The Annals of Applied Statistics
2016, Vol. 10, No. 3, 1397–1426
DOI: 10.1214/16-AOAS940

SPARSE MEDIAN GRAPHS ESTIMATION IN A HIGH-DIMENSIONAL SEMIPARAMETRIC MODEL

BY FANG HAN*, XIAOYAN HAN†, HAN LIU† AND BRIAN CAFFO*

Johns Hopkins University and Princeton University†*

We propose a unified framework for conducting inference on complex aggregated data in high-dimensional settings. We assume the data are a collection of multiple non-Gaussian realizations with underlying undirected graphical structures. Using the concept of median graphs in summarizing the commonality across these graphical structures, we provide a novel semiparametric approach to modeling such complex aggregated data, along with robust estimation of the median graph, which is assumed to be sparse. We prove the estimator is consistent in graph recovery and give an upper bound on the rate of convergence. We further provide thorough numerical analysis on both synthetic and real datasets to illustrate the empirical usefulness of the proposed models and methods.

1. Introduction.

Undirected graphs provide a powerful tool for understanding the interrelationships among random variables. Given a random vector, $\boldsymbol{X} = (X_1, \ldots, X_d)^T \in \mathbb{R}^d$, the associated conditional independence graph, $\mathcal{G} \in \{0, 1\}^{d \times d}$, is the undirected binary graph so that the entry $\mathcal{G}_{jk}$ (for $j \neq k$) is equal to 0 if and only if X_j is conditionally independent of X_k given the remaining variables, $\{X_{\backslash\{j,k\}}\}$. For estimation, it is typically assumed there are n independent and identically distributed realizations of $\boldsymbol{X}$ to infer independence relationships and, thus, the associated graph $\mathcal{G}$.

When $\boldsymbol{X} \sim N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is Gaussian distributed with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$, the nonzero entries of the concentration matrix $\boldsymbol{\Omega} := \boldsymbol{\Sigma}^{-1}$ encode the conditional independence structure of $\boldsymbol{X}$, and hence define the graph $\mathcal{G}$ [Dempster (1972)]. In other words, $\mathcal{G}_{jk} = I(\boldsymbol{\Omega}_{jk} \neq 0)$, where $I(\cdot)$ is an indicator function. Estimation of $\boldsymbol{\Omega}$ becomes problematic in high dimensions where $d > n$, thus leading to an active collection of research using sparsity constraints to obtain identifiability [see for example Cai, Liu and Luo (2011), Friedman, Hastie and Tibshirani (2007), Liu and Luo (2012), Ravikumar et al. (2009), Scheinberg, Ma and Glodfarb (2010), Yuan (2010), Banerjee, El Ghaoui and d'Aspremont (2008), Li and Toh (2010), Hsieh et al. (2011), Rothman et al. (2008), Lam and Fan (2009), Peng et al. (2009), Meinshausen and Bühlmann (2006)].

However, these papers all assume the object of inference is a single graph estimated from a single set of realizations of $\boldsymbol{X}$. In contrast, little work exists on

Received July 2015; revised April 2016.

Key words and phrases. Graphical model, median graph, complex aggregated data, semiparametric model, high-dimensional statistics.

estimation and inference from a population of graphs. Such a setting arises frequently in the sometimes controversial and rapidly evolving arenas of image- and electrophysiologically-based estimates of functional and structural brain connectivity [Bullmore and Sporns (2009), Fingelkurts and Kähkönen (2005), Friston (2011), Horwitz et al. (2003), Rubinov and Sporns (2010)]. Here, each subject-specific graph is an estimate of subject-specific brain connectivity.

In addition, frequently the assumption that the data are independently and identically drawn from a Gaussian distribution is too strong. Recently, Gaussian assumptions are relaxed via the *nonparanormal* distribution family [Liu, Lafferty and Wasserman (2009)]. A random vector, $\boldsymbol{X}$, is said to be nonparanormally distributed if, after an unspecified monotone transformation, it is Gaussian distributed. Moreover, an optimal graph recovery procedure is obtained by exploiting the rank-based estimator Kendall's tau [Liu et al. (2012)]. On the other hand, however, little has been done in high-dimensional graph estimation when the data are actually not identically drawn from a certain distribution.

This paper investigates a specific non-i.i.d. setting where the data arise from multiple datasets, each of which is assumed to be distributed according to a different distribution. This idea is central in fields, such as epidemiology, where population summaries are desired over collections of independently but not identically distributed datasets. A canonical example is the common odds ratio estimate from a collection of individual odds ratios [see for example Liu and Agresti (1996)]. In the motivating application, each dataset is a seed-based or region of interest summary of functional magnetic resonance imaging (fMRI) scans where a graphical representation of brain connectivity is of interest. The proposed approach does not assume a common underlying graph for each subject. Instead, the population graph defined is a summary, looking at commonalities in graphical structure across a population of heterogeneous graphs. Thus, it is proposed that, under the presumption of variation in brain graphical network structure, the investigation of a population graph is of conceptual and practical interest, especially when comparing population graphs across clinical diagnoses.

To best summarize the information from aggregated network datasets, the idea of "median graphs" from the pattern recognition field [Bunke and Shearer (1998), Jiang, Munger and Bunke (2001)] is employed. However, it is herein extended to *sparse median graphs*. A sparse median graph is defined as the sparse graph that has the smallest sum of Hamming distances to all graphs in a given sample. Combined with the strength of the nonparanormal modeling, a new method for estimating sparse median graphs is proposed. It is then proven that the obtained estimator is consistent. The upper bound on the convergence rate with respect to the Hamming distance is established, thus giving more understanding on the estimator's behavior.

In the neuroimaging literature, one relevant paper on summarizing multiple graphical models is Ramsay et al. (2009). There are three main differences between our proposed procedure and the one in Ramsay et al. (2009): (i) On the graph of interest, we focus on the undirected graphical models, while their focus is on the directed graphical models. (ii) On defining the summary graph combining the information from multiple datasets, Ramsay et al. (2009) propose a BIC-based data aggregation criterion, while we propose a median graph-based criterion. Our proposed method is shown to motivate a more robust estimation procedure. (iii) On conducting the algorithm, Ramsay et al. (2009) exploit a greedy search-based algorithm (GES), while we exploit a convex optimization-based algorithm (CLIME).

The rest of the paper is organized as follows. In Section 2, we introduce the notation and review the nonparanormal distribution and rank-based estimators. In Section 3, we introduce the model and give the definition of sparse median graphs. In Section 4, we propose the rank-based estimation procedures. Section 5 gives the theoretical properties of the proposed procedure for graph recovery. Section 6 demonstrates experimental results on both synthetic and real-world datasets. Discussions are in the last section.

2. Background.

Let $\mathbf{M} = [M_{jk}] \in \mathbb{R}^{d\times d}$ and $\boldsymbol{v} = (v_1, \dots, v_d)^T \in \mathbb{R}^d$. Let $\boldsymbol{v}_I$ denote the subvector of $\boldsymbol{v}$ with entries indexed by set I. Similarly, let the submatrix of $\mathbf{M}$ with rows indexed by set I and columns indexed by set J be denoted by $\mathbf{M}_{IJ}$. Let $\mathbf{M}_{I*}$ and $\mathbf{M}_{*J}$ be the submatrix of $\mathbf{M}$ with rows in I and the submatrix of $\mathbf{M}$ with columns in J. For $0 < q < \infty$, define the ℓ_q and ℓ_∞ vector norms as

$$\|\boldsymbol{v}\|_q := \left(\sum_{i=1}^d |v_i|^q\right)^{1/q} \quad \text{and} \quad \|\boldsymbol{v}\|_\infty := \max_{1\le i\le d} |v_i|,$$

and we define

$$\|\boldsymbol{v}\|_0 := \sum_{i=1}^d I(v_i \neq 0),$$

where $I(\cdot)$ denotes the indicator function. Likewise, for matrix norms, we define

$$\|\mathbf{M}\|_q := \max_{\|\boldsymbol{v}\|_q=1} \|\mathbf{M}\boldsymbol{v}\|_q, \qquad \|\mathbf{M}\|_{\max} := \max\{|M_{ij}|\} \quad \text{and}$$
$$\|\mathbf{M}\|_H := \sum_{j>k} I(\mathbf{M}_{jk} \neq 0).$$

We define diag($\mathbf{M}$) to be a diagonal matrix with diagonal values the same as that of $\mathbf{M}$ and with off-diagonal values zero.

2.1. *The nonparanormal.*

Liu, Lafferty and Wasserman (2009) and Liu et al. (2012) show the Gaussian graphical model can be relaxed to the nonparanormal graphical model without significant loss of inference power when the data are actually Gaussian distributed and with significant gain of inference power when they are not. This observation plays a role in our proposed model for relaxing the Gaussian assumption. In this section, the nonparanormal distribution family is introduced with the corresponding graphical model, following definitions in Liu et al. (2012).

DEFINITION 2.1 (The nonparanormal). Let $f = \{f_j\}_{j=1}^d$ be a set of univariate strictly increasing functions. A d-dimensional random vector $\boldsymbol{X} = (X_1, \dots, X_d)^T$ is said to follow a nonparanormal distribution, denoted $NPN_d(\boldsymbol{\Sigma}, f)$, if and only if

$$f(\boldsymbol{X}) := \{f_1(X_1), \dots, f_d(X_d)\}^T \sim N_d(\mathbf{0}, \boldsymbol{\Sigma}) \quad \text{where } \operatorname{diag}(\boldsymbol{\Sigma}) = \mathbf{I}_d,$$

where $\mathbf{I}_d \in \mathbb{R}^{d \times d}$ is the identity matrix. $\boldsymbol{\Sigma}$ is called the *latent correlation matrix*, and $\boldsymbol{\Omega} := \boldsymbol{\Sigma}^{-1}$ is called the *latent concentration matrix*.

Although the nonparanormal is strictly larger than the Gaussian, Liu, Lafferty and Wasserman (2009) show the conditional independence property of the nonparanormal is still encoded in the latent concentration matrix $\boldsymbol{\Omega}$.

2.2. *Rank-based estimator.* Liu et al. (2012) and Xue and Zou (2012) exploit the rank-based estimator, Kendall's tau, in inferring the latent concentration matrix $\boldsymbol{\Omega}$ in the nonparanormal family. Let $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n \in \mathbb{R}^d$, with $\boldsymbol{x}_i = (x_{i1}, \dots, x_{id})^T$ for $i = 1, \dots, n$, be n observed data points of a random vector $\boldsymbol{X}$. The Kendall's tau statistic is defined as

$$\widehat{\tau}_{jk}(\boldsymbol{x}_1, \dots, \boldsymbol{x}_n) := \frac{2}{n(n-1)} \sum_{1 \le i < i' \le n} \operatorname{sign}(x_{ij} - x_{i'j}) \cdot \operatorname{sign}(x_{ik} - x_{i'k}). \tag{2.1}$$

The Kendall's tau statistic is a monotone, transformation-invariant correlation between the empirical realizations of X_j and X_k for any $j, k \in \{1, \dots, d\}$. Let $\widehat{\mathbf{R}} = [\widehat{\mathbf{R}}_{jk}] \in \mathbb{R}^{d \times d}$, with

$$\widehat{\mathbf{R}}_{jk} = \sin\left(\frac{\pi}{2}\widehat{\tau}_{jk}(\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)\right), \tag{2.2}$$

be the Kendall's tau matrix. Liu et al. (2012) show that, if $\boldsymbol{X}$ is nonparanormally distributed, $\widehat{\mathbf{R}}$ is a consistent estimator of the latent correlation matrix $\boldsymbol{\Sigma}$ (with respect to element-wise sup norm $\|\cdot\|_{\max}$), even when the order of d is nearly exponentially larger than n.

Since the latent concentration matrix, $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$, fully encodes the nonparanormal graphical model, and $\widehat{\mathbf{R}}$ is a consistent estimator of $\boldsymbol{\Sigma}$, Kendall's tau is a good estimate of the nonparanormal graphical model, as it directly estimates the latent concentration matrix. Based on Kendall's tau matrix, Liu et al. (2012) propose the nonparanormal SKEPTIC by directly plugging $\widehat{\mathbf{R}}$ into any statistical methods in calculating the inverse covariance/correlation matrix. In this paper, we focus on one particular statistical method, CLIME [Cai, Liu and Luo (2011)]. Further details of the nonparanormal SKEPTIC are given in Section 4.

REMARK 2.2. If the underlying distribution is Gaussian, there is very little cost in using the rank-based procedure. In detail, empirically, Liu et al. (2012) show that the graphical model estimates based on Pearson's sample correlation and Kendall's tau have comparable performance with insignificant differences in this setting. Theoretically, in a low dimension, Xu et al. (2010) show the asymptotic variance of Kendall's tau estimates is very close to the Gaussian-based counterpart; in high dimensions, Liu et al. (2012) and Han and Liu (2014) further show in high dimensions the theoretical performances are also comparable.

3. Models and concepts.

3.1. *Models*. This section gives the proposed approach for modeling complex aggregated data. Assume the data are aggregated from multiple datasets, each of which is distributed according to a different nonparanormal distribution.

More specifically, let $\boldsymbol{X}_1, \dots, \boldsymbol{X}_T$ be T different random vectors with $\boldsymbol{X}_t := (X_{t1}, \dots, X_{td})^T$ satisfying

$$\boldsymbol{X}_t \sim NPN_d\big(\boldsymbol{\Sigma}^t, f^t\big) \qquad \text{for } t = 1, \dots, T.$$

Let $\boldsymbol{\Theta}^t := [\boldsymbol{\Sigma}^t]^{-1}$ denote the concentration matrix of $\boldsymbol{X}_t$. Based on $\boldsymbol{\Theta}^t$, we define $\mathcal{G}^t = [\mathcal{G}^t_{jk}] \in \{0, 1\}^{d\times d}$ where

$$\mathcal{G}^t_{jk} = 0 \quad \text{if and only if} \quad \boldsymbol{\Theta}^t_{jk} = 0.$$

Here $\mathcal{G}^t$ represents the Markov graph associated with $\boldsymbol{X}_t$. In detail, the pair (j, k) such that $\mathcal{G}^t_{jk} \neq 0$ indicates the conditional independence of X_{tj} and X_{tk} given all the rest in $\boldsymbol{X}_t$.

3.2. *Sparse median graphs*. This section introduces the concept of a *sparse median graph*, combining the ideas of median graphs from Jiang, Munger and Bunke (2001) and the sparsity concept commonly adopted in high-dimensional statistics [Bühlmann and van de Geer (2011)]. In the following, we write

$$\mathcal{B}(d) := \big\{\mathcal{G} \in \{0, 1\}^{d\times d}, \mathcal{G} \text{ is symmetric with diagonal entries all equal to } 0\big\}.$$

Let $\boldsymbol{d}(\cdot) : \mathcal{B}(d) \times \mathcal{B}(d) \to [0, \infty)$ be a distance function on the graph space. Jiang, Munger and Bunke (2001) define the median graph (reproduced in Definition 3.1 below) as the graph that has the smallest sum of distances to all graphs in a given set.

DEFINITION 3.1 (Median graph). Let $\mathcal{G}^1, \dots, \mathcal{G}^T$ be T different binary graphs in $\mathcal{B}(d)$, and the median graph $\mathcal{G}^*$ is defined by

$$\mathcal{G}^* := \underset{\mathcal{G} \in \mathcal{B}(d)}{\operatorname{argmin}} \sum_{t=1}^{T} \boldsymbol{d}(\mathcal{G}, \mathcal{G}^t). \tag{3.1}$$

When T is large, $\mathcal{G}^*$ will not be sparse and, therefore, the resulting median graph may not be interpretable. To attack this problem, consider the concept of a "sparse median graph." The sparse median graph is the graph that has the smallest sum of distances to all graphs in a given set. In addition, we require the number of nonzero entries in the graph to be less than or equal to a small value $s \ll d^2$.

We use the Hamming distance $\| \cdot \|_H$ in calculating the distance of any two graphs.

DEFINITION 3.2 (Sparse median graph). Let $\{\mathcal{G}^1, \dots, \mathcal{G}^T\}$ be T different binary graphs. The sparse median graph $\mathcal{G}_s^*$ is defined as

$$\mathcal{G}_s^* := \underset{\mathcal{G} \in \mathcal{B}(d), \|\mathcal{G}\|_H \le s}{\operatorname{argmin}} \sum_{t=1}^{T} \|\mathcal{G} - \mathcal{G}^t\|_H, \tag{3.2}$$

where $\| \cdot \|_H$ represents the number of nonzero entries in the upper triangle of the matrix of interest.

The next proposition presents an equivalent representation of $\mathcal{G}_s^*$ and further discusses identifiability conditions of the model. To this end, we first introduce some additional notation. For a series of numbers $\{a_1, \dots, a_n\}$, let $a^{(1)} \ge a^{(2)} \ge \cdots \ge a^{(n)}$ be an arbitrary sorted arrangement of $a_1, \dots, a_n$. Then the rank of a_i is defined as the set

$$\{j : a_i = a^{(j)} \text{ for some sorted arrangement of } \{a_1, \dots, a_n\}\}.$$

For a set $r = \{a_1, \dots, a_n\}$ and any number m, we write $r \preceq m$ if $a_i \le m$ for all i. We write $r \succ m$ if there exists at least one a_i such that $a_i > m$.

PROPOSITION 3.3. *Let $\mathcal{G}^t, t = 1, \dots, T$ and $\mathcal{G}_s^*$ be the sparse median graph. Let $\zeta_{jk} = \sum_t \mathcal{G}_{jk}^t$ and r_{jk} be the rank of ζ_{jk} in all values $\{\zeta_{j'k'}\}_{j'<k'}$. Then, if there are no ties around the rank s for the sequence $\{\zeta_{j'k'}\}_{j'<k'}$, we have*

$$[\mathcal{G}_s^*]_{jk} = [\mathcal{G}_s^*]_{kj} = \begin{cases} 1, & \text{if } r_{jk} \preceq s, \\ 0, & \text{if } r_{jk} \succ s. \end{cases} \tag{3.3}$$

Equivalently, given T graphs, $\{\mathcal{G}^t\}_{t=1}^T$, their s-sparse median graph, $\mathcal{G}_s^*$, is given by the indicator function of the s largest upper off-diagonal entries of their average. Moreover, the model is identifiable with respect to $\mathcal{G}_s^*$ if and only if there are no ties around the rank s for the sequence $\{\zeta_{j'k'}\}_{j'<k'}$.

PROOF. We first prove that (3.3) holds. Note that, for any $\mathcal{G}, \mathcal{G}^1, \dots, \mathcal{G}^T$, we have

$$\begin{aligned}
\operatorname*{argmin}_{\mathcal{G}\in\mathcal{B}(d), \|\mathcal{G}\|_H \le s} \sum_{t=1}^{T} \|\mathcal{G} - \mathcal{G}^t\|_H &= \operatorname*{argmin}_{\mathcal{G}\in\mathcal{B}(d), \|\mathcal{G}\|_H \le s} \sum_{1\le j<k\le d} \sum_{t=1}^{T} I(\mathcal{G}^t_{jk} \neq \mathcal{G}_{jk}) \\
&= \operatorname*{argmax}_{\mathcal{G}\in\mathcal{B}(d), \|\mathcal{G}\|_H \le s} \sum_{1\le j<k\le d} \sum_{t=1}^{T} I(\mathcal{G}^t_{jk} = \mathcal{G}_{jk}) \\
&= \operatorname*{argmax}_{\mathcal{G}\in\mathcal{B}(d), \|\mathcal{G}\|_H \le s} \sum_{\{(j,k):\mathcal{G}_{jk}\neq 0, j<k\}} \zeta_{jk}.
\end{aligned}$$

Hence, to minimize $\sum_{t=1}^{T} \|\mathcal{G} - \mathcal{G}^t\|_H$ over $\{\mathcal{G} \in \mathcal{B}(d), \|\mathcal{G}\|_H \le s\}$, it is equivalent to maximize $\sum_{j<k} \zeta_{jk}$ for all different s pairs of $\{(j,k), j<k\}$. In particular, when there are no ties around the rank s for the sequence $\{\zeta_{j'k'}\}_{j'<k'}$, the minimum of $\sum_{t=1}^{T} \|\mathcal{G} - \mathcal{G}^t\|_H$ over $\{\mathcal{G} \in \mathcal{B}(d), \|\mathcal{G}\|_H \le s\}$ is attained as in (3.3). This completes the proof of the first assertion.

We then turn to study the second assertion. For this, on one hand, assume there are ties around the rank s. In other words, there exist at least two sets of entries (j_1, k_1) and (j_2, k_2), such that $s \in r_{j_1,k_1} = r_{j_2,k_2}$. Then $\mathcal{G}^*_s$, picking either $[\mathcal{G}^*_s]_{j_1,k_1} = 1$ or $[\mathcal{G}^*_s]_{j_2,k_2} = 1$, attains the same minimum to equation (3.2). This is in contradiction to the model identifiability assumption. On the other hand, if there does not exist a tie around the rank s, then we can determine the unique top s pairs for the solution to (3.2), and hence the model is identifiable. □

REMARK 3.4. The population sparse median graph is defined as the optimum of a specified loss function with regard to the Hamming distance. This is a common approach for representing a summary of multiple, possibly heterogeneous, data points. In principle, there are potential issues by aggregation, such as averaging out effects when both positive and negative ones exist. However, since we focus only on undirected graphs taking values $\{0, 1\}$, such issues can be minimized. Actually, the robustness to aggregation issues is one strong advantage motivating the sparse median graph. In Section 6.2, we will further illustrate the empirical power of using the notion of the sparse median graph combined with robust estimation.

REMARK 3.5. The sparse median graph (SMG) is a summarization graph across different subjects. Therefore, instead of depicting the conditional independence structure among the covariates of each specific subject, the SMG depicts the edges that are present across most subjects' conditional independence graphs. In other words, there exists an edge between the jth and kth nodes in the SMG if and only if, for most subjects, these two nodes are dependent conditional on all the other nodes.

4. Methods. For $t = 1, \dots, T$, let $\boldsymbol{x}_i^t = (x_{i1}^t, \dots, x_{id}^t)^T, i = 1, \dots, n_t$ be n_t independent realizations of $\boldsymbol{X}_t$ (defined in Section 3.1). The observed data are $\{\boldsymbol{x}_i^t\}$ for $t = 1, \dots, T$ and $i = 1, \dots, n_t$, and the target is to estimate the sparse median graph $\mathcal{G}_s^*$, defined in (3.2). The proposed method is a two step procedure. In the first step, the nonparanormal SKEPTIC is used to obtain the estimators $\{\widehat{\mathcal{G}}^t\}_{t=1}^T$ of $\{\mathcal{G}^t\}_{t=1}^T$. In the second step, $\mathcal{G}_s^*$ is estimated based on the estimators $\{\widehat{\mathcal{G}}^t\}_{t=1}^T$ obtained in the first step.

More specifically, in the first step, for each $t \in \{1, 2, \dots, T\}$, let

$$\widehat{\mathbf{R}}_{jk}^t := \sin\left(\frac{\pi}{2}\widehat{\tau}_{jk}(\boldsymbol{x}_1^t, \dots, \boldsymbol{x}_{n_t}^t)\right),$$

where $\widehat{\tau}_{jk}(\cdot)$ is defined in (2.1). By using $\widehat{\mathbf{R}}^t = [\widehat{\mathbf{R}}_{jk}^t] \in \mathbb{R}^{d \times d}$ to estimate $\boldsymbol{\Sigma}^t$, one can plug $\widehat{\mathbf{R}}^t$ into CLIME to get estimates of $\boldsymbol{\Omega}^t$ and $\mathcal{G}^t$:

$$\text{(4.1)} \qquad \widehat{\boldsymbol{\Omega}}^t = \underset{\mathbf{M}}{\arg\min} \sum_{j,k} |\mathbf{M}_{jk}| \qquad \text{such that } \|\widehat{\mathbf{R}}^t \mathbf{M} - \mathbf{I}_d\|_{\max} \le \lambda_t,$$

where $\lambda_t > 0$ is a tuning parameter. Cai, Liu and Luo (2011) show this optimization can be decomposed into d vector minimization problems, each of which can be reformulated as a linear program. Thus, it has the potential to scale to very large problems. Once $\widehat{\boldsymbol{\Omega}}^t$ is obtained, one can apply an additional thresholding step to estimate $\mathcal{G}^t$. For this, the graph estimator $\widehat{\mathcal{G}}^t \in \mathcal{B}(d)$ is defined, in which a pair (j, k) satisfies $\widehat{\mathcal{G}}_{jk}^t \neq 0$ if and only if $\widehat{\boldsymbol{\Omega}}_{jk}^t > \gamma_t$. Here, γ_t is another tuning parameter. However, in practice, the CLIME algorithm works well without a second step truncation.

In the second step, provided the estimates $\{\widehat{\mathcal{G}}^t, t = 1, \dots, T\}$ have been obtained, the following equation is optimized to obtain $\widehat{\mathcal{G}}_s^*$:

$$\text{(4.2)} \qquad \widehat{\mathcal{G}}_s^* = \underset{\mathcal{G} \in \mathcal{B}(d), \|\mathcal{G}\|_H \le s}{\operatorname{argmin}} \sum_t \|\mathcal{G} - \widehat{\mathcal{G}}^t\|_H,$$

where the term $\|\mathcal{G}\|_H \le s$ controls the sparsity degree of $\mathcal{G}$. For presentation clearness, we assume s is known in the following. In Section 6, we will further discuss how to choose s.

Of note, let $\widehat{\zeta}_{jk}$ be defined as $\widehat{\zeta}_{jk} := \sum_t \widehat{\mathcal{G}}_{jk}^t$. Let $(j_1, k_1), (j_2, k_2), \dots$ be s pairs with the highest values in $\{\widehat{\zeta}_{jk}\}_{j<k}$. Using a similar argument as in Proposition 3.3 yields $\widehat{\mathcal{G}}_{jk} = 1$ if and only if $(j, k) \in \{(j_1, k_1), (j_2, k_2), \dots\}$.

REMARK 4.1. For simplicity, it is assumed there are no ties around the rank s for the sequence $\{\widehat{\zeta}_{jk}\}$. If the model discussed in Section 3 is identifiable and several mild conditions as shown in Section 5 hold, then there are indeed no ties with high probability.

5. Theoretical properties. In this section, the estimators from Section 4 are proven to be consistent for the true median graph. Notably, a nonasymptotic bound on the rate of convergence in estimating the sparse median graph with respect to the Hamming distance is provided.

Additional notation is required. Let M_d be a quantity which may scale with the dimension d. Define

$$\mathscr{S}_d(q, s, M_d) := \left\{ \mathbf{\Omega} : \|\mathbf{\Omega}\|_1 \le M_d \text{ and } \max_{1 \le j \le d} \sum_{k=1}^{d} |\mathbf{\Omega}_{jk}|^q \le s \right\}.$$

For $q = 0$, the class $\mathscr{S}_d(q, s, M_d)$ contains all the s-sparse matrices with the ℓ_1 norm bounded above by M_d. The next theorem provides the parameter estimation and graph estimation consistency results for the nonparanormal SKEPTIC estimator defined in (4.1).

THEOREM 5.1 [Liu et al. (2012)]. *Let $\boldsymbol{X}^t \sim NPN_d(\mathbf{\Sigma}^t, f^t)$ with $\mathbf{\Omega}^t := [\mathbf{\Sigma}^t]^{-1} \in \mathscr{S}_d(q, s_t, M_d)$ with $0 \le q < 1$. Let $\widehat{\mathbf{\Omega}}^t$ be defined in (4.1). There exist constants, C_0 and C_1, only depending on q, such that whenever one chooses the tuning parameter $\lambda_t = C_0 M_d \sqrt{\frac{\log d}{n_t}}$, with probability no less than $1 - d^{-2}$,*

$$\|\widehat{\mathbf{\Omega}}^t - \mathbf{\Omega}^t\|_2 \le C_1 M_d^{2-2q} \cdot s \cdot \left(\frac{\log d}{n_t}\right)^{(1-q)/2}.$$

Let $\widehat{\mathcal{G}}^t$ be the graph estimator defined in Section 4 with the second tuning parameter $\gamma_t := 4M_d\lambda_t$. If it is further assumed $\mathbf{\Omega} \in \mathscr{S}_d(0, s, M_d)$ and $\min_{j,k:\mathbf{\Omega}_{jk} \neq 0} |\mathbf{\Omega}_{jk}| \ge 2\gamma_t$, then

$$\mathbb{P}(\widehat{\mathcal{G}}^t \neq \mathcal{G}^t) \le 4d^{-\varepsilon_1},$$

where $\varepsilon_1 > 0$ is a constant that does not depend on (n_t, d, s_t).

PROOF. We combine Theorems 1 and 7 in Cai, Liu and Luo (2011) and Theorem 4.2 in Liu et al. (2012). □

THEOREM 5.2 (Consistency). *With the above notation, if the assumptions from Theorem 5.1 hold, the parameter $q = 0$, the parameters λ_t and γ_t are fixed, and the model in Section 3 is identifiable, then*

$$\mathbb{P}(\widehat{\mathcal{G}}_s^* \neq \mathcal{G}_s^*) \le 4Td^{-\varepsilon_1}, \tag{5.1}$$

where $\widehat{\mathcal{G}}_s^*$ is defined as in (4.2).

PROOF. If the model is identifiable, then one only needs to show, with high probability, all $\mathcal{G}^t$ can be recovered. Note the union bound in Theorem 5.1 yields

$$\mathbb{P}\left(\bigcup_{t=1}^{T} \{\widehat{\mathcal{G}}^t \neq \mathcal{G}^t\}\right) \le \sum_{t=1}^{T} \mathbb{P}(\widehat{\mathcal{G}}^t \neq \mathcal{G}^t) \le 4d^{-\varepsilon_1} \le 4Td^{-\varepsilon_1}.$$

This completes the proof. □

The next theorem provides an upper bound of the rate of convergence with respect to the Hamming distance. Such a result is based on the recent explorations in graph recovery with respect to the Hamming distance [Jin, Zhang and Zhang (2014), Ke, Jin and Fan (2014)].

THEOREM 5.3 (Rate of convergence). *Assume the above assumptions in Theorems 5.1 and 5.2 hold. Let $\mathcal{A}_t$ be the event*

$$\mathcal{A}_t := \{\|\widehat{\mathcal{G}}^t - \mathcal{G}^t\|_H \le \delta_t\},$$

and let δ_t be defined as a random number, depending on n_t, d, s_t, M_d, such that $\mathbb{P}(\mathcal{A}_t) = 1 - o(d^{-\varepsilon_2})$. Moreover, reorder $\{\zeta_{jk}\}_{j<k}$ to be $\zeta^{(1)} \ge \zeta^2 \ge \cdots \ge \zeta^{d(d-1)/2}$, and let $u^* = (\zeta^{(s)} - \zeta^{(s+1)})/2$. Then

$$\mathbb{P}\biggl(\|\mathcal{G}_s^* - \widehat{\mathcal{G}}_s^*\|_H \le \frac{2\sum_{t=1}^T \delta_t}{u^*}\biggr) = 1 - o(Td^{-\varepsilon_2}). \tag{5.2}$$

PROOF. Let $\kappa^* := (\zeta^{(s)} + \zeta^{(s+1)})/2$. We reorder $\{\widehat{\zeta}_{jk}\}_{j<k}$ to be $\widehat{\zeta}^{(1)} \ge \widehat{\zeta}^{(2)} \ge \cdots \ge \widehat{\zeta}^{d(d-1)/2}$, and let $\widehat{\kappa}^* := (\widehat{\zeta}^{(s)} + \widehat{\zeta}^{(s+1)})/2$. Then $[\mathcal{G}_s^*]_{jk} \ne [\widehat{\mathcal{G}}_s^*]_{jk}$ if and only if

$$\operatorname{sign}(\widehat{\zeta}_{jk} - \widehat{\kappa}^*) \cdot \operatorname{sign}(\zeta_{jk} - \kappa^*) < 0.$$

Recall $\delta_t \in \mathbb{R}$ is defined such that

$$\mathbb{P}(\mathcal{A}_t) = \mathbb{P}(\|\widehat{\mathcal{G}}^t - \mathcal{G}^t\|_H > \delta_t) = o(d^{-\varepsilon_2}). \tag{5.3}$$

Note such a bound has been established in some constrained situations, for example, in Jin, Zhang and Zhang (2014) (Theorem 1.2).

Let $\bar{\mathcal{G}}^*$ be the graph defined as

$$\bar{\mathcal{G}}_{jk}^* = \bar{\mathcal{G}}_{kj}^* = \begin{cases} 1, & \text{if } \widehat{\zeta}_{jk} \ge \kappa^*, \\ 0, & \text{if } \widehat{\zeta}_{jk} < \kappa^*. \end{cases}$$

First, we consider quantifying the difference between $\mathcal{G}_s^*$ and $\bar{\mathcal{G}}^*$. Let $u_{jk} := |\zeta_{jk} - \kappa^*|$. We reorder $\{u_{jk}\}$ from the smallest to the largest such that $u^{(1)} \le u^{(2)} \le \cdots \le u^{(d(d-1)/2)}$. Let N^* be defined as

$$\sum_{t=1}^{N^*} u^{(t)} \le \sum_t \delta_t \quad \text{and} \quad \sum_{t=1}^{N^*+1} u^{(t)} > \sum_t \delta_t.$$

Then, conditioning on the event $\bigcap_t \mathcal{A}_t$, we have the difference between $\widehat{\mathcal{G}}^t$ and $\mathcal{G}^t$ with regard to the Hamming distance is at most $\sum_t \delta_t$, and therefore

$$\|\mathcal{G}_s^* - \bar{\mathcal{G}}^*\|_H \le N^*.$$

In particular, reminding that $u^* := (\zeta^{(s)} - \zeta^{(s+1)})/2 = \min_{(j,k)} u_{jk}$, we have

$$\|\mathcal{G}_s^* - \bar{\mathcal{G}}^*\|_H \le \frac{\sum_{t=1}^T \delta_t}{u^*}. \tag{5.4}$$

Consider now quantifying the difference between $\bar{\mathcal{G}}^*$ and $\widehat{\mathcal{G}}_s^*$. Using (5.4) and the fact $\|\mathcal{G}_s^*\|_H = s$, one obtains

$$\min\left(0, s - \frac{\sum \delta_t}{u^*}\right) \le \|\bar{\mathcal{G}}^*\|_H \le s + \frac{\sum \delta_t}{u^*}.$$

Combining with the fact $\|\widehat{\mathcal{G}}_s^*\|_H = s$, then

$$\|\widehat{\mathcal{G}}_s^* - \bar{\mathcal{G}}^*\|_H \le \frac{\sum_{t=1}^T \delta_t}{u^*}.$$

Accordingly, by the triangle inequality, with high probability,

$$\|\widehat{\mathcal{G}}_s^* - \mathcal{G}_s^*\|_H \le \|\widehat{\mathcal{G}}_s^* - \bar{\mathcal{G}}^*\|_H + \|\bar{\mathcal{G}}^* - \mathcal{G}_s^*\|_H \le \frac{2\sum_{t=1}^T \delta_t}{u^*}.$$

This completes the proof. □

REMARK 5.4. The bound constructed in (5.2) is to balance the difference of $\{\mathcal{G}^t\}_{t=1}^T$ to $\mathcal{G}_s^*$ and the estimation error of $\widehat{\mathcal{G}}^t$ to $\mathcal{G}^t$. In other words, the better it is to differentiate $\{\mathcal{G}^t\}$ with $\mathcal{G}_s^*$ in the population level and the more accurately $\widehat{\mathcal{G}}^t$ can approach $\mathcal{G}^t$, the better the final estimator can recover the sparse median graph.

6. Empirical results.

In this section, we investigate the performance of the proposed method compared to the performances of alternative methods on synthetic and real-world datasets. Since we aim to estimate a summary graph throughout multiple, possibly non-i.i.d., datasets, our estimation procedure involves two steps: In the first step, for each specific dataset, we employ a graphical model estimation procedure; in the second step, based on the calculated graph estimates, we obtain a single estimate of the summary graph. We call the former step the "estimation of graphs" part and the latter step the "combination[1] of datasets" part. In the following simulations and experiments, we will compare our methods with multiple candidates using different graph estimation and dataset combination approaches, and reveal the advantage of our proposed one.

6.1. *Estimation methods.*

In our simulations and experiments, we consider the methods Kendall, Pearson and LW (detailed definitions provided later) to estimate graphs (or correlation matrices) on individual datasets. To combine multiple datasets, we employ SMG, Naive and Average (detailed definitions provided later).

[1] We use the terms "combination," "aggregation," and "summarization" synonymously in this work.

Therefore, we will compare a total of nine methods, each of which is denoted by first stating the aggregation method and then the graph estimation method. For example, our proposed method corresponds to SMG Kendall. We elaborate the details of the competing methods as follows.

6.1.1. *Estimation of graphs*. For any individual dataset, we consider the following approaches for graph estimation:

Kendall: This method calculates the Kendall's tau correlation matrix and plugs the matrix into CLIME. Details are in Section 4.

Pearson: This method follows the same steps as Kendall except we plug the Pearson sample correlation matrix into CLIME instead.

Ledoit–Wolf (LW): Using the `tawny` package [Rowe (2014)], this method calculates the Ledoit–Wolf shrinkage estimation [Ledoit and Wolf (2003)] of the covariance matrix of the dataset, $\widehat{\mathbf{\Sigma}}$, and a corresponding precision matrix, $\widehat{\mathbf{\Theta}} = \widehat{\mathbf{\Sigma}}^{-1}$. We employ a threshold on $\widehat{\mathbf{\Theta}}$ such that the sparsity of the induced graph is as close as possible to the sparsities of the corresponding Kendall- and Pearson-based graphs.

We select the tuning parameters $\{\lambda_t\}$ in CLIME[2] using the StARS stability-based approach [Liu, Roeder and Wasserman (2010)]. StARS selects a tuning parameter that simultaneously makes a graph sparse and replicable under random sampling. The detailed procedure can be found in Section 3.2 in Liu, Roeder and Wasserman (2010).

6.1.2. *Combination of datasets*. Our experiments involve inference on T datasets, where each dataset corresponds to a different subject. We consider the following approaches to estimate one sparse graph across the multiple datasets:

Sparse Median Graph (SMG): We estimate a graph for each of the T datasets. Then, given some sparsity[3] s, we combine these graphs with the method proposed in Section 3.2 to obtain a sparse median graph.

Naive: We concatenate the T datasets into one centered dataset on which we estimate a graph using the techniques from Section 6.1.1.

Average: For each of the T datasets, we calculate an associated correlation or precision matrix. We average these matrices and threshold such that only the s entries in the averaged matrix with the largest magnitudes correspond to edges in the estimated graph.

[2] Recall the formal definition of CLIME also requires a set of thresholding parameters $\{\gamma_t\}$. We choose to set $\gamma_t = 0$ for all t. While thresholding by some small $\gamma_t > 0$ is indeed an option, we have found this has very little impact on the output of the method compared to $\gamma_t = 0$.

[3] In our experiments, where s is unknown, we set s to be the median of the sparsities of the graphs estimated on each of the T datasets.

6.2. *Synthetic data simulations.* In this simulation, we examine the estimation performance of the proposed method on synthetically generated data. In particular, we generate $T = 15$ different datasets with $n_t = 100$ samples in each dataset. Each dataset follows a different nonparanormal distribution, corresponding to a different undirected graph $\mathcal{G}^t$. For each method, we use a sequence of uniformly spaced sparsity parameters $\widehat{s}$ from 0 to $\binom{d-1}{2}$ to estimate a sequence of graphs, over which we plot a ROC curve. In addition, we repeat this simulation for $d = 50$, 100 and 250. Our results show the SMG Kendall exhibits better estimation performance than the competing methods.

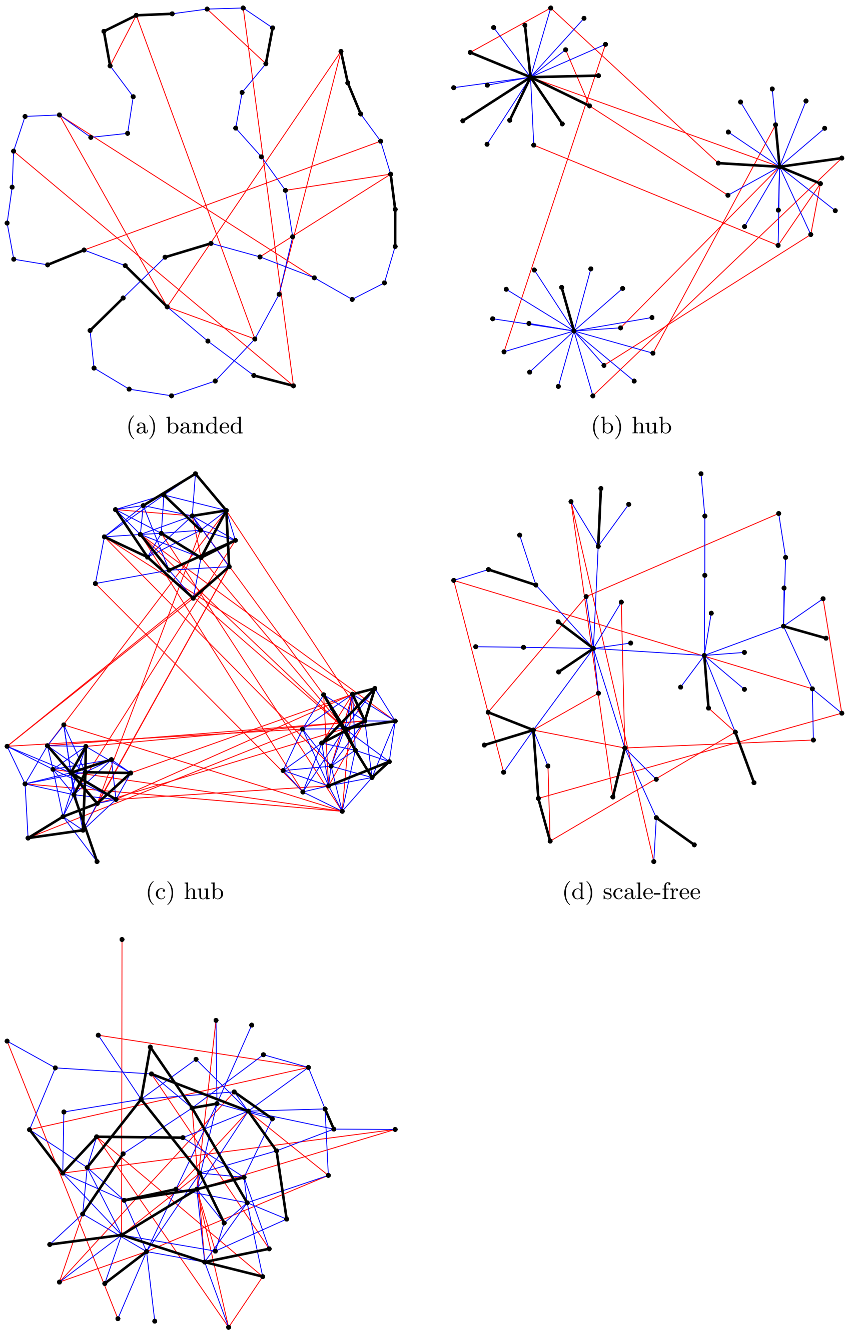
More specifically, we conduct the simulation with the following procedure:

1. Using the huge package [Zhao et al. (2012)], we generate a sparse graph $\mathcal{G}_s^*$ with sparsity s, along with a corresponding covariance matrix $\mathbf{\Sigma}$. We will use this as the oracle graph of the population. We adopt the following five models for $\mathcal{G}_s^*$: banded, clustered, hub, random and scale-free [definitions provided in Zhao et al. (2012)]. We examine $\mathcal{G}_s^*$ at $d = 50$, 100 and 250.
2. For each subject $t = 1, 2, \dots, T$, we construct a perturbed graph $\mathcal{G}^t$ to reflect the difference among different subjects. In particular, we add $\lfloor 0.001 \times (\binom{d-1}{2} - s) \rfloor$ edges and remove $\lfloor 0.75 \times s \rfloor$ edges from $\mathcal{G}_s^*$. We illustrate a typical run of the generated graphs $\mathcal{G}^t$ for a specific t in Figures 1, 2 and 3. In each figure, the black edges represent the ones present in both $\mathcal{G}_s^*$ and $\mathcal{G}^t$, the blue edges represent the ones only present in $\mathcal{G}_s^*$, and the red edges represent the ones only present in $\mathcal{G}^t$.
3. Using each $\mathcal{G}^t$, we generate a corresponding covariance matrix $\mathbf{\Sigma}^t$ with an algorithm identical to the one implemented in the huge package.
4. For $t = 1, \dots, T$, we generate a $(n_t \times d)$ dataset[4] $\mathcal{D}^t$ from $NPN_d(\mathbf{\Sigma}^t, f)$, where[5] $f_1(x) = \cdots = f_d(x) = x^5$. Thus, the population dataset is
$$\mathcal{D} = \{\mathcal{D}^t : t = 1, \dots, T\}.$$
5. Applying the nine methods described in Section 6.1 to $\mathcal{D}$, we estimate a sparse graph, $\widehat{\mathcal{G}_s^*}$, and calculate the true positive and true negative rates.
6. We repeat the simulation 100 times and plot an averaged ROC curve over the range of $\widehat{s}$. We show the results in Figures 4, 5 and 6.

From the curves in Figures 4, 5 and 6, we clearly see our proposed method exhibits a higher estimation performance than the competing methods. This is as expected because the proposed method is the only consistent estimator of $\mathcal{G}_s^*$, while all the competing methods deviate from the truth. In addition, observe that the Kendall-based methods tend to outperform the Pearson-based methods—a pattern

[4] Each $\mathcal{D}^t$ corresponds to a realization $\{\boldsymbol{x}_i^t\}$ for $i = 1, \dots, n_t$ from Section 4.

[5] In the case where the transformation is the identity, we found that all the Kendall- and Pearson-based methods perform almost identically (similar to the results cited in Remark 2.2). For conciseness, we omit this case from our presentation.



(a) banded

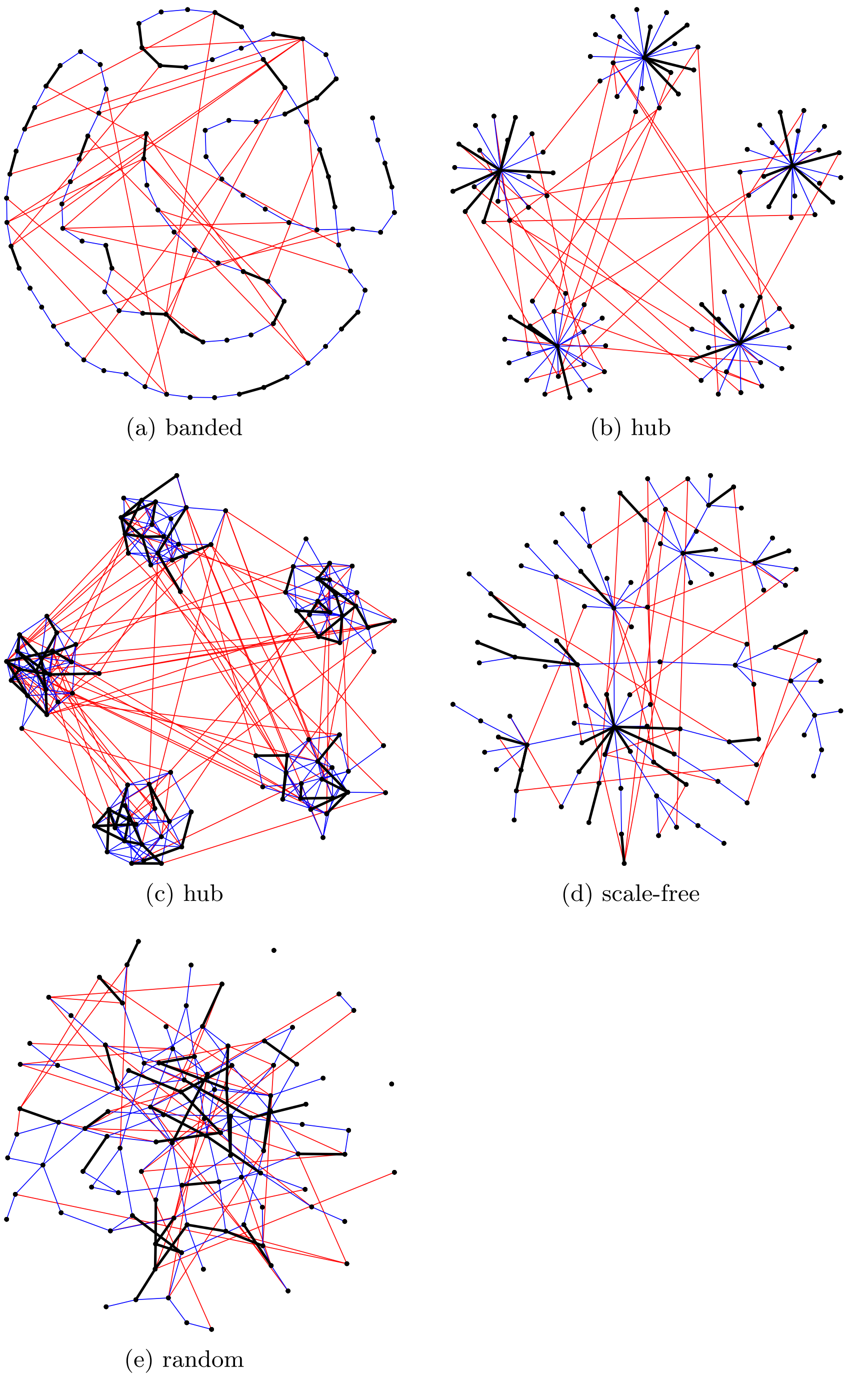
(b) hub

(c) hub

(d) scale-free

(e) random

FIG. 1. *An illustration of the five graph patterns with perturbations for $d = 50$. The black edges represent the ones present in both $\mathcal{G}_s^*$ and $\mathcal{G}^t$, the blue edges represent the ones only present in $\mathcal{G}_s^*$, and the red edges represent the ones only present in $\mathcal{G}^t$.*



(a) banded

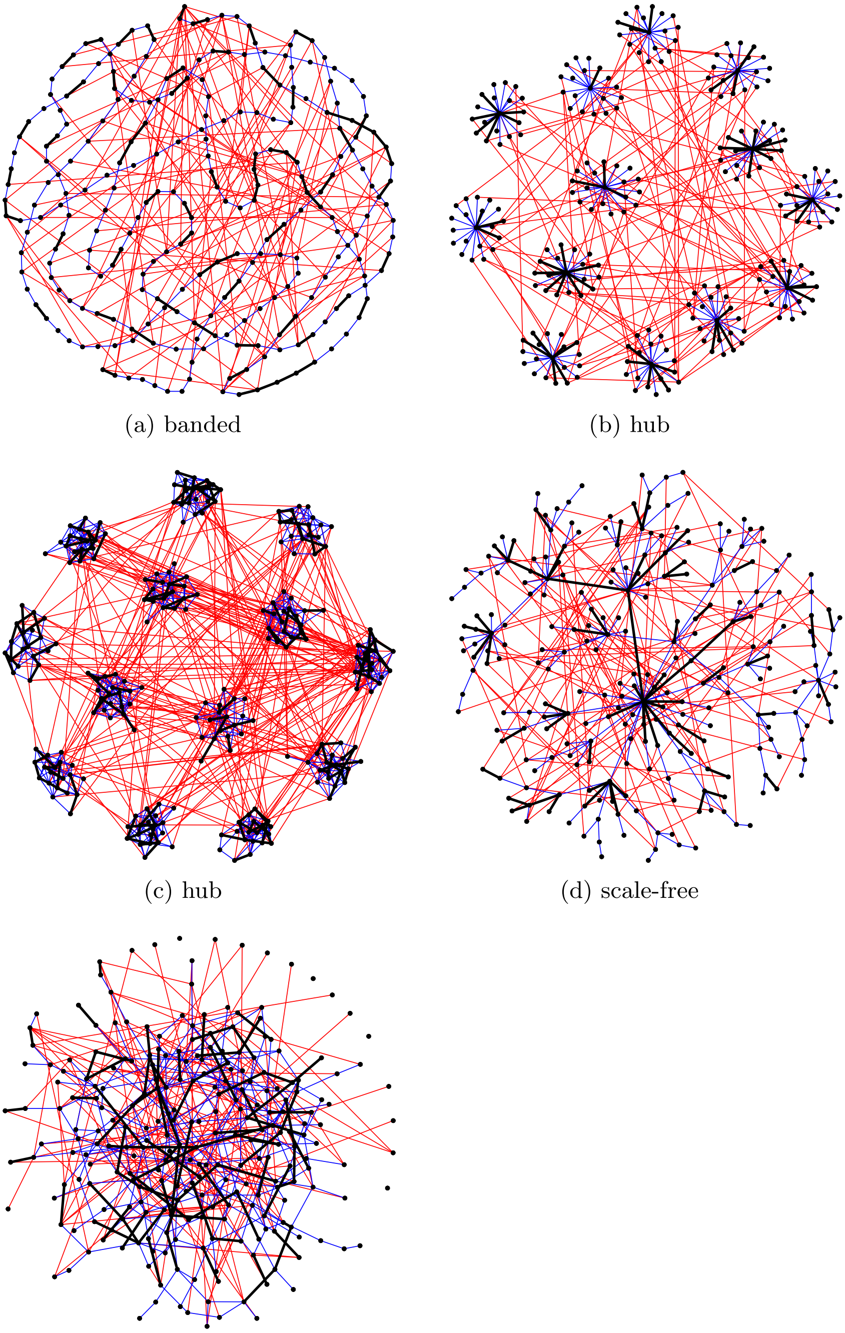
(b) hub

(c) hub

(d) scale-free

(e) random

FIG. 2. *An illustration of the five graph patterns with perturbations for $d = 100$. The black edges represent the ones present in both $\mathcal{G}_s^*$ and $\mathcal{G}^t$, the blue edges represent the ones only present in $\mathcal{G}_s^*$, and the red edges represent the ones only present in $\mathcal{G}^t$.*



(a) banded

(b) hub

(c) hub

(d) scale-free

(e) random

FIG. 3. *An illustration of the five graph patterns with perturbations for $d = 250$. The black edges represent the ones present in both $\mathcal{G}_s^*$ and $\mathcal{G}^t$, the blue edges represent the ones only present in $\mathcal{G}_s^*$, and the red edges represent the ones only present in $\mathcal{G}^t$.*

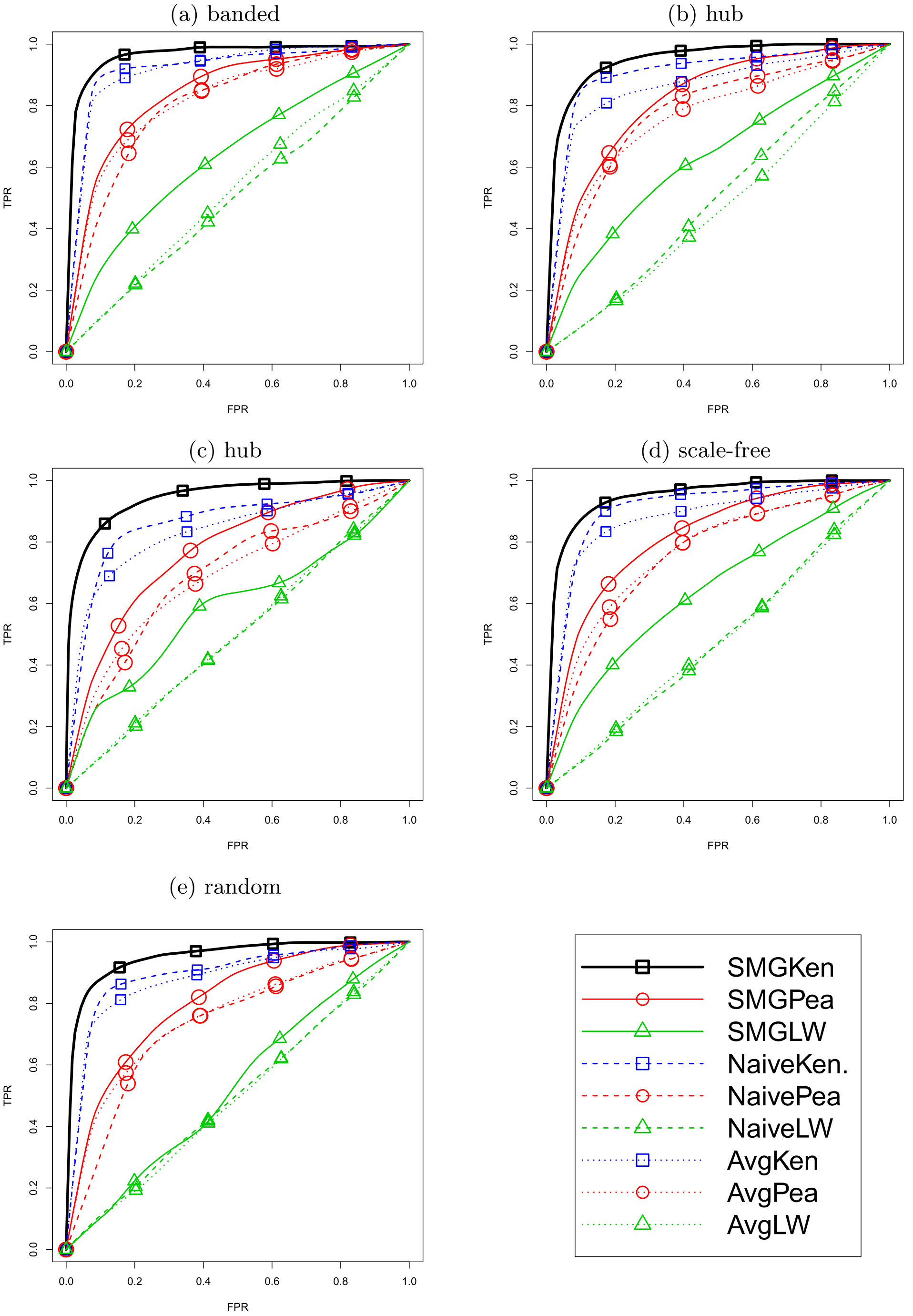


FIG. 4. *ROC curves in estimating the graphical models for different methods in five different graph patterns. Here,* $d = 50$ *and* $n_t = 100$ *for all* $t = 1, 2, \dots, 15$.

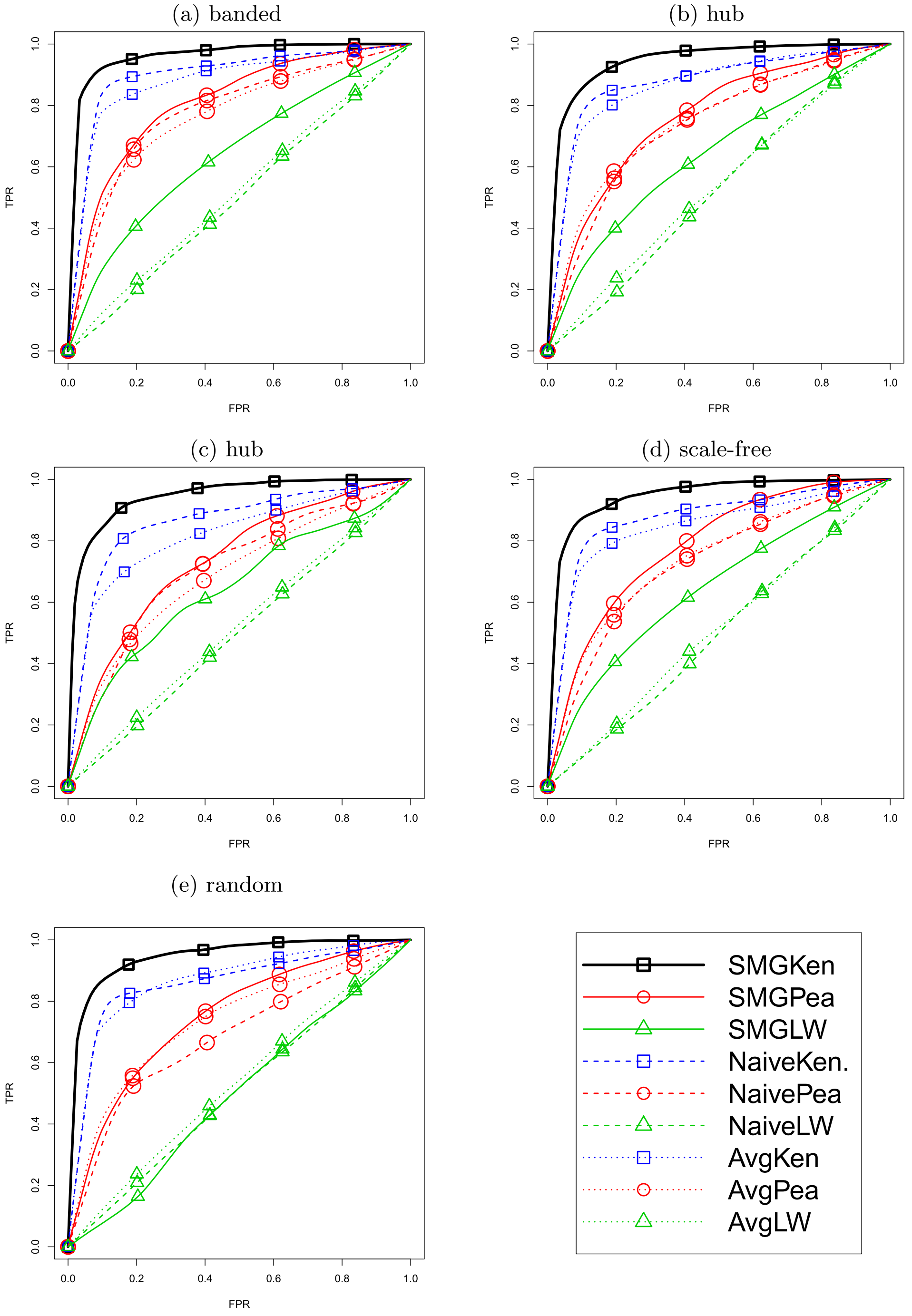


FIG. 5. *ROC curves in estimating the graphical models for different methods in five different graph patterns. Here*, $d = 100$ *and* $n_t = 100$ *for all* $t = 1, 2, \ldots, 15$.

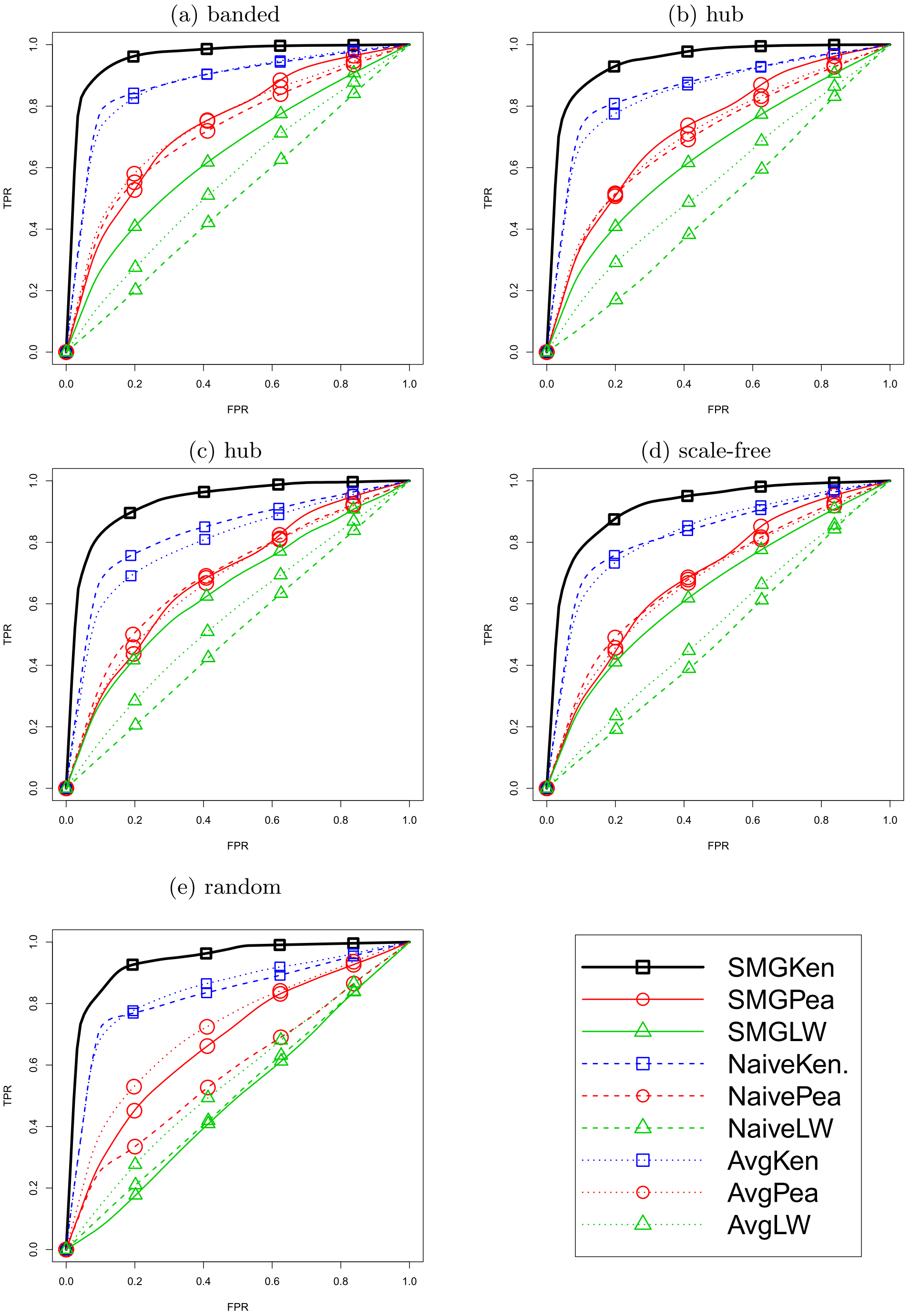


FIG. 6. *ROC curves in estimating the graphical models for different methods in five different graph patterns. Here,* $d = 250$ *and* $n_t = 100$ *for all* $t = 1, 2, \ldots, 15$.

that becomes more distinct in larger dimensions. This result confirms the claim that using Kendall's tau leads to optimal graph recovery rates [Liu et al. (2012)]. Furthermore, the poor performance of the LW-based methods (worse than both Kendall and Pearson) suggests that, while covariance shrinkage demonstrates potential in financial applications, its benefits do not carry over to graph estimation, at least under our simulation settings.

REMARK 6.1. An alternative simulation setting involves varying n_t. In our experiments, we found that this setting produces results almost identical to the case of constant n_t as long as they are in the same magnitude. (This phenomenon is a consequence of Theorem 5.3.) In addition, maintaining a constant n_t provides a better indication of the effects of increasing d. Therefore, for clarity, we omit the varying n_t case from the simulation section.

6.3. *ADHD data experiments.* In practice, there exists no gold standard for the structure of the oracle graph of brain imaging data. Therefore, in addition to the above simulation on synthetic data, we investigate the estimation performance, predictive power and stability of the proposed method on the ADHD-200 brain imaging dataset [Eloyan et al. (2012), Milham et al. (2012)].

The ADHD-200 dataset is a landmark study compiling over 1000 functional and structural scans including subjects with and without attention deficit hyperactive disorder (ADHD). The data used in the analysis are from 739 unique subjects: 478 controls and 261 children diagnosed with ADHD of various subtypes. Each has at least one blood oxygen level dependent (BOLD) resting state functional MRI scans. The scans were measured with different time resolutions (TR), different scan lengths and possibly during multiple sessions—causing the number of scans associated with one particular subject to range from 78 to 456. The varying TR and length of scanning stress the importance of addressing subject-level heteroscedasticity in graph estimates. The data also include demographic variables as predictors. These include age, IQ, gender and handedness. These demographic variables are combined into a matrix with dimensions 4×739. We follow the procedure in Eloyan et al. (2012) for data preprocessing with one additional step: The data collected from the same patient are concatenated together.

We construct our predictors by extracting 264 voxels from each image that broadly cover major functional regions of the cerebral cortex and cerebellum following Power et al. (2011). The locations of these 264 voxels are illustrated in Figure 7, and the value of each voxel is calculated as the mean of all data points inside each small seed region. Therefore, each subject t corresponds to a matrix of size $(n_t \times d)$, where n_t is the number of images and $d = 264$.

6.3.1. *ADHD data simulation.* Here, we examine the estimation performance of the proposed method on real brain imaging data. This involves three steps: first, we generate a "true graph"; second, we simulate datasets associated with different

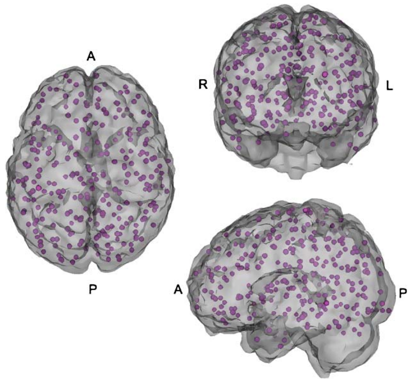


FIG. 7. *The illustration of the locations of the* 264 *nodes.*

"subjects"; third, we examine the estimation performance based on the simulated multiple datasets.

In detail, we first estimate a sparse graph on a homogeneous dataset and use this graph as the true graph. Then we simulate not identically distributed "subjects" by partitioning the homogeneous dataset and adding perturbations to each partition. Using these simulated datasets, we assess the estimation performance of the nine methods from Section 6.1 with a simulation similar to that in Section 6.2. Our results confirm SMG Kendall continues to exhibit better estimation performance than the competing methods when the data originate from the ADHD-200 dataset.

More specifically, we use the brain imaging data of the subject with the patient ID 15002 as our homogeneous dataset. This patient possesses the largest number of scans in the dataset with 456 images. We denote this dataset by $\mathcal{D}$. Then we implement the following simulation procedure:

1. Using the Kendall method described in Section 6.1.1, we estimate an oracle sparse median graph $\mathcal{G}_s^*$ on $\mathcal{D}$ with the s parameter chosen using StARS.
2. To simulate different datasets, we randomly partition $\mathcal{D}$ into $T = 10$ smaller datasets $\{\mathcal{D}_t : t = 1, \dots, T\}$. This creates six sets of 46 scans and four sets of 45 scans, each with $d = 264$.
3. For each "patient," $t = 1, 2, \dots, T$, we generate a graph $\mathcal{G}^t$ for the patient by removing edges from $\mathcal{G}_s^*$. More specifically, we select a $p_r := 50\%$ of the d vertices in $\mathcal{G}_s^*$ randomly, and delete all edges incident to these vertices.
4. Let μ and σ denote the mean and standard deviation of the vectorized $\mathcal{D}$. Note each of the $\lfloor p_r \times d \rfloor$ randomly selected vertices corresponds to a column in the datasets. We perturb each $\mathcal{D}_t$ to match $\mathcal{G}^t$ by replacing each entry of the randomly selected columns with a number randomly generated from the distribution $N(\mu, \sigma)$. Let us denote this perturbed dataset as $\widetilde{\mathcal{D}}_t$.

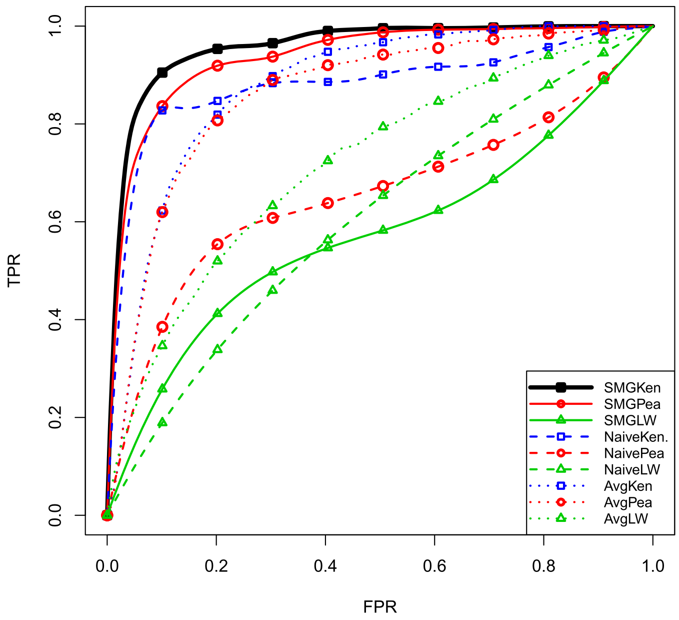


FIG. 8. *ROC curves in estimating the summary graphical models using data based on the dataset of the subject of ID* 15002 *in the* ADHD-200 *dataset*. *Here*, $d = 264$ *and* $T = 10$.

5. To simulate the effects of outliers, we choose 30% of the rows in each $\widetilde{\mathcal{D}}_t$ and apply the following transformation to each entry in a chosen row, i, in the dataset:

$$[\widehat{\mathcal{D}}_t]_{ij} = [\widetilde{\mathcal{D}}_t]_{ij}^5 \times \frac{\sum_{k=1}^d [\widetilde{\mathcal{D}}_t]_{ik}}{\sum_{k=1}^d [\widetilde{\mathcal{D}}_t]_{ik}^5} \qquad (\text{for} j = 1, 2, \ldots, d).$$

In rows that were not chosen, the entries of $\widehat{\mathcal{D}}_t$ and $\widetilde{\mathcal{D}}_t$ are identical. Therefore, $\widehat{\mathcal{D}}_t$ is the final perturbed dataset for one particular "subject," and the dataset of all simulated "subjects" is

$$\widehat{\mathcal{D}} = \{\widehat{\mathcal{D}}_t : t = 1, \ldots, T\}.$$

6. Applying the nine methods described in Section 6.1 to $\widehat{\mathcal{D}}$, we estimate a sparse graph, $\widehat{\mathcal{G}}_{\widehat{s}}^*$, and calculate the true and false positive rates.

7. We repeat the simulation 100 times and plot an averaged ROC curve over the range of $\widehat{s}$. The results are in Figure 8.

Comparing the results from Figure 8 to those in Section 6.2, we see the proposed method continues to demonstrate the best estimation performance, and the LW-based methods continue to perform the worst among the competing methods. However, in this simulation, SMG Kendall and SMG Pearson outperform Naive Kendall and Naive Pearson, where each Kendall-based method still outperforms the corresponding Pearson-based method. This suggests that the benefits of sparse median graphs tend to dominate when estimating graphs on real brain imaging data—unlike the synthetic setting where the benefits of utilizing Kendall's tau tend to dominate. Nonetheless, the results from this simulation and Section 6.2 both demonstrate the potential of the proposed method to improve the estimation accuracy of population-level networks.

6.3.2. *Predictive power experiment*. In this section, we compare the predictive power of our proposed method to that of the competing methods.[6] To this end, we examine the difference between summary graphs of different subpopulations.[7] In the sequel, we focus on SMG Kendall, SMG Pearson and Naive Kendall, which have performed the best in the previous simulations. (Avg Kendall and Avg Pearson also performed well, but we omit them because they are not robust to outliers.)

Several population sparse graph contrasts of interest are investigated and include the following: ADHD case status (denoted by *Case* and *Control*), gender (denoted by *Female* and *Male*) and age. Given the pediatric population in the ADHD study, we investigate young adults versus children using a cutoff of 12 years. Subjects having ages larger than 12 years are denoted by *Senior* and those less than or equal to 12 years denoted by *Junior*. Figure 9 provides a visual comparison of the brain connectivity graphs obtained using the three methods on *Case* and *Control* data. We observe that the *Case* and *Control* graphs show the most edge disagreements when estimated with SMG Kendall. This is consistent with the simulation results and strongly indicates the sparse median graph concept coupled with the Kendall's tau estimation procedure improves actual estimation in this context.

For the remaining subject classes (including ADHD case status), we provide more detailed analyses. We apply the three methods on subpopulations to compare graphs at different covariate levels. For example, graphs of cases and controls are investigated within gender. Summary statistics for these subpopulation differences are presented in Table 1. In addition, we include a case ("null" in Table 1) where we randomly divide the patients into two subpopulations, estimate a graph for each subpopulation, and calculate the difference between the two graphs.

For all the approaches, we observe that the differences in the real subpopulation splits are consistently larger than the difference in the "null" split. This provides

[6] We consider the predictive power of the methods in classification. Because classification power increases with greater separation between different classes, our experiment measures the predictive power by calculating the scaled Hamming distance between the sparse graphs estimated over two classes of data.

[7] All the remaining experiments use all patients in the ADHD dataset. For selecting tuning parameters—since we must estimate a graph for each patient, and parameter selection is computationally expensive—we randomly sample 100 subjects from the 739 subjects and apply StARS to estimate the CLIME parameter for each subject. Then, we find the median valued parameter among the 100 selected parameters and use it as the universal parameter for all applications of CLIME in the following experiments. We find the median parameter from using both Kendall and Pearson is $\lambda = 0.171$.

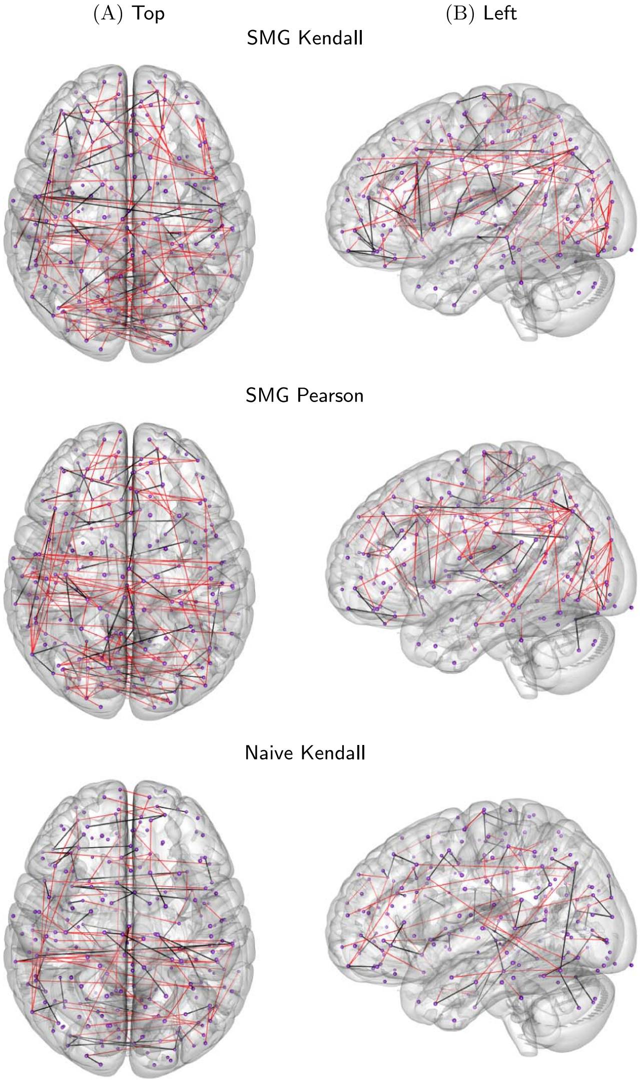


FIG. 9. *The difference between the estimated sparse graphs of the cases and controls subjects using* SMG Kendall, SMG Pearson *and* Naive Kendall. *Here, the black color represents the edges only present in the graph for cases but not in controls persons, while the red represents the opposite.*

evidence for the population-level network's ability to capture differences in different subpopulations. Furthermore, we observe that this disparity between the real subpopulation cases and the "null" case is most pronounced with SMG Kendall and least pronounced with Naive Kendall—demonstrating the sparse median graph's potential advantage in predictive tasks.

TABLE 1

Predictive power. Predictive power of SMG Kendall *and competing methods. We measure predictive power by the Hamming distance between patients of two classes divided by* $\binom{d-1}{2}$. *We use* $\lambda = 0.171$ *for the CLIME parameter. This table represents values at* 10^{-3} *scale*

Data	**SMG Ken.**	**SMG Pea.**	**Naive Ken.**
	Randomized data difference		
Null	3.74	4.46	3.95
	ADHD case and control difference		
Whole	5.18	5.10	4.35
Male	10.57	10.05	4.90
Female	6.60	5.67	4.90
Junior	6.22	6.31	4.35
Senior	9.02	7.89	5.10
	Male and female difference		
ADHD case	9.07	8.64	5.67
ADHD control	7.81	6.63	4.35
	Junior and senior difference		
ADHD case	9.45	8.93	5.67
ADHD control	9.36	9.25	6.19

In addition, within the contrasting classes, we observe SMG Kendall estimates the greatest difference between any two classes. Furthermore, while SMG Pearson performs very closely to SMG Kendall in most cases, there is a larger difference between the two methods in the tests that separate or compare the subjects by gender. This suggests SMG Kendall is more sensitive to the differences between male and female brains than SMG Pearson. Moreover, both SMG-based methods show higher predictive powers than Naive Kendall. This provides further evidence of the predictive advantage of non-i.i.d. population models.

6.3.3. *Stability*: *CLIME parameter perturbations.* In this experiment, we compare the stability of the proposed method to those of the competing methods under parameter perturbations. In particular, we examine the stability by measuring the scaled Hamming distance between a sparse graph estimated with the CLIME parameter $\lambda = 0.171$ and the sparse graph estimated using a perturbed CLIME parameter.

To this end, we conduct the experiment as follows for each of the three methods:

1. Using the CLIME parameter $\lambda = 0.171$, we estimate a population-level sparse graph $\widehat{\mathcal{G}}^*_{s_\lambda,\lambda}$. Here, we select s_λ by setting s_λ to be the median number of edges of the graphs estimated for each individual subject. More specifically, recall the algorithm first applies Kendall or Pearson (see Section 6.1.1) to each individual

TABLE 2

Stability with respect to the clime parameter. Stability of SMG Kendall *and competing methods with respect to perturbations to the CLIME parameter. Stability is measured as Hamming distance, divided by s_λ, between the graph estimated with the perturbed parameter and the graph estimated with the unperturbed parameter, λ. Here, s_λ is the number edges in the graph estimated with the unperturbed parameter. We use $\lambda = 0.171$ for the CLIME parameter. This table represents values at 10^{-1} scale*

	Variation					
	Mean	**Std. Dev.**	**Mean**	**Std. Dev.**	**Mean**	**Std. Dev.**
	0.9λ		0.95λ		0.99λ	
SMG Kendall	3.46	0.137	2.14	0.109	1.68	0.148
SMG Pearson	3.23	0.240	2.05	0.142	1.51	0.209
Naive Kendall	3.83	0.381	2.46	0.279	1.57	0.145
	1.01λ		1.05λ		1.1λ	
SMG Kendall	1.51	0.128	1.77	0.266	2.34	0.066
SMG Pearson	1.48	0.143	1.80	0.096	2.41	0.097
Naive Kendall	1.49	0.231	2.42	0.182	3.17	0.174

graph. Each one of these individual graphs possesses some number of edges. We choose s_λ to be the median among that set of numbers.

2. We repeat the procedure but estimate the graph of each individual subject with a perturbed CLIME parameter. In particular, we use $p \times \lambda$ for the values of $p = 0.9$, 0.95, 0.99, 1.01, 1.05 and 1.1.

3. We examine the Hamming distance between each $\widehat{\mathcal{G}}^*_{s_{(p\times\lambda)},(p\times\lambda)}$ and $\widehat{\mathcal{G}}^*_{s_\lambda,\lambda}$ divided by s_λ. The results are in Table 2.

Table 2 shows our proposed method is comparable in stability to SMG Pearson. In addition, Naive Kendall tends to display significantly higher instability than the SMG-based methods for 1.05λ and 1.1λ. Since CLIME outputs more sparse graphs for larger λ parameters, this supports the claim that the sparse median graph provides a more stable estimator of graphs in sparse settings than models assuming the population data arise from i.i.d. settings.

6.3.4. *Stability*: *Data perturbations*. In this experiment, we consider the stability of the proposed method when the dataset is perturbed. In particular, we take K subsamples of the $T = 739$ subjects in the ADHD-200 data, $\mathcal{D} = \{\mathcal{D}_t\}_{t=1}^T$, to create[8] $\{\widehat{\mathcal{D}}^k\}_{k=1}^K$. We estimate a sparse population graph on each $\widehat{\mathcal{D}}^k$ and measure the instability by examining the differences in the resulting graphs.

More specifically, we apply the following procedure using each of the three methods:

[8] Here, $\mathcal{D}$ is the entire ADHD-200 dataset, $\mathcal{D}_t$ is the data corresponding to the tth subject in $\mathcal{D}$, and $\widehat{\mathcal{D}}^k$ is the kth dataset found by taking a subsample of the subjects from $\mathcal{D}$.

TABLE 3

Stability with respect to data. Stability of SMG Kendall *and competing methods with respect to perturbations to the data via subsampling. Here, we measure the total instability as the mean of the disagreements on the presence of each edge* [*Liu, Roeder and Wasserman* (2010)]. *Here,* $n = 100$ *samples were taken. We use* $\lambda = 0.171$ *for the CLIME parameter. This table represents values at* 10^{-3} *scale*

	Sampling ratio					
	Instability	**Std. Err.**	**Instability**	**Std. Err.**	**Instability**	**Std. Err.**
	0.65		0.8		0.9	
SMG Kendall	2.39	0.152	1.55	0.122	0.941	0.096
SMG Pearson	2.30	0.152	1.54	0.126	1.02	0.102
Naive Kendall	3.00	0.171	2.70	0.162	2.46	0.157

1. We randomly draw $K = 100$ subsamples of $T^k = \lfloor p \times T \rfloor$ patients at subsampling ratios of $p = 0.65$, 0.8 and 0.9:

$$I^k = \{t_i\}_{i=1}^{T^k} \subset \{1, 2, \ldots, T\}.$$

In other words, each subsampled dataset, $\widehat{\mathcal{D}}^k = \{\mathcal{D}_t\}_{t \in I_k}$, contains the data corresponding to T^k subjects from the entire ADHD dataset.

2. Using the CLIME parameter $\lambda = 0.171$, we estimate a sparse population graph $\widehat{\mathcal{G}}_s^{*,k}$ from $\widehat{\mathcal{D}}^k$, where s is the stability parameter chosen using the same method as in Section 6.3.3.

3. We measure the instability by averaging the disagreements on the presence of edges in $\{\widehat{\mathcal{G}}_s^{*,k}\}_{k=1}^{K}$. We refer to Section 3.2 of Liu, Roeder and Wasserman (2010) for a detailed description of this measure. The results are in Table 3, where larger values correspond to more instability.

Table 3 shows our proposed method is comparable in stability to SMG Pearson under data perturbations. In addition, observe Naive Kendall displays significantly more instability than either of the methods that employ the sparse median graph approach. This demonstrates the resistance of the sparse median graph approach to the characteristics of individual subjects when estimating a population-level graph.

7. Discussion. In this paper, we discuss the concept of the sparse median graph that estimates a population-level graph under nonparanormal assumptions. This new approach combines two new developments in graph estimation literature—namely, (i) employing sparsity constraints in high-dimensional settings for identifiability [for example Banerjee, El Ghaoui and d'Aspremont (2008), Friedman, Hastie and Tibshirani (2007)] and (ii) increasing graph recovery rates by using nonparanormal assumptions [Liu, Lafferty and Wasserman (2009), Liu et al. (2012)]—with the idea of median graphs from pattern recognition literature [Bunke and Shearer (1998), Jiang, Munger and Bunke (2001)]. The resulting method, which we analyze both theoretically and empirically, allows us to estimate a graph that emphasizes the commonalities and downplays the individual outliers within a population.

In particular, we theoretically prove the consistency of this method and bound its rate of convergence. Then in two simulations—one with the synthetic and one with the ADHD-200 brain imaging data—we demonstrate our proposed method displays higher estimation performance than potential competing methods. We observe the benefits of the nonparanormal model with Kendall's tau tend to dominate in the synthetic data simulations, while the benefits of the sparse median graph aggregation method tend to dominate in the simulations with the real data. One possible explanation is that the "biggest challenge" in estimating the graphs from the synthetic data is the data's non-Gaussianity (which is addressed by utilizing Kendall's tau), while the biggest challenge in estimating the graphs from the real brain imaging data stems from the individual outlier and heterogeneous characteristics of patients and scans (which are downplayed by the sparse median graph). However, the consistent optimal performance of the proposed method in both simulations demonstrates its value as an estimator of choice for both highly non-Gaussian data as well as complex aggregated datasets with large variation in individual characteristics.

In addition, we perform experiments using the ADHD-200 brain imaging dataset. The experiments demonstrate the proposed method possesses the highest predictive power for classification tasks among its competitors. Furthermore, stability experiments on the same dataset show the sparse median graph summarization provides much more stable estimators than the Naive Kendall method that assumes the homogeneity of the entire dataset.

These results offer compelling evidence that the proposed method possesses the potential to become a unified framework for conducting inference on complex datasets of aggregated data. While the current analysis is primarily illustrative, we have demonstrated its value for applications in brain imaging fields where interest lies primarily in population characteristics. Therefore, we believe this investigation would justify a more thorough inferential investigation of median graph properties and network modification with disease in future works.

REFERENCES

BANERJEE, O., EL GHAOUI, L. and D'ASPREMONT, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *J. Mach. Learn. Res.* **9** 485–516. MR2417243

BÜHLMANN, P. and VAN DE GEER, S. (2011). *Statistics for High-Dimensional Data*. Springer, Heidelberg. MR2807761

BULLMORE, E. and SPORNS, O. (2009). Complex brain networks: Graph theoretical analysis of structural and functional systems. *Nat. Rev. Neurosci.* **10** 186–198.

BUNKE, H. and SHEARER, K. (1998). A graph distance metric based on the maximal common subgraph. *Pattern Recognition Letters* **19** 255–259.

CAI, T., LIU, W. and LUO, X. (2011). A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *J. Amer. Statist. Assoc.* **106** 594–607. MR2847973

DEMPSTER, A. P. (1972). Covariance selection. *Biometrics* **28** 157–175.

ELOYAN, A., MUSCHELLI, J., NEBEL, M. B., LIU, H., HAN, F., ZHAO, T., BARBER, A., JOEL, S., PEKAR, J. J., MOSTOFSKY, S. and CAFFO, B. (2012). Automated diagnoses of attention deficit hyperactive disorder using magnetic resonance imaging. *Frontiers in Systems Neuroscience* **6** 61.

FINGELKURTS, A. A. and KÄHKÖNEN, S. (2005). Functional connectivity in the brain—Is it an elusive concept? *Neuroscience and Biobehavioral Reviews* **28** 827–836.

FRIEDMAN, J. H., HASTIE, T. and TIBSHIRANI, R. (2007). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9** 432–441.

FRISTON, K. J. (2011). Functional and effective connectivity: A review. *Brain Connect.* **1** 13–36.

HAN, F. and LIU, H. (2014). Distribution-free tests of independence with applications to testing more structures. Preprint. Available at arXiv:1410.4179.

HORWITZ, B. (2003). The elusive concept of brain connectivity. *Neuroimage* **19** 466–470.

HSIEH, C. J., SUSTIK, M. A., RAVIKUMAR, P. and DHILLON, I. S. (2011). Sparse inverse covariance matrix estimation using quadratic approximation. In *Advances in Neural Information Processing Systems* (*NIPS*) **24**. Granada, Spain.

JIANG, X., MUNGER, A. and BUNKE, H. (2001). On median graphs: Properties, algorithms, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **23** 1144–1151.

JIN, J., ZHANG, C.-H. and ZHANG, Q. (2014). Optimality of graphlet screening in high dimensional variable selection. *J. Mach. Learn. Res.* **15** 2723–2772. MR3270749

KE, T., JIN, J. and FAN, J. (2014). Covariance assisted screening and estimation. *Ann. Statist.* **42** 2202.

LAM, C. and FAN, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *Ann. Statist.* **37** 4254–4278. MR2572459

LEDOIT, O. and WOLF, M. (2003). Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of Empirical Finance* **10** 603–621.

LI, L. and TOH, K.-C. (2010). An inexact interior point method for L_1-regularized sparse covariance selection. *Math. Program. Comput.* **2** 291–315. MR2741488

LIU, I. and AGRESTI, A. (1996). Mantel–Haenszel-type infererence for cumulative odds ratios with a stratified ordinal response. *Biometrics* **52** 1223–1234. MR1422076

LIU, H., LAFFERTY, J. and WASSERMAN, L. (2009). The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *J. Mach. Learn. Res.* **10** 2295–2328. MR2563983

LIU, W. and LUO, X. (2012). High-dimensional sparse precision matrix estimation via sparse column inverse operator. Preprint. Available at arXiv:1203.3896.

LIU, H., ROEDER, K. and WASSERMAN, L. (2010). Stability approach to regularization selection (StARS) for high dimensional graphical models. In *Advances in Neural Information Processing Systems* 1432–1440. Vancouver, Canada.

LIU, H., HAN, F., YUAN, M., LAFFERTY, J. and WASSERMAN, L. (2012). High-dimensional semiparametric Gaussian copula graphical models. *Ann. Statist.* **40** 2293–2326. MR3059084

MEINSHAUSEN, N. and BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34** 1436–1462. MR2278363

MILHAM, M. P., FAIR, D., MENNES, M. and MOSTOFSKY, S. H. (2012). The ADHD-200 consortium: A model to advance the translational potential of neuroimaging in clinical neuroscience. *Frontiers in Systems Neuroscience* **6** 62.

PENG, J., WANG, P., ZHOU, N. and ZHU, J. (2009). Partial correlation estimation by joint sparse regression models. *J. Amer. Statist. Assoc.* **104** 735–746. MR2541591

POWER, J. D., COHEN, A. L., NELSON, S. M., WIG, G. S., BARNES, K. A., CHURCH, J. A., VOGEL, A. C., LAUMANN, T. O., MIEZIN, F. M., SCHLAGGAR, B. L. and PETERSON, S. (2011). Functional network organization of the human brain. *Neuron* **72** 665–678.

RAMSAY, J. D., HANSON, S. J., HANSON, C., HALCHENKO, Y., POLDRACK, R. and GLYMOUR, C. (2009). Six problems for causal inference from fMRI. *NeuroImage* **49** 1545–1558.

RAVIKUMAR, P., WAINWRIGHT, M., RASKUTTI, G. and YU, B. (2009). Model selection in Gaussian graphical models: High-dimensional consistency of ℓ_1-regularized MLE. In *Advances in Neural Information Processing Systems* **22**. Vancouver, Canada.

ROTHMAN, A. J., BICKEL, P. J., LEVINA, E. and ZHU, J. (2008). Sparse permutation invariant covariance estimation. *Electron. J. Stat.* **2** 494–515. MR2417391

ROWE, B. L. Y. (2014). tawny: Provides various portfolio optimization strategies including random matrix theory and shrinkage estimators. R package version 2.1.2.

RUBINOV, M. and SPORNS, O. (2010). Complex network measures of brain connectivity: Uses and interpretations. *Neuroimage* **52** 1059–1069.

SCHEINBERG, K., MA, S. and GLODFARB, D. (2010). Sparse inverse covariance selection via alternating linearization methods. In *Advances in Neural Information Processing Systems* (*NIPS*) **23**. Vancouver, Canada.

XU, W., HOU, Y., HUNG, Y. S. and ZOU, Y. (2010). Comparison of Spearman's rho and Kendall's tau in normal and contaminated normal models. Preprint. Available at arXiv:1011.2009.

XUE, L. and ZOU, H. (2012). Regularized rank-based estimation of high-dimensional nonparanormal graphical models. *Ann. Statist.* **40** 2541–2571. MR3097612

YUAN, M. (2010). High dimensional inverse covariance matrix estimation via linear programming. *J. Mach. Learn. Res.* **11** 2261–2286. MR2719856

ZHAO, T., LIU, H., ROEDER, K., LAFFERTY, J. and WASSERMAN, L. (2012). The huge package for high-dimensional undirected graph estimation in R. *J. Mach. Learn. Res.* **13** 1059–1062. MR2930633

F. HAN
DEPARTMENT OF STATISTICS
UNIVERSITY OF WASHINGTON
SEATTLE, WASHINGTON 98195
USA
E-MAIL: fanghan@uw.edu

X. HAN
H. LIU
DEPARTMENT OF OPERATIONS RESEARCH
AND FINANCIAL ENGINEERING
PRINCETON UNIVERSITY
PRINCETON, NEW JERSEY 08544
USA
E-MAIL: xiaoyanh@princeton.edu
hanliu@princeton.edu

B. CAFFO
DEPARTMENT OF BIOSTATISTICS
JOHNS HOPKINS UNIVERSITY
BALTIMORE, MARYLAND 21205
USA
E-MAIL: bcaffo@jhsph.edu