hat{\Sigma} \hat{\Theta}) \hat{\Sigma}_{\mathcal{G}_2} (I - \hat{\Theta} \hat{\Sigma}) M_{j*}^\top] (P_{k*} \hat{\Sigma}_{\mathcal{G}} P_{k*}^\top). \end{aligned} \quad (14)$$

Then under the conditions in Theorem 4.2, $\hat{\xi}_{jk}/\xi_{jk}$ converges in probability to 1.

Combining Theorem 4.2 and Lemma 4.1 with Slutsky's theorem, we can obtain the final version of the asymptotic normality result which does not involve any population quantity.

Theorem 4.3. Suppose the conditions in Theorem 4.2 hold. Let $\hat{\Theta}^u$ be the de-biased estimator with $\lambda' = C' \sqrt{\log d/n}$, where C' is a sufficiently large constant. For any $1 \leq j \leq d_1$ and $d_1 + 1 \leq k \leq d$, under the scaling condition $s^* \log d/\sqrt{n} = o(1)$, we have

$$\sqrt{n} \cdot (\hat{\theta}_{jk}^u - \theta_{jk}^*)/\hat{\xi}_{jk} \rightsquigarrow N(0, 1).$$

Applying Theorem 4.3, it is easy to construct asymptotically valid confidence intervals and test functions. For any $1 \leq j \leq d_1$ and $d_1 + 1 \leq k \leq d$ and the significance level $\alpha \in (0, 1)$, let

$$\begin{aligned} I_{jk}(\alpha) &= [\hat{\theta}_{jk}^u - \delta(\alpha, n), \hat{\theta}_{jk}^u + \delta(\alpha, n)], \quad \text{where} \\ \delta(\alpha, n) &= \frac{\hat{\xi}_{jk}}{\sqrt{n}} \Phi^{-1}\left(1 - \frac{\alpha}{2}\right). \end{aligned} \quad (15)$$

Also for the null hypothesis $H_0 : \theta_{jk}^* = 0$, we can construct the following test function

$$T_{jk}(\alpha) = \begin{cases} 1 & \text{if } |\sqrt{n} \cdot \hat{\theta}_{jk}^u/\hat{\xi}_{jk}| > \Phi^{-1}(1 - \alpha/2), \\ 0 & \text{if } |\sqrt{n} \cdot \hat{\theta}_{jk}^u/\hat{\xi}_{jk}| \leq \Phi^{-1}(1 - \alpha/2), \end{cases} \quad (16)$$

where $\alpha \in (0, 1)$ is the significance level of the test. Corollary 4.1 proves the validity of the confidence interval and the test function.

Corollary 4.1. Suppose the conditions in Theorem 4.3 hold. The confidence interval in (15) is asymptotically valid and the Type I error of (16) is asymptotically α , that is,

$$\begin{aligned} \lim_{n \rightarrow \infty} \mathbb{P}(\theta_{jk}^* \in I_{jk}(\alpha)) &= 1 - \alpha \quad \text{and} \\ \lim_{n \rightarrow \infty} \mathbb{P}_{\theta_{jk}^*=0}(T_{jk}(\alpha) = 1) &= \alpha. \end{aligned}$$

5. Numerical Experiments

In this section, we conduct numerical experiments on both simulated and real data to validate our STRINGS estimator and the ‘‘untangle and chord’’ procedure. We also compare our method with ℓ_1 regularized maximum likelihood estimator (GLASSO), partial Gaussian graphical model (pGGM) in Yuan and Zhang (2014) and the density ratio estimator (KLIEP) in Liu et al. (2016).

5.1. Simulated Data

For each dimension d and probability $p \in (0, 1)$ which governs the sparsity s , we generate the precision matrix Ω^* as follows. First, we generate a symmetric matrix A , where it has zeros in the diagonal and each off-diagonal entry of A is set to one with probability p . Then, $\delta \cdot I_d$ is added to A to make its condition number equal to d . Third, we add all-one matrices to the groups $\mathcal{G}_1 = \{1, \dots, d/2\}$ and $\mathcal{G}_2 = \{d/2 + 1, \dots, d\}$ to represent the dense within-group connections. Finally, the precision matrix is standardized so that the diagonal entries of Ω^* are all ones. It is straightforward to see that the sparsity s is approximately equal to $d^2 p/4$. Under this model, we generate $n = 4d$ training samples from the multivariate normal distribution with mean 0 and covariance $\Sigma^* = (\Omega^*)^{-1}$.

5.1.1. Estimation Quality

Since we are estimating the support of Ω_{12}^* , we adopt standard graphical plot, receiver operating characteristic curve (ROC curve) to demonstrate the performance of different estimators. ROC curve plots the true positive rate (TPR) against the false positive rate (FPR) for different values of the regularization parameter λ .

For $\hat{\Omega}_{12}$, absolute values above 1×10^{-3} are considered to be nonzeros since we set the optimization accuracy to be 1×10^{-4} . We consider different values of $d \in \{50, 100, 200\}$ and $p \in \{0.04, 0.08, 0.12\}$. The ROC curves are plotted in Figure 3. In addition, for each configuration, we report the mean and the standard error of the area under the ROC curve (AUC) over 100 replications in Table 1.

We can see that the STRINGS estimator outperforms other estimators uniformly over all configurations of (d, p) . This is expected since (a) GLASSO is not tailored for estimation under dense precision matrices; (b) pGGM can only tolerate one dense block matrix.

5.1.2. Inference Quality

For inference, we only report the results for our ‘‘untangle and chord’’ procedure since (a) GLASSO, not designed for partially sparse graphical model estimation, performs poorly in terms of estimation accuracy; and (b) pGGM and KLIEP do not provide

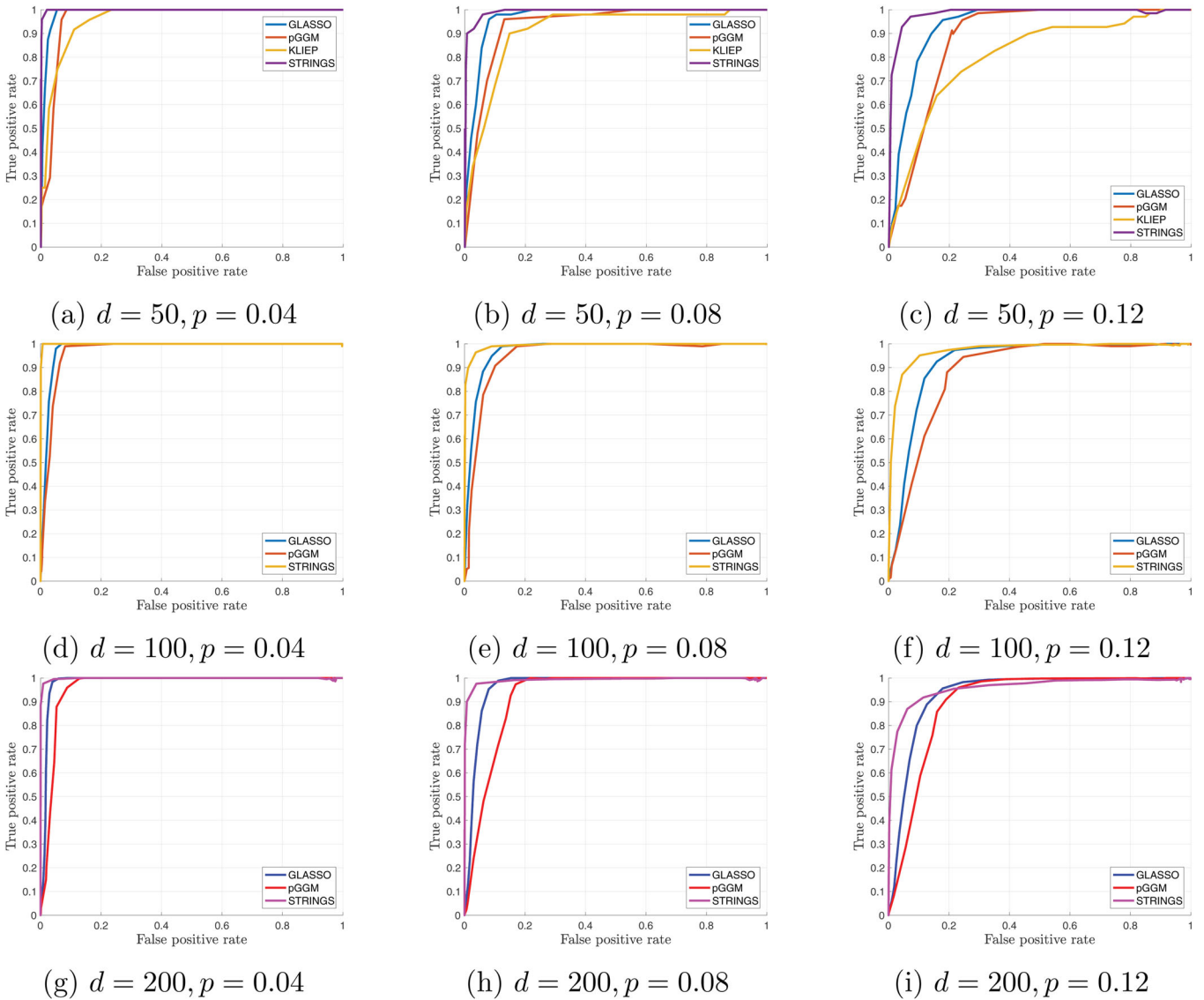


Figure 3. ROC curves for different configurations of dimension d and sparsity.

Table 1. Mean (standard error) of AUC over 100 replications.

d	p	$s \approx d^2 p/4$	GLASSO	pGGM	KLIEP	STRINGS
50	0.04	25	0.986(0.004)	0.959(0.012)	0.875(0.016)	0.997(0.002)
50	0.08	50	0.958(0.006)	0.909(0.020)	0.820(0.018)	0.986(0.008)
50	0.12	75	0.932(0.007)	0.877(0.028)	0.777(0.016)	0.974(0.009)
d	p	$s \approx d^2 p/4$	GLASSO	pGGM	KLIEP	STRINGS
100	0.04	100	0.978(0.002)	0.967(0.004)	nan(nan)	0.998(0.002)
100	0.08	200	0.969(0.003)	0.944(0.010)	nan(nan)	0.990(0.003)
100	0.12	300	0.922(0.005)	0.880(0.012)	nan(nan)	0.959(0.007)
d	p	$s \approx d^2 p/4$	GLASSO	pGGM	KLIEP	STRINGS
200	0.04	400	0.981(0.001)	0.958(0.007)	nan(nan)	0.994(0.004)
200	0.08	800	0.963(0.001)	0.922(0.007)	nan(nan)	0.988(0.003)
200	0.12	1200	0.934(0.002)	0.891(0.007)	nan(nan)	0.961(0.003)

NOTE: The bold values highlight the superior performance of STRINGS over other methods.

any inferential procedure. We generate another sample of size $4d$ to de-bias the initial STRINGS estimator. Guided by [Theorem 4.2](#), the tuning parameter λ' in CLIME-type procedure is chosen to be $0.5\sqrt{\log d/n}$. By [Corollary 4.1](#), the $(1 - \alpha) \times 100\%$

asymptotic confidence interval for parameter θ_{jk}^* is given by

$$I_{jk}(\alpha) = \left[\hat{\theta}_{jk}^u - \frac{\hat{\xi}_{jk}}{\sqrt{n}} \Phi^{-1}\left(1 - \frac{\alpha}{2}\right), \hat{\theta}_{jk}^u + \frac{\hat{\xi}_{jk}}{\sqrt{n}} \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \right],$$

where $\hat{\Theta}^u$ is the de-biased estimator and $\hat{\xi}_{jk}$ is specified in (14). Throughout this section, we set $\alpha = 0.05$. For every parameter θ_{jk}^* , we estimate the probability that the true value θ_{jk}^* is covered by the confidence interval $I_{jk}(\alpha)$ using its empirical version, that is, $\hat{\alpha}_{jk}$ is the percentage of times that θ_{jk}^* is covered by $I_{jk}(\alpha)$ in 100 replications. Next for $S = \text{supp}(\Omega_{12}^*)$, we define the overall coverage probability, the average coverage probability over S and over S^c to be

$$\begin{aligned} \text{Avcov} &= \frac{1}{(d/2)^2} \sum_{(j,k)} \hat{\alpha}_{jk}, & \text{Avcov}_S &= \frac{1}{|S|} \sum_{(j,k) \in S} \hat{\alpha}_{jk}, \\ \text{Avcov}_{S^c} &= \frac{1}{|S^c|} \sum_{(j,k) \in S^c} \hat{\alpha}_{jk}, \end{aligned} \quad (17)$$

respectively. The result of these three quantities over 100 replications can be seen in [Table 2](#). The coverage probabilities over the

Table 2. Average coverage probabilities over 100 replications.

d	p	$s \approx d^2 p/4$	Avgcov	Avgcov _{s}	Avgcov _{s^c}
50	0.04	25	0.952(0.009)	0.942(0.046)	0.952(0.010)
50	0.08	50	0.945(0.012)	0.927(0.037)	0.947(0.012)
50	0.12	75	0.939(0.012)	0.921(0.033)	0.942(0.012)
100	0.04	100	0.947(0.006)	0.930(0.025)	0.948(0.006)
100	0.08	200	0.946(0.006)	0.934(0.018)	0.947(0.006)
100	0.12	300	0.941(0.005)	0.928(0.014)	0.943(0.006)
200	0.04	400	0.930(0.007)	0.907(0.022)	0.931(0.006)
200	0.08	800	0.926(0.007)	0.907(0.015)	0.927(0.007)
200	0.12	1200	0.927(0.004)	0.915(0.009)	0.930(0.004)

support S and the nonsupport S^c are around 95%, which matches the significance level $\alpha = 0.05$. And the coverage probability over S decreases as the dimension d increases, as expected.

In Figure 4, we show the QQ-plot of $\sqrt{n} \cdot (\hat{\theta}_{jk}^u - \theta_{jk}^*) / \hat{\xi}_{jk}$ when $d = 200$ and $p = 0.04$. We choose $(j, k) \in \{(101, 1), (101, 2), (101, 3)\}$ to present. As we can see, the scattered points of $\sqrt{n} \cdot (\hat{\theta}_{jk}^u - \theta_{jk}^*) / \hat{\xi}_{jk}$ in 100 replications are close to the line with zero intercept and unit slope.

5.2. fMRI Data

In this section, we apply our estimation and inference methods to an fMRI data studied in Chen et al. (2017). This dataset includes fMRI measurements of 17 subjects while they were watching a 23-min movie (BBC’s “Sherlock”). The fMRI measurements were made every 1.5 sec, thus in total we have 945 brain images for each subject. As described in Chen et al. (2017), the 23-min movie is divided into 26 scenes for further analysis. For the original fMRI data, there are 271,633 voxels measured. We adopt the method introduced in Baldassano, Beck, and Fei-Fei (2015) to reduce the dimension to 172 regions of interest (ROIs). We use the average of the first eight subjects as X_{G_1} and the average of the remaining nine subjects as X_{G_2} for conducting ISA. For preprocessing, each ROI is standardized to have zero mean and unit variance. For estimation, the tuning parameter λ is chosen through cross-validation. In inference, we threshold the de-biased estimator at level $\Phi^{-1}(1 - 4\alpha/d^2) \cdot \hat{\xi}_{jk} / \sqrt{n}$, where $\alpha = 0.05$ and $4/d^2$ accounts for the Bonferroni correction in multiple hypothesis testing. We pick the eighth scene and

the fifteenth scene for presentation. Scene 8 represents a press conference held by the police department to describe the recent suicides. Scene 15 contains the first meeting of Sherlock and Watson during which Sherlock shows his deduction talent to Watson.

In Figure 5, we show the brain networks for these two scenes estimated by our method. Each purple circle represents an ROI. We also show the snapshots of both the left and the right brain hemispheres in Figure 6. The color represents the degree of the ROIs in the inter-subject conditional independence graph. And a redder area corresponds to the ROI with higher degree. As we can see, for the eighth scene when the press conference took place, the visual cortex and auditory cortex are highly activated since the subjects were mostly receiving audio and visual information from the press conference. The high activation of the visual and auditory cortices are ubiquitous in all 26 scenes. This makes sense since the subjects were under an audio-visual stimulus (“BBC’s Sherlock”). This also matches the results in Chen et al. (2017). More specifically, during the fifteenth scene when Sherlock and Watson met, we can see that the prefrontal cortex especially the dorsolateral prefrontal cortex (DL-PFC) has a large degree. DL-PFC is known for its function in working memory and abstract reasoning (Miller and Cummings 2007). And this coincides with scene 15 since the subjects might reason about Sherlock’s deduction about Watson’s job.

6. Extensions to Multi-Subject Analysis

In this section, we discuss the extension of ISA to multiple subjects. As a motivating example, let us revisit the fMRI data considered in Section 5.2. In total, there are 17 subjects and we artificially divide them into 8 and 9 to perform ISA. A more principled way to analyze this data would be conducting multi-subject analysis, which we detail below. Let $X = (X_1, \dots, X_{Ld})^\top$ be an Ld -dimensional random vector, where L is the number of subjects and d is the number of features for each subject. Let $\mathcal{G}_1, \dots, \mathcal{G}_L$ be L disjoint subsets of $\{1, \dots, Ld\}$ with cardinality $|\mathcal{G}_\ell| = d$, each corresponding to a single subject. Denote by $X_{\mathcal{G}_\ell}$ the features of the ℓ th subject. Let $\Sigma^* \in \mathbb{R}^{Ld \times Ld}$ be the covariance matrix of X , with $\Sigma_{jk}^* \in \mathbb{R}^{d \times d}$ being the covariance between $X_{\mathcal{G}_j}$ and $X_{\mathcal{G}_k}$. For the precision matrix $\Omega^* = (\Sigma^*)^{-1}$, we use Ω_{jk}^* to denote the dependency between $X_{\mathcal{G}_j}$ and $X_{\mathcal{G}_k}$.

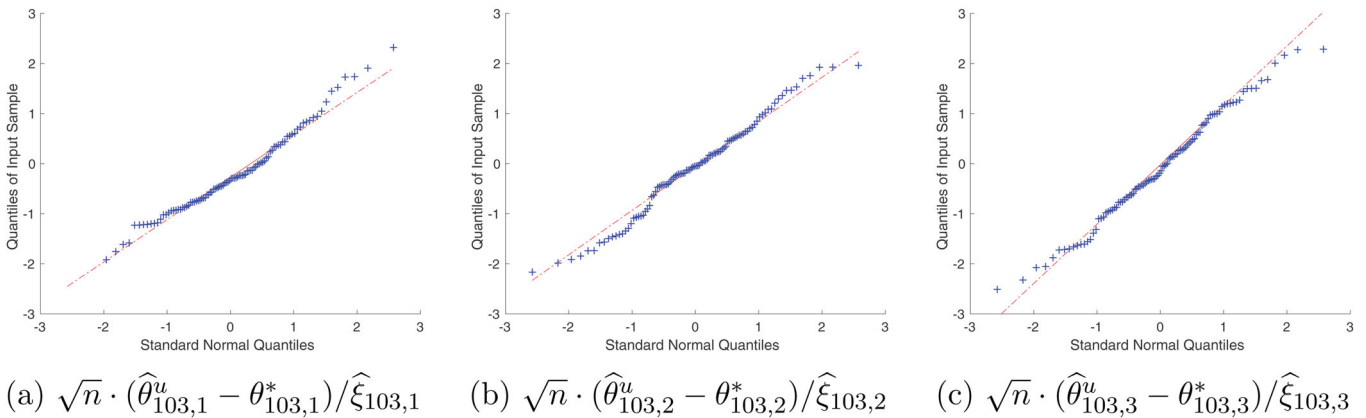


Figure 4. The QQ-plot of $\sqrt{n} \cdot (\hat{\theta}_{jk}^u - \theta_{jk}^*) / \hat{\xi}_{jk}$ for $(j, k) = (103, 1), (103, 2),$ and $(103, 3)$.

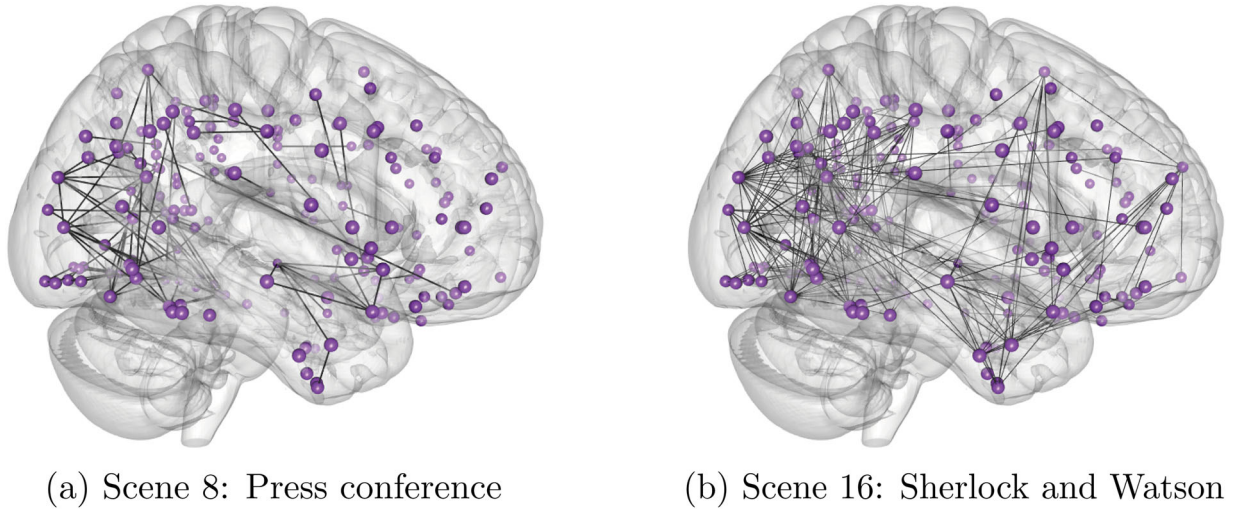


Figure 5. The brain networks for two different scenes. Each purple circle represents an ROI and the black edges represent the graph edges.

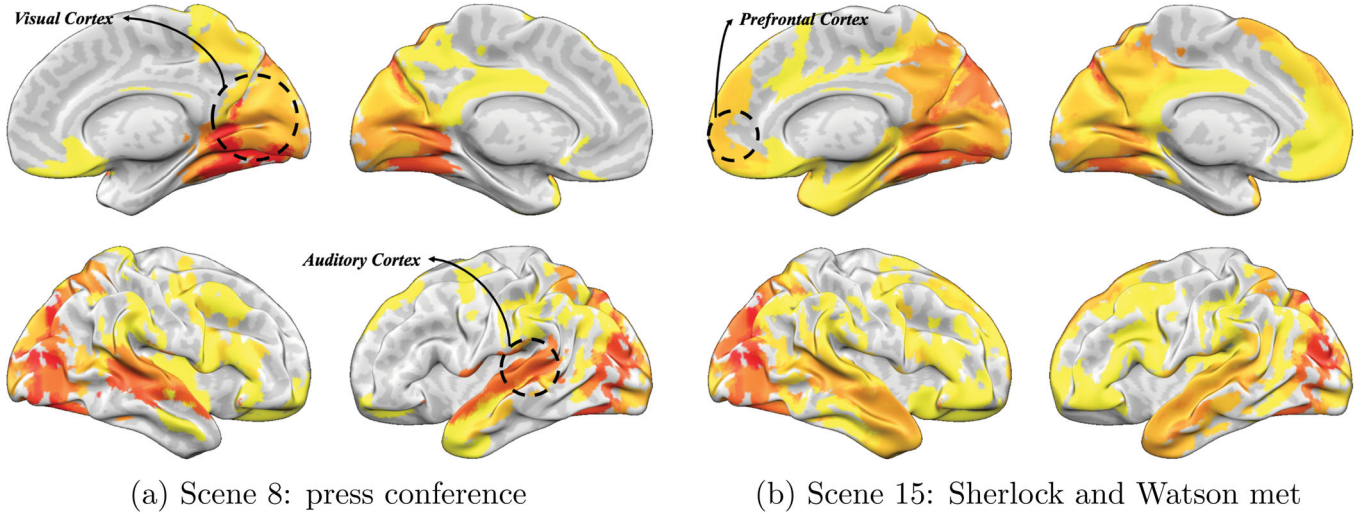


Figure 6. The brain images for both left and right hemisphere.

Define $\Sigma_G^* = \text{diag}(\Sigma_{11}^*, \dots, \Sigma_{LL}^*)$ and $\Theta^* = \Omega^* - (\Sigma_G^*)^{-1}$. We have a similar observation to that in the two-subject case.

Proposition 6.1. If $\|\Omega_{jk}^*\|_0 \leq s$ for all $j \neq k$, then $\|\Theta^*\|_0 = \mathcal{O}(Ls^2)$. Further assume that $L = \mathcal{O}(1)$, we have $\|\Theta^*\|_0 = \mathcal{O}(s^2)$.

In words, this justifies the estimation of Θ^* , instead of Ω^* , when we only know the off-diagonal parts of Ω^* are sparse. In addition, in cases like the fMRI data, it is often reasonable to assume that all the off-diagonal blocks of Ω^* are the same. This is because all the subjects are under the same stimuli, that is, the movie Sherlock. In this case, the multi-subject STRINGS estimator amounts to solving the following optimization problem:

$$\begin{aligned} \text{minimize}_{\Theta \in Ld \times Ld} \quad & \text{Tr}(\Theta \hat{\Sigma}) - \log |\hat{\Sigma}_G \Theta \hat{\Sigma}_G + \hat{\Sigma}_G| \\ & + \lambda \|\Theta\|_{1,1} \quad (18) \\ \text{subject to} \quad & \Theta_{ij} = \Theta_{kl} \quad \text{for any } i \neq j, k \neq l. \end{aligned}$$

In words, this formulation maximizes the regularized log-likelihood under the constraint that all the off-diagonal blocks of Θ are the same. It is straightforward to see that this multi-block

estimator reduces to the STRINGS estimator in the two-block case. This is indeed a computationally efficient estimator that can be solved by off-the-shelf software packages, however, its theoretical guarantees are elusive. First, the establishment of the estimation guarantees is complicated by the equality constraint. Second, the debiasing technique in this multi-subject case is not a priori clear. Last but not least, the number of parameters to estimate is enlarged from s to Ls^2 , which could be a challenge for real world applications that only contain limited amount of data. Addressing these issues is of great importance and we leave it as a future work.

7. Discussion

In this work, we consider the problem of ISA, where the goal is to study the inter-subject dependency while the intra-subject dependence is treated as nuisance. Under the framework of Gaussian graphical models, we propose a consistent estimator, STRINGS estimator, and a debiasing technique called *Untangle and Chord* to construct confidence intervals in the high-dimensional regime. There are numerous questions that are interesting for future investigation and here we single out a few.

- *Handle moderate sparsity.* Our result (cf. Theorem 4.1) guarantees a consistent estimator of Ω_{12}^* in the high-dimensional regime when $s \ll n \ll d^2$ with s being the sparsity of Ω_{12}^* . This excludes the case with $s = O(d)$. How to construct a consistent estimator in this challenging regime is an interesting question to investigate.
- *Inference without sample splitting.* The current work utilizes sample splitting to decouple the dependency between the de-biasing matrices and the estimate $\hat{\Theta}$. This may not be efficient since only half of the data is used for estimation and may result in low power. It is of great importance to study de-biasing techniques without sample splitting in the high-dimensional regime beyond simple linear models.

Supplementary Materials

In the supplementary materials, we provide proofs for the theoretical results in the main text.

Funding

J. Lu is supported in part by NSF1916211, NIH funding: NIH1R35CA220523-01A1, and NIH5U01CA209414-02.

References

- Baldassano, C., Beck, D. M., and Fei-Fei, L. (2015), “Parcelating Connectivity in Spatial Maps,” *PeerJ*, 3, e784. [8]
- Banerjee, O., Ghaoui, L. E., and d’Aspremont, A. (2008), “Model Selection Through Sparse Maximum Likelihood Estimation for Multivariate Gaussian or Binary Data,” *Journal of Machine Learning Research*, 9, 485–516. [3]
- Bartels, A., and Zeki, S. (2004), “Functional Brain Mapping During Free Viewing of Natural Scenes,” *Human Brain Mapping*, 21, 75–85. [1]
- Belitski, A., Gretton, A., Magri, C., Murayama, Y., Montemurro, M. A., Logothetis, N. K., and Panzeri, S. (2008), “Low-Frequency Local Field Potentials and Spikes in Primary Visual Cortex Convey Independent Visual Information,” *The Journal of Neuroscience*, 28, 5696–5709. [1]
- Bickel, P. J., and Levina, E. (2008), “Regularized Estimation of Large Covariance Matrices,” *The Annals of Statistics*, 36, 199–227. [5]
- Cai, T., Liu, W., and Luo, X. (2011), “A Constrained ℓ_1 Minimization Approach to Sparse Precision Matrix Estimation,” *Journal of the American Statistical Association*, 106, 594–607. [3,4,5]
- Cai, T. T., Liu, W., and Zhou, H. H. (2016), “Estimating Sparse Precision Matrix: Optimal Rates of Convergence and Adaptive Estimation,” *The Annals of Statistics*, 44, 455–488. [6]
- Chen, J., Leong, Y. C., Honey, C. J., Yong, C. H., Norman, K. A., and Hasson, U. (2017), “Shared Memories Reveal Shared Structure in Neural Activity Across Individuals,” *Nature Neuroscience*, 20, 115–125. [8]
- Friedman, J., Hastie, T., and Tibshirani, R. (2008), “Sparse Inverse Covariance Estimation With the Graphical Lasso,” *Biostatistics*, 9, 432–441. [3]
- Gu, Q., Cao, Y., Ning, Y., and Liu, H. (2015), “Local and Global Inference for High Dimensional Gaussian Copula Graphical Models,” arXiv no. 1502.02347. [3]
- Hartley, T., Maguire, E. A., Spiers, H. J., and Burgess, N. (2003), “The Well-Worn Route and the Path Less Traveled: Distinct Neural Bases of Route Following and Wayfinding in Humans,” *Neuron*, 37, 877–888. [1]
- Hasson, U., Nir, Y., Levy, I., Fuhrmann, G., and Malach, R. (2004), “Inter-subject Synchronization of Cortical Activity During Natural Vision,” *Science*, 303, 1634–1640. [1]
- Horowitz, B., and Rapoport, S. I. (1988), “Partial Correlation Coefficients Approximate the Real Intrasubject Correlation Pattern in the Analysis of Interregional Relations of Cerebral Metabolic Activity,” *Journal of Nuclear Medicine*, 29, 392–399. [1]
- Huang, S., Li, J., Sun, L., Ye, J., Fleisher, A., Wu, T., Chen, K., Reiman, E., and Alzheimer’s Disease Neuroimaging Initiative (2010), “Learning Brain Connectivity of Alzheimer’s Disease by Sparse Inverse Covariance Estimation,” *NeuroImage*, 50, 935–949. [1]
- Isserlis, L. (1916), “On Certain Probable Errors and Correlation Coefficients of Multiple Frequency Distributions With Skew Regression,” *Biometrika*, 11, 185–190. [6]
- Jankova, J., and van de Geer, S. (2015), “Confidence Intervals for High-Dimensional Inverse Covariance Estimation,” *Electronic Journal of Statistics*, 9, 1205–1229. [3]
- Javanmard, A., and Montanari, A. (2014), “Confidence Intervals and Hypothesis Testing for High-Dimensional Regression,” *Journal of Machine Learning Research*, 15, 2869–2909. [3]
- Lee, H., Lee, D. S., Kang, H., Kim, B.-N., and Chung, M. K. (2011), “Sparse Brain Network Recovery Under Compressed Sensing,” *IEEE Transactions on Medical Imaging*, 30, 1154–1165. [1]
- Lindquist, M. A. (2008), “The Statistical Analysis of fMRI Data,” *Statistical Science*, 23, 439–464. [1]
- Liu, S., Suzuki, T., Sugiyama, M., and Fukumizu, K. (2016), “Structure Learning of Partitioned Markov Networks,” in *International Conference on Machine Learning*. [3,6]
- Marrelec, G., Krainik, A., Duffau, H., Péligrini-Issac, M., Lehericy, S., Doyon, J., and Benali, H. (2006), “Partial Correlation for Functional Brain Interactivity Investigation in Functional MRI,” *Neuroimage*, 32, 228–237. [1]
- Mechler, F., Victor, J. D., Purpura, K. P., and Shapley, R. (1998), “Robust Temporal Coding of Contrast by VI Neurons for Transient But Not for Steady-State Stimuli,” *Journal of Neuroscience*, 18, 6583–6598. [1]
- Meinshausen, N., and Bühlmann, P. (2006), “High-Dimensional Graphs and Variable Selection With the Lasso,” *The Annals of Statistics*, 34, 1436–1462. [3,5]
- Miller, B. L., and Cummings, J. L. (2007), *The Human Frontal Lobes: Functions and Disorders*, New York: Guilford Press. [8]
- Ning, Y., and Liu, H. (2017), “A General Theory of Hypothesis Tests and Confidence Regions for Sparse High Dimensional Models,” *The Annals of Statistics*, 45, 158–195. [3]
- Ravikumar, P., Wainwright, M. J., Raskutti, G., and Yu, B. (2011), “High-Dimensional Covariance Estimation by Minimizing ℓ_1 -Penalized Log-Determinant Divergence,” *Electronic Journal of Statistics*, 5, 935–980. [3]
- Rothman, A. J., Bickel, P. J., Levina, E., and Zhu, J. (2008), “Sparse Permutation Invariant Covariance Estimation,” *Electronic Journal of Statistics*, 2, 494–515. [3,5]
- Simony, E., Honey, C. J., Chen, J., Lositsky, O., Yeshurun, Y., Wiesel, A., and Hasson, U. (2016), “Dynamic Reconfiguration of the Default Mode Network During Narrative Comprehension,” *Nature Communications*, 7, 12141. [1]
- van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. (2014), “On Asymptotically Optimal Confidence Regions and Tests for High-Dimensional Models,” *The Annals of Statistics*, 42, 1166–1202. [3,4]
- Varoquaux, G., Gramfort, A., Poline, J. B., and Thirion, B. (2012), “Markov Models for fMRI Correlation Structure: Is Brain Functional Connectivity Small World, or Decomposable Into Networks?,” *Journal of Physiology-Paris*, 106, 212–221. [1]
- Yao, H., Shi, L., Han, F., Gao, H., and Dan, Y. (2007), “Rapid Learning in Cortical Coding of Visual Scenes,” *Nature Neuroscience*, 10, 772–778. [1]
- Yuan, M. (2010), “High Dimensional Inverse Covariance Matrix Estimation via Linear Programming,” *Journal of Machine Learning Research*, 11, 2261–2286. [3,5]
- Yuan, M., and Lin, Y. (2007), “Model Selection and Estimation in the Gaussian Graphical Model,” *Biometrika*, 94, 19–35. [3]
- Yuan, X.-T., and Zhang, T. (2014), “Partial Gaussian Graphical Model Estimation,” *IEEE Transactions on Information Theory*, 60, 1673–1687. [3,6]
- Zacks, J. M., Braver, T. S., Sheridan, M. A., Donaldson, D. I., Snyder, A. Z., Ollinger, J. M., Buckner, R. L., and Raichle, M. E. (2001), “Human Brain Activity Time-Locked to Perceptual Event Boundaries,” *Nature Neuroscience*, 4, 651–655. [1]
- Zhang, C.-H., and Zhang, S. S. (2014), “Confidence Intervals for Low Dimensional Parameters in High Dimensional Linear Models,” *Journal of the Royal Statistical Society, Series B*, 76, 217–242. [3,4]
- Zhao, S. D., Cai, T. T., and Li, H. (2014), “Direct Estimation of Differential Networks,” *Biometrika*, 101, 253–268. [5]