Journal of Machine Learning Research 15 (2014) 1713-1750 Submitted 5/13; Revised 3/14; Published 5/14

Graph Estimation From Multi-Attribute Data

Mladen Kolar MKOLAR@CHICAGOBOOTH.EDU
The University of Chicago Booth School of Business
Chicago, Illinois 60637, USA

Han Liu HANLIU@PRINCETON.EDU
Department of Operations Research and Financial Engineering
Princeton University
Princeton, New Jersey 08544, USA

Eric P. Xing EPXING@CS.CMU.EDU
Machine Learning Department
Carnegie Mellon University
Pittsburgh, Pennsylvania 15213, USA

Editor: Yuan (Alan) Qi

Abstract

Undirected graphical models are important in a number of modern applications that involve exploring or exploiting dependency structures underlying the data. For example, they are often used to explore complex systems where connections between entities are not well understood, such as in functional brain networks or genetic networks. Existing methods for estimating structure of undirected graphical models focus on scenarios where each node represents a scalar random variable, such as a binary neural activation state or a continuous mRNA abundance measurement, even though in many real world problems, nodes can represent multivariate variables with much richer meanings, such as whole images, text documents, or multi-view feature vectors. In this paper, we propose a new principled framework for estimating the structure of undirected graphical models from such multivariate (or multi-attribute) nodal data. The structure of a graph is inferred through estimation of non-zero partial canonical correlation between nodes. Under a Gaussian model, this strategy is equivalent to estimating conditional independencies between random vectors represented by the nodes and it generalizes the classical problem of covariance selection (Dempster, 1972). We relate the problem of estimating non-zero partial canonical correlations to maximizing a penalized Gaussian likelihood objective and develop a method that efficiently maximizes this objective. Extensive simulation studies demonstrate the effectiveness of the method under various conditions. We provide illustrative applications to uncovering gene regulatory networks from gene and protein profiles, and uncovering brain connectivity graph from positron emission tomography data. Finally, we provide sufficient conditions under which the true graphical structure can be recovered correctly.

Keywords: graphical model selection, multi-attribute data, network analysis, partial canonical correlation

1. Introduction

Gaussian graphical models are commonly used to represent and explore conditional independencies between variables in a complex system. An edge between two nodes is present

in the graph if and only if the corresponding variables are conditionally independent given all the other variables. Current approaches to estimating the Markov network structure underlying a Gaussian graphical model focus on cases where nodes in a network represent scalar variables such as the binary voting actions of actors (Banerjee et al., 2008; Kolar et al., 2010) or the continuous mRNA abundance measurements of genes (Peng et al., 2009). However, in many modern problems, we are interested in studying a network where nodes can represent more complex and informative vector-variables or multi-attribute entities. For example, due to advances of modern data acquisition technologies, researchers are able to measure the activities of a single gene in a high-dimensional space, such as an image of the spatial distribution of the gene expression, or a multi-view snapshot of the gene activity such as mRNA and protein abundances; when modeling a social network, a node may correspond to a person for which a vector of attributes is available, such as personal information, demographics, interests, and other features. Therefore, there is a need for methods that estimate the structure of an undirected graphical model from such multi-attribute data.

In this paper, we develop a new method for estimating the structure of undirected graphical models of which the nodes correspond to vectors, that is, multi-attribute entities. We consider the following setting. Let $X = (X_1^T, ..., X_p^T)^T$ where $X_1 \in \mathbb{R}^{k_1}, \ldots, X_p \in \mathbb{R}^{k_p}$ are random vectors that jointly follow a multivariate Gaussian distribution with mean $\mu = (\mu_1^T, \ldots, \mu_p^T)^T$ and covariance matrix Σ^*, which is partitioned as

$$\Sigma^* = \begin{pmatrix} \Sigma^*_{11} & \cdots & \Sigma^*_{1p} \\ \vdots & \ddots & \vdots \\ \Sigma^*_{p1} & \cdots & \Sigma^*_{pp} \end{pmatrix}, \tag{1}$$

with $\Sigma^*_{ij} = \mathrm{Cov}(X_i, X_j)$. Without loss of generality, we assume $\mu = 0$. Let $G = (V, E)$ be a graph with the vertex set $V = \{1, \ldots, p\}$ and the set of edges $E \subseteq V \times V$ that encodes the conditional independence relationships among $(X_a)_{a \in V}$. That is, each node $a \in V$ of the graph G corresponds to the random vector X_a and there is no edge between nodes a and b in the graph if and only if X_a is conditionally independent of X_b given all the vectors corresponding to the remaining nodes, $X_{\neg ab} = \{X_c \ : \ c \in V \backslash \{a, b\}\}$. Such a graph is known as a *Markov network* (of Markov graph), which we shall emphasize in this paper to contrast an alternative graph over V known as the *association network*, which is based on pairwise marginal independence. Conditional independence can be read from the inverse of the covariance matrix of X, as the block corresponding to X_a and X_b will be equal to zero when they are conditionally independent, whereas marginal independencies are captured by the covariance matrix itself. It is well known that estimating an association network from data can result in a hard-to-interpret dense graph due to prevalent indirect correlations among variables (for example, multiple nodes each influenced by a common single node could result in a clique over all these nodes), which can be avoided in estimating a Markov network.

Let $\mathcal{D}_n = \{x_i\}_{i=1}^n$ be a sample of n independent and identically distributed (*iid*) vectors drawn from $\mathrm{N}(0, \Sigma^*)$. For a vector x_i, we denote $x_{i,a} \in \mathbb{R}^{k_a}$ the component corresponding to the node $a \in V$. Our goal is to estimate the structure of the graph G from the sample $\mathcal{D}_n$. Note that we allow for different nodes to have different number of attributes, which is useful in many applications, for example, when a node represents a gene pathway (of

different sizes) in a regulatory network, or a brain region (of different volumes) in a neural activation network.

Learning the structure of a Gaussian graphical model, where each node represents a scalar random variable, is a classical problem, known as the covariance selection (Dempster, 1972). One can estimate the graph structure by estimating the sparsity pattern of the precision matrix $\Omega = \Sigma^{-1}$. For high dimensional problems, Meinshausen and Bühlmann (2006) propose a parallel Lasso approach for estimating Gaussian graphical models by solving a collection of sparse regression problems. This procedure can be viewed as a pseudo-likelihood based method. In contrast, Banerjee et al. (2008), Yuan and Lin (2007), and Friedman et al. (2008) take a penalized likelihood approach to estimate the sparse precision matrix Ω. To reduce estimation bias, Lam and Fan (2009), Johnson et al. (2012), and Shen et al. (2012) developed the non-concave penalties to penalize the likelihood function. More recently, Yuan (2010) and Cai et al. (2011) proposed the graphical Dantzig selector and CLIME, which can be solved by linear programming and are more amenable to theoretical analysis than the penalized likelihood approach. Under certain regularity conditions, these methods have proven to be graph-estimation consistent (Ravikumar et al., 2011; Yuan, 2010; Cai et al., 2011) and scalable software packages, such as *glasso* and *huge*, were developed to implement these algorithms (Zhao et al., 2012). For a recent survey see Pourahmadi (2011). However, these methods cannot be extended to handle multi-attribute data we consider here in an obvious way. For example, if the number of attributes is the same for each node, one may naively estimate one graph per attribute, however, it is not clear how to combine such graphs into a summary graph with a clear statistical interpretation. The situation becomes even more difficult when nodes correspond to objects that have different number of attributes.

In a related work, Katenka and Kolaczyk (2011) use canonical correlation to estimate association networks from multi-attribute data, however, such networks have different interpretation to undirected graphical models. In particular, as mentioned above, association networks are known to confound the direct interactions with indirect ones as they only represent marginal associations, whereas Markov networks represent conditional independence assumptions that are better suited for separating direct interactions from indirect confounders. Our work is related to the literature on simultaneous estimation of multiple Gaussian graphical models under a multi-task setting (Guo et al., 2011; Varoquaux et al., 2010; Honorio and Samaras, 2010; Chiquet et al., 2011; Danaher et al., 2014). However, the model given in (1) is different from models considered in various multi-task settings and the optimization algorithms developed in the multi-task literature do not extend to handle the optimization problem given in our setting.

Unlike the standard procedures for estimating the structure of Gaussian graphical models, for example, neighborhood selection (Meinshausen and Bühlmann, 2006) or glasso (Friedman et al., 2008), which infer the partial correlations between pairs of nodes, our proposed method estimates the graph structure based on the partial canonical correlation, which can naturally incorporate complex nodal observations. Under that the Gaussian model in (1), the estimated graph structure has the same probabilistic independence interpretations as the Gaussian graphical model over univariate nodes. The main contributions of the paper are the following. First, we introduce a new framework for learning structure of undirected graphical models from multi-attribute data. Second, we develop an efficient

algorithm that estimates the structure of a graph from the observed data. Third, we provide extensive simulation studies that demonstrate the effectiveness of our method and illustrate how the framework can be used to uncover gene regulatory networks from gene and protein profiles, and to uncover brain connectivity graph from functional magnetic resonance imaging data. Finally, we provide theoretical results, which give sufficient conditions for consistent structure recovery.

2. Methodology

In this section, we propose to estimate the graph by estimating non-zero partial canonical correlation between the nodes. This leads to a penalized maximum likelihood objective, for which we develop an efficient optimization procedure.

2.1 Preliminaries

Let X_a and X_b be two multivariate random vectors. Canonical correlation is defined between X_a and X_b as

$$\rho_c(X_a, X_b) = \max_{u \in \mathbb{R}^{k_a}, v \in \mathbb{R}^{k_b}} \text{corr}(u^T X_a, v^T X_b).$$

That is, computing canonical correlation between X_a and X_b is equivalent to maximizing the correlation between two linear combinations $u^T X_a$ and $v^T X_b$ with respect to vectors u and v. Canonical correlation can be used to measure association strength between two nodes with multi-attribute observations. For example, in Katenka and Kolaczyk (2011), a graph is estimated from multi-attribute nodal observations by elementwise thresholding the canonical correlation matrix between nodes, but such a graph estimator may confound the direct interactions with indirect ones.

In this paper, we exploit the partial canonical correlation to estimate a graph from multi-attribute nodal observations. A graph is going to be formed by connecting nodes with non-zero partial canonical correlation. Let $\widehat{A} = \text{argmin } E\left(||X_a - AX_{\neg ab}||_2^2\right)$ and $\widehat{B} = \text{argmin } E\left(||X_b - BX_{\neg ab}||_2^2\right)$, then the partial canonical correlation between X_a and X_b is defined as

$$\rho_c(X_a, X_b; X_{\neg ab}) = \max_{u \in \mathbb{R}^{k_a}, v \in \mathbb{R}^{k_b}} \text{corr}\{u^T(X_a - \widehat{A}X_{\neg ab}), v^T(X_b - \widehat{B}X_{\neg ab})\}, \tag{2}$$

that is, the partial canonical correlation between X_a and X_b is equal to the canonical correlation between the residual vectors of X_a and X_b after the effect of $X_{\neg ab}$ is removed (Rao, 1969).[1]

Let $\Omega^* = (\Sigma^*)^{-1}$ denote the precision matrix. A simple calculation, given in Appendix B.3, shows that

$$\rho_c(X_a, X_b; X_{\neg ab}) \neq 0 \quad \text{if and only if} \quad \max_{u \in \mathbb{R}^{k_a}, v \in \mathbb{R}^{k_b}} u^T \Omega^*_{ab} v \neq 0. \tag{3}$$

This implies that estimating whether the partial canonical correlation is zero or not can be done by estimating whether a block of the precision matrix is zero or not. Furthermore,

1. The operator $E(\cdot)$ denotes the expectation and $X_{\neg ab} = \{X_c : c \in V \backslash \{a, b\}\}$ denotes all the variables except for X_a and X_b.

under the Gaussian model in (1), vectors X_a and X_b are conditionally independent given $X_{\neg ab}$ if and only if partial canonical correlation is zero. A graph built on this type of inter-nodal relationship is known as a Markov network, as it captures both local and global Markov properties over all arbitrary subsets of nodes in the graph, even though the graph is built based on pairwise conditional independence properties. In Section 2.2, we use the above observations to design an algorithm that estimates the non-zero partial canonical correlation between nodes from data $\mathcal{D}_n$ using the penalized maximum likelihood estimation of the precision matrix.

Based on the relationship given in (3), we can motivate an alternative method for estimating the non-zero partial canonical correlation. Let $\overline{a} = \{b \ : \ b \in V \backslash \{a\}\}$ denote the set of all nodes minus the node a. Then

$$E\left(X_a \mid X_{\overline{a}} = x_{\overline{a}}\right) = \Sigma^*_{a,\overline{a}} \Sigma^{*,-1}_{\overline{a},\overline{a}} x_{\overline{a}}.$$

Since $\Omega^*_{a,\overline{a}} = -(\Sigma^*_{aa} - \Sigma^*_{a,\overline{a}} \Sigma^{*,-1}_{\overline{a},\overline{a}} \Sigma^*_{\overline{a},a})^{-1} \Sigma^*_{a,\overline{a}} \Sigma^{*,-1}_{\overline{a},\overline{a}}$, we observe that a zero block Ω_{ab} can be identified from the regression coefficients when each component of X_a is regressed on $X_{\overline{a}}$. We do not build an estimation procedure around this observation, however, we note that this relationship shows how one would develop a regression based analogue of the work presented in Katenka and Kolaczyk (2011).

2.2 Penalized Log-Likelihood Optimization

Based on the data $\mathcal{D}_n$, we propose to minimize the penalized negative Gaussian log-likelihood under the model in (1),

$$\min_{\Omega > 0} \Big\{ \operatorname{tr} S\Omega - \log |\Omega| + \lambda \sum_{a,b} ||\Omega_{ab}||_F \Big\}, \tag{4}$$

where $S = n^{-1} \sum_{i=1}^n x_i x_i^T$ is the sample covariance matrix, $||\Omega_{ab}||_F$ denotes the Frobenius norm of Ω_{ab} and λ is a user defined parameter that controls the sparsity of the solution $\widehat{\Omega}$. The first two terms in (4) correspond to the negative Gaussian log-likelihood, while the second term is the Frobenius norm penalty, which encourages blocks of the precision matrix to be equal to zero, similar to the way that the ℓ_2 penalty is used in the group Lasso (Yuan and Lin, 2006). Here we assume that the same number of samples is available per attribute. However, the same method can be used in cases when some samples are obtained on a subset of attributes. Indeed, we can simply estimate each element of the matrix S from available samples, treating non-measured attributes as missing completely at random (for more details see Kolar and Xing, 2012).

The dual problem to (4) is

$$\max_{\Sigma} \sum_{j \in V} k_j + \log |\Sigma| \qquad \text{subject to} \qquad \max_{a,b} \ ||S_{ab} - \Sigma_{ab}||_F \leqslant \lambda, \tag{5}$$

where k_j is the number attributes of node j, Σ is the dual variable to Ω and $|\Sigma|$ denotes the determinant of Σ. Note that the primal problem gives us an estimate of the precision matrix, while the dual problem estimates the covariance matrix. The proposed optimization procedure, described below, will simultaneously estimate the precision matrix and covariance matrix, without explicitly performing an expensive matrix inversion.

We propose to optimize the objective function in (4) using an inexact block coordinate descent procedure, inspired by Mazumder and Agarwal (2011). The block coordinate descent is an iterative procedure that operates on a block of rows and columns while keeping the other rows and columns fixed. We write

$$\Omega = \begin{pmatrix} \Omega_{aa} & \Omega_{a,\overline{a}} \\ \Omega_{\overline{a},a} & \Omega_{\overline{a},\overline{a}} \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{aa} & \Sigma_{a,\overline{a}} \\ \Sigma_{\overline{a},a} & \Sigma_{\overline{a},\overline{a}} \end{pmatrix}, \quad S = \begin{pmatrix} S_{aa} & S_{a,\overline{a}} \\ S_{\overline{a},a} & S_{\overline{a},\overline{a}} \end{pmatrix},$$

and suppose that $(\widetilde{\Omega}, \widetilde{\Sigma})$ are the current estimates of the precision matrix and covariance matrix. With the above block partition, we have $\log|\Omega| = \log(\Omega_{\overline{a},\overline{a}}) + \log(\Omega_{aa} - \Omega_{a,\overline{a}}(\Omega_{\overline{a},\overline{a}})^{-1}\Omega_{\overline{a},a})$. In the next iteration, $\widehat{\Omega}$ is of the form

$$\widehat{\Omega} = \widetilde{\Omega} + \begin{pmatrix} \Delta_{aa} & \Delta_{a,\overline{a}} \\ \Delta_{\overline{a},a} & 0 \end{pmatrix} = \begin{pmatrix} \widehat{\Omega}_{aa} & \widehat{\Omega}_{a,\overline{a}} \\ \widehat{\Omega}_{\overline{a},a} & \widetilde{\Omega}_{\overline{a},\overline{a}} \end{pmatrix},$$

and is obtained by minimizing

$$\operatorname{tr} S_{aa}\Omega_{aa} + 2\operatorname{tr} S_{a,\overline{a}}\Omega_{\overline{a},a} - \log|\Omega_{aa} - \Omega_{a,\overline{a}}(\widetilde{\Omega}_{\overline{a},\overline{a}})^{-1}\Omega_{\overline{a},a}| + \lambda||\Omega_{aa}||_F + 2\lambda\sum_{b\neq a}||\Omega_{ab}||_F. \quad (6)$$

Exact minimization over the variables Ω_{aa} and $\Omega_{a,\overline{a}}$ at each iteration of the block coordinate descent procedure can be computationally expensive. Therefore, we propose to update Ω_{aa} and $\Omega_{a,\overline{a}}$ using one generalized gradient step update (see Beck and Teboulle, 2009) in each iteration. Note that the objective function in (6) is a sum of a smooth convex function and a non-smooth convex penalty so that the gradient descent method cannot be directly applied. Given a step size t, generalized gradient descent optimizes a quadratic approximation of the objective at the current iterate $\widetilde{\Omega}$, which results in the following two updates

$$\widehat{\Omega}_{aa} = \underset{\Omega_{aa}}{\operatorname{argmin}} \Big\{ \operatorname{tr}(S_{aa} - \widetilde{\Sigma}_{aa})\Omega_{aa} + \frac{1}{2t}||\Omega_{aa} - \widetilde{\Omega}_{aa}||_F^2 + \lambda||\Omega_{aa}||_F \Big\}, \quad \text{and} \quad (7)$$

$$\widehat{\Omega}_{ab} = \underset{\Omega_{ab}}{\operatorname{argmin}} \Big\{ \operatorname{tr}(S_{ab} - \widetilde{\Sigma}_{ab})\Omega_{ba} + \frac{1}{2t}||\Omega_{ab} - \widetilde{\Omega}_{ab}||_F^2 + \lambda||\Omega_{ab}||_F \Big\}, \quad \forall b \in \overline{a}. \quad (8)$$

If the resulting estimator $\widehat{\Omega}$ is not positive definite or the update does not decrease the objective, we halve the step size t and find a new update. Once the update of the precision matrix $\widehat{\Omega}$ is obtained, we update the covariance matrix $\widehat{\Sigma}$. Updates to the precision and covariance matrices can be found efficiently, without performing expensive matrix inversion. First, note that the solutions to (7) and (8) can be computed in a closed form as

$$\widehat{\Omega}_{aa} = (1 - t\lambda/||\widetilde{\Omega}_{aa} + t(\widetilde{\Sigma}_{aa} - S_{aa})||_F)_+(\widetilde{\Omega}_{aa} + t(\widetilde{\Sigma}_{aa} - S_{aa})), \quad \text{and} \quad (9)$$

$$\widehat{\Omega}_{ab} = (1 - t\lambda/||\widetilde{\Omega}_{ab} + t(\widetilde{\Sigma}_{ab} - S_{ab})||_F)_+(\widetilde{\Omega}_{ab} + t(\widetilde{\Sigma}_{ab} - S_{ab})), \quad \forall b \in \overline{a}, \quad (10)$$

where $(x)_+ = \max(0, x)$. Next, the estimate of the covariance matrix can be updated efficiently, without inverting the whole $\widehat{\Omega}$ matrix, using the matrix inversion lemma as follows

$$\begin{aligned} \widehat{\Sigma}_{\overline{a},\overline{a}} &= (\widetilde{\Omega}_{\overline{a},\overline{a}})^{-1} + (\widetilde{\Omega}_{\overline{a},\overline{a}})^{-1}\widehat{\Omega}_{\overline{a},a}(\widehat{\Omega}_{aa} - \widehat{\Omega}_{a,\overline{a}}(\widetilde{\Omega}_{\overline{a},\overline{a}})^{-1}\widehat{\Omega}_{\overline{a},a})^{-1}\widehat{\Omega}_{a,\overline{a}}(\widetilde{\Omega}_{\overline{a},\overline{a}})^{-1}, \\ \widehat{\Sigma}_{a,\overline{a}} &= -\widehat{\Omega}_{aa}\widehat{\Omega}_{a,\overline{a}}\widehat{\Sigma}_{\overline{a},\overline{a}}, \\ \widehat{\Sigma}_{aa} &= (\widehat{\Omega}_{aa} - \widehat{\Omega}_{a,\overline{a}}(\widetilde{\Omega}_{\overline{a},\overline{a}})^{-1}\widehat{\Omega}_{\overline{a},a})^{-1}, \end{aligned} \quad (11)$$

with $(\widetilde{\Omega}_{\overline{a},\overline{a}})^{-1} = \widetilde{\Sigma}_{\overline{a},\overline{a}} - \widetilde{\Sigma}_{\overline{a},a}\widetilde{\Sigma}_{aa}^{-1}\widetilde{\Sigma}_{a,\overline{a}}$.

Combining all three steps we get the following algorithm:

1. Set the initial estimator $\widetilde{\Omega} = \mathrm{diag}(S)$ and $\widetilde{\Sigma} = \widetilde{\Omega}^{-1}$. Set the step size $t = 1$.

2. For each $a \in V$ perform the following:

   Update $\widehat{\Omega}$ using (9) and (10).
   If $\widehat{\Omega}$ is not positive definite, set $t \leftarrow t/2$ and repeat the update.
   Update $\widehat{\Sigma}$ using (11).

3. Repeat Step 2 until the duality gap

$$\left|\mathrm{tr}(S\widehat{\Omega}) - \log|\widehat{\Omega}| + \lambda\sum_{a,b}||\widehat{\Omega}_{ab}||_F - \sum_{j\in V} k_j - \log|\Sigma|\right| \leqslant \epsilon,$$

   where ϵ is a prefixed precision parameter (for example, $\epsilon = 10^{-3}$).

Finally, we form a graph $\widehat{G} = (V, \widehat{E})$ by connecting nodes with $||\widehat{\Omega}_{ab}||_F \neq 0$.

Computational complexity of the procedure is given in Appendix A. Convergence of the above described procedure to the unique minimum of the objective function in (4) does not follow from the standard results on the block coordinate descent algorithm (Tseng, 2001) for two reasons. First, the minimization problem in (6) is not solved exactly at each iteration, since we only update Ω_{aa} and $\Omega_{a,\overline{a}}$ using one generalized gradient step update in each iteration. Second, the blocks of variables, over which the optimization is done at each iteration, are not completely separable between iterations due to the symmetry of the problem. The proof of the following convergence result is given in Appendix B.

Lemma 1 *For every value of $\lambda > 0$, the above described algorithm produces a sequence of estimates $\left\{\widetilde{\Omega}^{(t)}\right\}_{t\geqslant 1}$ of the precision matrix that monotonically decrease the objective values given in* (4)*. Every element of this sequence is positive definite and the sequence converges to the unique minimizer $\widehat{\Omega}$ of* (4)*.*

2.3 Efficient Identification of Connected Components

When the target graph $\widehat{G}$ is composed of smaller, disconnected components, the solution to the problem in (4) is block diagonal (possibly after permuting the node indices) and can be obtained by solving smaller optimization problems. That is, the minimizer $\widehat{\Omega}$ can be obtained by solving (4) for each connected component independently, resulting in massive computational gains. We give necessary and sufficient condition for the solution $\widehat{\Omega}$ of (4) to be block-diagonal, which can be easily checked by inspecting the empirical covariance matrix S.

Our first result follows immediately from the Karush-Kuhn-Tucker conditions for the optimization problem (4) and states that if $\widehat{\Omega}$ is block-diagonal, then it can be obtained by solving a sequence of smaller optimization problems.

Lemma 2 *If the solution to* (4) *takes the form* $\widehat{\Omega} = \mathrm{diag}(\widehat{\Omega}_1, \widehat{\Omega}_2, \ldots, \widehat{\Omega}_l)$*, that is,* $\widehat{\Omega}$ *is a block diagonal matrix with the diagonal blocks* $\widehat{\Omega}_1, \ldots, \widehat{\Omega}_l$*, then it can be obtained by solving*

$$\min_{\Omega_{l'} > 0} \left\{ \mathrm{tr}\, S_{l'} \Omega_{l'} - \log |\Omega_{l'}| + \lambda \sum_{a,b} ||\Omega_{ab}||_F \right\}$$

separately for each $l' = 1, \ldots, l$*, where* $S_{l'}$ *are submatrices of* S *corresponding to* $\Omega_{l'}$*.*

Next, we describe how to identify diagonal blocks of $\widehat{\Omega}$. Let $\mathcal{P} = \{P_1, P_2, \ldots, P_l\}$ be a partition of the set V and assume that the nodes of the graph are ordered in a way that if $a \in P_j$, $b \in P_{j'}$, $j < j'$, then $a < b$. The following lemma states that the blocks of $\widehat{\Omega}$ can be obtained from the blocks of a thresholded sample covariance matrix.

Lemma 3 *A necessary and sufficient condition for* $\widehat{\Omega}$ *to be block diagonal with blocks* $P_1, P_2, \ldots, P_l$ *is that* $||S_{ab}||_F \leqslant \lambda$ *for all* $a \in P_j$, $b \in P_{j'}$, $j \neq j'$*.*

Blocks $P_1, P_2, \ldots, P_l$ can be identified by forming a $p \times p$ matrix Q with elements $q_{ab} = \mathbb{1}\{||S_{ab}||_F > \lambda\}$ and computing connected components of the graph with adjacency matrix Q. The lemma states also that given two penalty parameters $\lambda_1 < \lambda_2$, the set of unconnected nodes with penalty parameter λ_1 is a subset of unconnected nodes with penalty parameter λ_2. The simple check above allows us to estimate graphs on data sets with large number of nodes, if we are interested in graphs with small number of edges. However, this is often the case when the graphs are used for exploration and interpretation of complex systems. Lemma 3 is related to existing results established for speeding-up computation when learning single and multiple Gaussian graphical models (Witten et al., 2011; Mazumder and Hastie, 2012; Danaher et al., 2014). Each condition is different, since the methods optimize different objective functions.

3. Consistent Graph Identification

In this section, we provide theoretical analysis of the estimator described in Section 2.2. In particular, we provide sufficient conditions for consistent graph recovery. For simplicity of presentation, we assume that $k_a = k$, for all $a \in V$, that is, we assume that the same number of attributes is observed for each node. For each $a = 1, \ldots, kp$, we assume that $(\sigma^*_{aa})^{-1/2} X_a$ is sub-Gaussian with parameter γ, where σ^*_{aa} is the ath diagonal element of Σ^*. Recall that Z is a sub-Gaussian random variable if there exists a constant $\sigma \in (0, \infty)$ such that

$$E\left(\exp(tZ)\right) \leqslant \exp(\sigma^2 t^2), \text{ for all } t \in \mathbb{R}.$$

Our assumptions involve the Hessian of the function $f(A) = \mathrm{tr}\, SA - \log |A|$ evaluated at the true Ω^*, $\mathcal{H} = \mathcal{H}(\Omega^*) = (\Omega^*)^{-1} \otimes (\Omega^*)^{-1} \in \mathbb{R}^{(pk)^2 \times (pk)^2}$, with $\otimes$ denoting the Kronecker product, and the true covariance matrix Σ^*. The Hessian and the covariance matrix can be thought of as block matrices with blocks of size $k^2 \times k^2$ and $k \times k$, respectively. We will make use of the operator $\mathcal{C}(\cdot)$ that operates on these block matrices and outputs a smaller matrix with elements that equal to the Frobenius norm of the original blocks. For example, $\mathcal{C}(\Sigma^*) \in \mathbb{R}^{p \times p}$ with elements $\mathcal{C}(\Sigma^*)_{ab} = ||\Sigma^*_{ab}||_F$. Let $\mathcal{T} = \{(a, b) : ||\Omega_{ab}||_F \neq 0\}$ and

$\mathcal{N} = \{(a,b) : ||\Omega_{ab}||_F = 0\}$. With this notation introduced, we assume that the following irrepresentable condition holds. There exists a constant $\alpha \in [0,1)$ such that

$$\|\!|\mathcal{C}\left(\mathcal{H}_{\mathcal{N}\mathcal{T}}(\mathcal{H}_{\mathcal{T}\mathcal{T}})^{-1}\right)|\!\|_\infty \leqslant 1-\alpha, \tag{12}$$

where $|\!|\!|A|\!|\!|_\infty = \max_i \sum_j |A_{ij}|$. We will also need the following quantities to specify the results $\kappa_{\Sigma *} = |\!|\!|\mathcal{C}(\Sigma^*)|\!|\!|_\infty$ and $\kappa_{\mathcal{H}} = |\!|\!|\mathcal{C}(\mathcal{H}_{\mathcal{T}\mathcal{T}}^{-1})|\!|\!|_\infty$. These conditions extend the conditions specified in Ravikumar et al. (2011) needed for estimating graphs from single attribute observations.

We have the following result that provides sufficient conditions for the exact recovery of the graph.

Proposition 4 *Let $\tau > 2$. We set the penalty parameter λ in (4) as*

$$\lambda = 8k\alpha^{-1}\left(128(1+4\gamma^2)^2(\max_a(\sigma^*_{aa})^2)n^{-1}(2\log(2k)+\tau\log(p))\right)^{1/2}.$$

If $n > C_1 s^2 k^2(1+8\alpha^{-1})^2(\tau\log p + \log 4 + 2\log k)$, where s is the maximal degree of nodes in G, $C_1 = (48\sqrt{2}(1+4\gamma^2)(\max_a \sigma^*_{aa})\max(\kappa_{\Sigma*}\kappa_{\mathcal{H}}, \kappa^3_{\Sigma*}\kappa^2_{\mathcal{H}}))^2$ and

$$\min_{(a,b)\in\mathcal{T}, a\neq b} ||\Omega_{ab}||_F > 16\sqrt{2}(1+4\gamma^2)(\max_a \sigma^*_{aa})(1+8\alpha^{-1})\kappa_{\mathcal{H}} k\left(\frac{\tau\log p+\log 4+2\log k}{n}\right)^{1/2},$$

then $\text{pr}\left(\widehat{G} = G\right) \geqslant 1 - p^{2-\tau}$.

The proof of Proposition 4 is given in Appendix B. We extend the proof of Ravikumar et al. (2011) to accommodate the Frobenius norm penalty on blocks of the precision matrix. This proposition specifies the sufficient sample size and a lower bound on the Frobenius norm of the off-diagonal blocks needed for recovery of the unknown graph. Under these conditions and correctly specified tuning parameter λ, the solution to the optimization problem in (4) correctly recovers the graph with high probability. In practice, one needs to choose the tuning parameter in a data dependent way. For example, using the Bayesian information criterion. Even though our theoretical analysis obtains the same rate of convergence as that of Ravikumar et al. (2011), our method has a significantly improved finite-sample performance, as will be shown in Section 5. It remains an open question whether the sample size requirement can be improved as in the case of group Lasso (see, for example, Lounici et al., 2011). The analysis of Lounici et al. (2011) relies heavily on the special structure of the least squares regression. Hence, their method does not carry over to the more complicated objective function as in (4).

4. Interpreting Edges

We propose a post-processing step that will allow us to quantify the strength of links identified by the method proposed in Section 2.2, as well as identify important attributes that contribute to the existence of links.

For any two nodes a and b for which $\Omega_{ab} \neq 0$, we define $\mathcal{N}(a,b) = \{c \in V\backslash\{a,b\} \ : \ \Omega_{ac} \neq 0 \text{ or } \Omega_{bc} \neq 0\}$, which is the Markov blanket for the set of nodes $\{X_a, X_b\}$. Note that the

conditional distribution of $(X_a^T, X_b^T)^T$ given $X_{\neg ab}$ is equal to the conditional distribution of $(X_a^T, X_b^T)^T$ given $X_{\mathcal{N}(a,b)}$. Now,

$$\begin{aligned}\rho_c(X_a, X_b; X_{\neg ab}) &= \rho_c(X_a, X_b; X_{\mathcal{N}(a,b)}) \\ &= \max_{w_a \in \mathbb{R}^{k_a}, w_b \in \mathbb{R}^{k_b}} \operatorname{corr}(u^T(X_a - \widetilde{A} X_{\mathcal{N}(a,b)}), v^T(X_b - \widetilde{B} X_{\mathcal{N}(a,b)})),\end{aligned}$$

where $\widetilde{A} = \operatorname{argmin}\ E\left(||X_a - A X_{\mathcal{N}(a,b)}||_2^2\right)$ and $\widetilde{B} = \operatorname{argmin}\ E\left(||X_b - B X_{\mathcal{N}(a,b)}||_2^2\right)$. Let $\overline{\Sigma}(a,b) = \operatorname{var}(X_a, X_b \mid X_{\mathcal{N}(a,b)})$. Now we can express the partial canonical correlation as

$$\rho_c(X_a, X_b; X_{\mathcal{N}(a,b)}) = \max_{w_a \in \mathbb{R}^{k_a}, w_b \in \mathbb{R}^{k_a}} \frac{w_a^T \overline{\Sigma}_{ab} w_b}{\left(w_a^T \overline{\Sigma}_{aa} w_a\right)^{1/2} \left(w_b^T \overline{\Sigma}_{bb} w_b\right)^{1/2}},$$

where

$$\overline{\Sigma}(a,b) = \begin{pmatrix} \overline{\Sigma}_{aa} & \overline{\Sigma}_{ab} \\ \overline{\Sigma}_{ba} & \overline{\Sigma}_{bb} \end{pmatrix}.$$

The weight vectors w_a and w_b can be easily found by solving the system of eigenvalue equations

$$\begin{cases} \overline{\Sigma}_{aa}^{-1} \overline{\Sigma}_{ab} \overline{\Sigma}_{bb}^{-1} \overline{\Sigma}_{ba} w_a = \phi^2 w_a \\ \overline{\Sigma}_{bb}^{-1} \overline{\Sigma}_{ba} \overline{\Sigma}_{aa}^{-1} \overline{\Sigma}_{ab} w_b = \phi^2 w_b \end{cases} \tag{13}$$

with w_a and w_b being the vectors that correspond to the maximum eigenvalue ϕ^2. Furthermore, we have $\rho_c(X_a, X_b; X_{\mathcal{N}(a,b)}) = \phi$. Following Katenka and Kolaczyk (2011), the weights w_a, w_b can be used to access the relative contribution of each attribute to the edge between the nodes a and b. In particular, the weight $(w_{a,i})^2$ characterizes the relative contribution of the ith attribute of node a to $\rho_c(X_a, X_b; X_{\mathcal{N}(a,b)})$.

Given an estimate $\widehat{\mathcal{N}}(a,b) = \{c \in V \backslash \{a,b\} \ :\ \widehat{\Omega}_{ac} \neq 0 \text{ or } \widehat{\Omega}_{bc} \neq 0\}$ of the Markov blanket $\mathcal{N}(a,b)$, we form the residual vectors

$$r_{i,a} = x_{i,a} - \breve{A} x_{i,\widehat{\mathcal{N}}(a,b)}, \qquad r_{i,b} = x_{i,b} - \breve{B} x_{i,\widehat{\mathcal{N}}(a,b)},$$

where $\breve{A}$ and $\breve{B}$ are the least square estimators of $\widetilde{A}$ and $\widetilde{B}$. Given the residuals, we form $\breve{\Sigma}(a,b)$, the empirical version of the matrix $\overline{\Sigma}(a,b)$, by setting

$$\breve{\Sigma}_{aa} = \operatorname{corr}\left(\{r_{i,a}\}_{i \in [n]}\right), \quad \breve{\Sigma}_{bb} = \operatorname{corr}\left(\{r_{i,b}\}_{i \in [n]}\right), \quad \breve{\Sigma}_{ab} = \operatorname{corr}\left(\{r_{i,a}\}_{i \in [n]}, \{r_{i,a}\}_{i \in [n]}\right).$$

Now, solving the eigenvalue system in (13) will give us estimates of the vectors w_a, w_b and the partial canonical correlation.

Note that we have described a way to interpret the elements of the off-diagonal blocks in the estimated precision matrix. The elements of the diagonal blocks, which correspond to coefficients between attributes of the same node, can still be interpreted by their relationship to the partial correlation coefficients.

5. Simulation Studies

In this section, we perform a set of simulation studies to illustrate finite sample performance of our method. We demonstrate that the scalings of (n, p, s) predicted by the theory are sharp. Furthermore, we compare against three other methods: 1) a method that uses the glasso first to estimate one graph over each of the k individual attributes and then creates an edge in the resulting graph if an edge appears in at least one of the single attribute graphs, 2) the method of Guo et al. (2011) and 3) the method of Danaher et al. (2014). We have also tried applying the glasso to estimate the precision matrix for the model in (1) and then post-processing it, so that an edge appears in the resulting graph if the corresponding block of the estimated precision matrix is non-zero. The result of this method is worse compared to the first baseline, so we do not report it here.

All the methods above require setting one or two tuning parameters that control the sparsity of the estimated graph. We select these tuning parameters by minimizing the Bayesian information criterion (Schwarz, 1978), which balances the goodness of fit of the model and its complexity, over a grid of parameter values. For our multi-attribute method, the Bayesian information criterion takes the following form

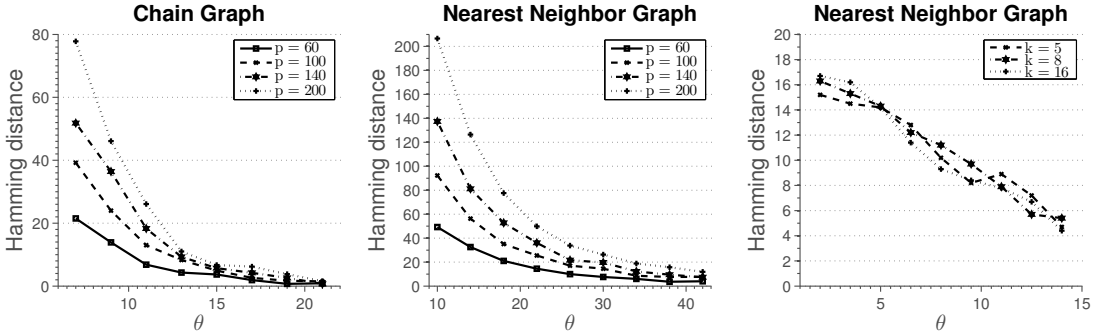
$$\mathrm{BIC}(\lambda) = \mathrm{tr}(S\widehat{\Omega}) - \log|\widehat{\Omega}| + \sum_{a<b} \mathbb{1}\{\widehat{\Omega}_{ab} \neq 0\} k_a k_b \log(n).$$

Other methods for selecting tuning parameters are possible, like minimization of cross-validation or Akaike information criterion (Akaike, 1974). However, these methods tend to select models that are too dense.

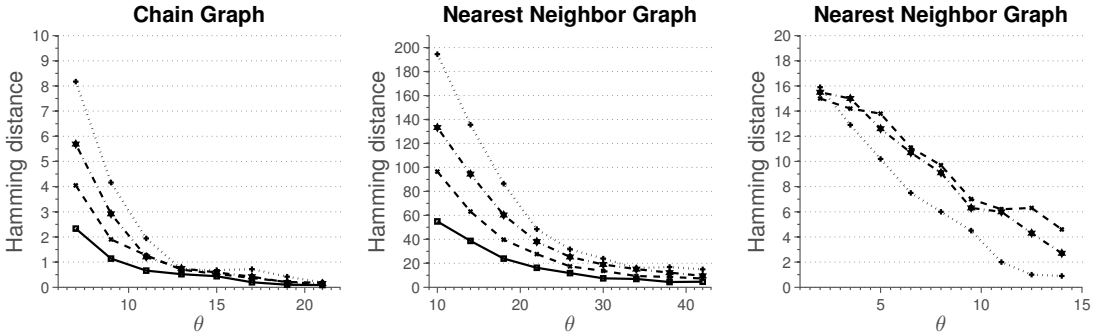
Theoretical results given in Section 3 characterize the sample size needed for consistent recovery of the underlying graph. In particular, Proposition 4 suggests that we need $n = \theta s^2 k^2 \log(pk)$ samples to estimate the graph structure consistently, for some control parameter $\theta > 0$. Therefore, if we plot the hamming distance between the true and recovered graph against θ, we expect the curves to reach zero distance for different problem sizes at a same point. We verify this on randomly generated chain and nearest-neighbors graphs.

Simulation 1. We generate data as follows. A random graph with p nodes is created by first partitioning nodes into $p/20$ connected components, each with 20 nodes, and then forming a random graph over these 20 nodes. A chain graph is formed by permuting the nodes and connecting them in succession, while a nearest-neighbor graph is constructed following the procedure outlined in Li and Gui (2006). That is, for each node, we draw a point uniformly at random on a unit square and compute the pairwise distances between nodes. Each node is then connected to $s = 4$ closest neighbors. Since some of nodes will have more than 4 adjacent edges, we randomly remove edges from nodes that have degree larger than 4 until the maximum degree of a node in a graph is 4. Once the graph is created, we construct a precision matrix, with non-zero blocks corresponding to edges in the graph. Elements of diagonal blocks are set as $0.5^{|a-b|}$, $0 \leqslant a, b \leqslant k$, while off-diagonal blocks have elements with the same value, 0.2 for chain graphs and $0.3/k$ for nearest-neighbor graph. Finally, we add ρI to the precision matrix, so that its minimum eigenvalue is equal to 0.5. Note that $s = 2$ for the chain graph and $s = 4$ for the nearest-neighbor graph. Simulation results are averaged over 100 replicates.

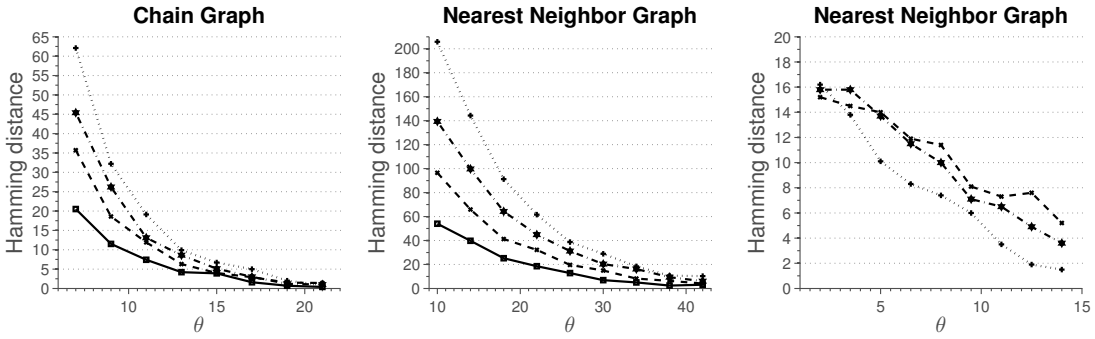
Figure 1 shows simulation results. Each row in the figure reports results for one method, while each column in the figure represents a different simulation setting. For the first two



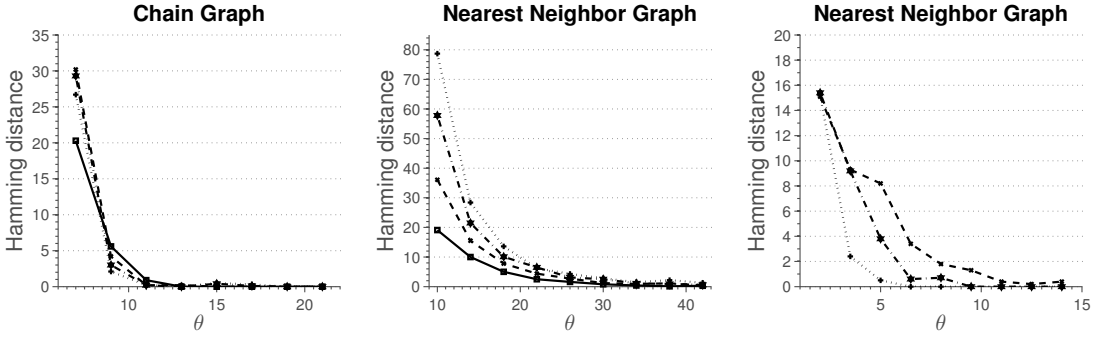
(a) glasso procedure



(b) Procedure of Danaher et al. (2014)



(c) Procedure of Guo et al. (2011)



(d) Multi-attribute procedure

Figure 1: Results of Simulation 1. Average hamming distance plotted against the rescaled sample size. Off-diagonal blocks are full matrices.

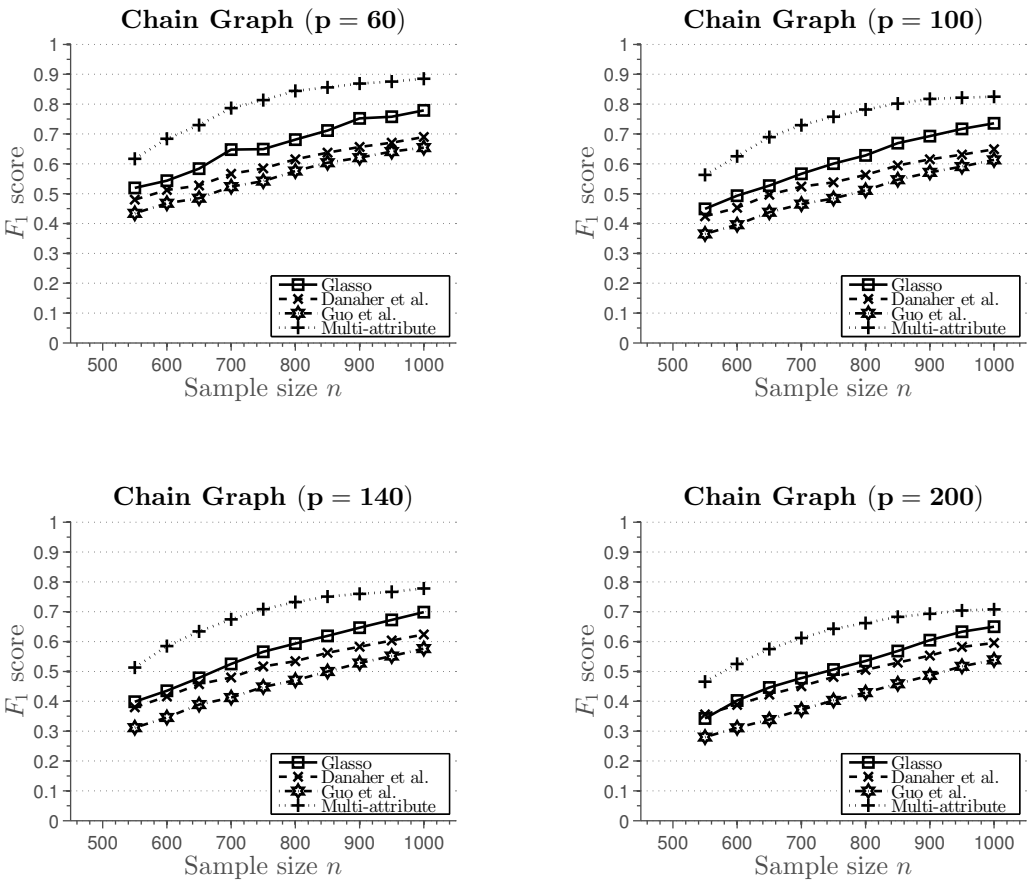


Figure 2: Results of Simulation 1 for smaller sample size. The number of attributes is $k = 2$. Average F_1 score plotted against the sample size.

columns, we set $k = 3$ and vary the total number of nodes in the graph. The third simulation setting sets the total number of nodes $p = 20$ and changes the number of attributes k. In the case of the chain graph, we observe that for small sample sizes the method of Danaher et al. (2014) outperforms all the other methods. We note that the multi-attribute method is estimating many more parameters, which require large sample size in order to achieve high accuracy. However, as the sample size increases, we observe that multi-attribute method starts to outperform the other methods. In particular, for the sample size indexed by $\theta = 13$ all the graph are correctly recovered, while other methods fail to recover the graph consistently at the same sample size. In the case of nearest-neighbor graph, none of the methods recover the graph well for small sample sizes. However, for moderate sample sizes, multi-attribute method outperforms the other methods. Furthermore, as the sample size increases none of the other methods recover the graph exactly. This suggests that the conditions for consistent graph recovery may be weaker in the multi-attribute setting.

From Figure 1 we can observe that for sufficiently large sample size n, multi-attribute method recovers the graph structure exactly. Next, we evaluate performance of the methods

for smaller sample sizes, with the number of attributes $k = 2$. Figure 2 shows average F_1 score plotted against the sample size.[2] This figure shows more precisely performance of the methods for smaller sample sizes that are not sufficient for perfect graph recovery. Again, even though none of the methods perform well, we can observe somewhat better performance of the multi-attribute procedure.

5.1 Alternative Structure of Off-diagonal Blocks

In this section, we investigate performance of different estimation procedures under different assumptions on the elements of the off-diagonal blocks of the precision matrix.

Simulation 2. First, we investigate a situation where the multi-attribute method does not perform as well as the methods that estimate multiple graphical models. One such situation arises when different attributes are conditionally independent. To simulate this situation, we use the data generating approach as before, however, we make each block Ω_{ab} of the precision matrix Ω a diagonal matrix. Figure 3 summarizes results of the simulation. We see that the methods of Danaher et al. (2014) and Guo et al. (2011) perform better, since they are estimating much fewer parameters than the multi-attribute method. glasso does not exploit any structural information underlying the estimation problem and requires larger sample size to correctly estimate the graph than other methods.

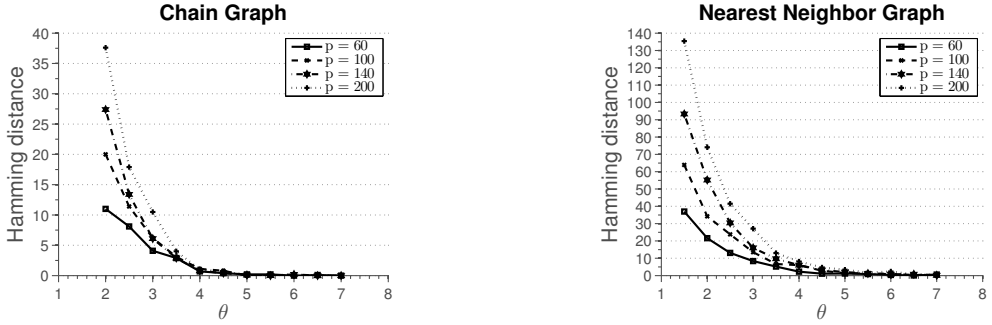
Simulation 3. A completely different situation arises when the edges between nodes can be inferred only based on inter-attribute data, that is, when a graph based on any individual attribute is empty. To generate data under this situation, we follow the procedure as before, but with the diagonal elements of the off-diagonal blocks Ω_{ab} set to zero. Figure 4 summarizes results of the simulation. In this setting, we clearly see the advantage of the multi-attribute method, compared to other three methods. Furthermore, we can see that glasso does better than multi-graph methods of Danaher et al. (2014) and Guo et al. (2011). The reason is that glasso can identify edges based on inter-attribute relationships among nodes, while multi-graph methods rely only on intra-attribute relationships. This simulation illustrates an extreme scenario where inter-attribute relationships are important for identifying edges.

Simulation 4. So far, off-diagonal blocks of the precision matrix were constructed to have constant values. Now, we use the same data generating procedure, but generate off-diagonal blocks of a precision matrix in a different way. Each element of the off-diagonal block Ω_{ab} is generated independently and uniformly from the set $[-0.3, -0.1] \bigcup [0.1, 0.3]$. The results of the simulation are given in Figure 5. Again, qualitatively, the results are similar to those given in Figure 1, except that in this setting more samples are needed to recover the graph correctly.

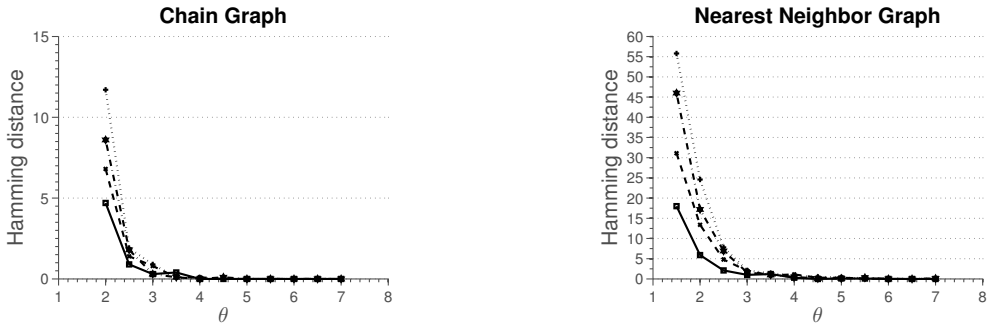
5.2 Different Number of Samples per Attribute

In this section, we show how to deal with a case when different number of samples is available per attribute. As noted in Section 2.2, we can treat non-measured attributes as missing completely at random (see Kolar and Xing, 2012, for more details).

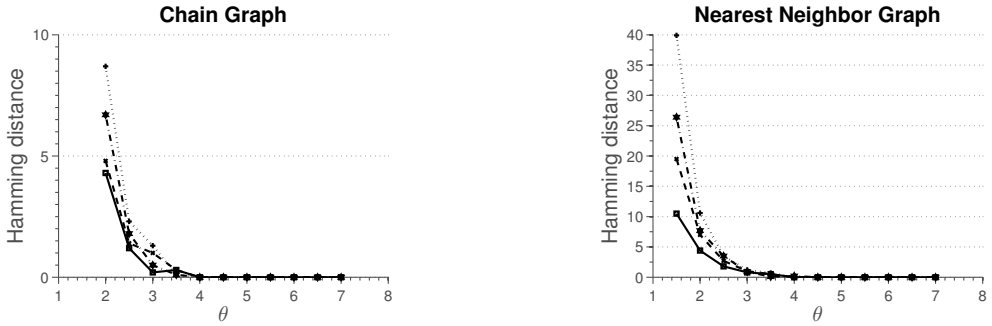
2. The F_1 score is a measure commonly used in information retrieval and is defined as the harmonic mean of precision and recall, that is, $F_1 := 2 * \text{precision} * \text{recall}/(\text{precision} + \text{recall})$. The precision is defined as $\text{precision} := |\widehat{E} \cap E|/|\widehat{E}|$ and the recall is defined as $\text{recall} := |\widehat{E} \cap E|/|E|$.



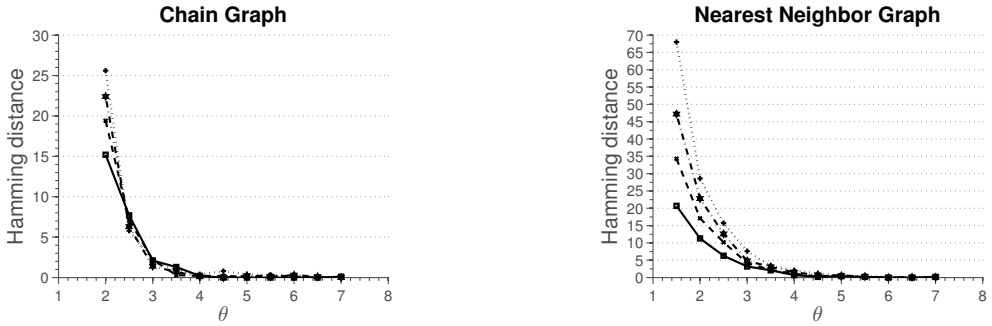
(a) glasso procedure



(b) Procedure of Danaher et al. (2014)

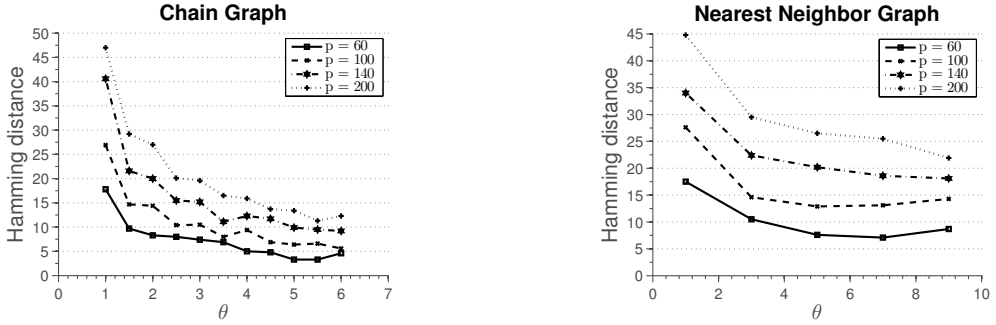


(c) Procedure of Guo et al. (2011)

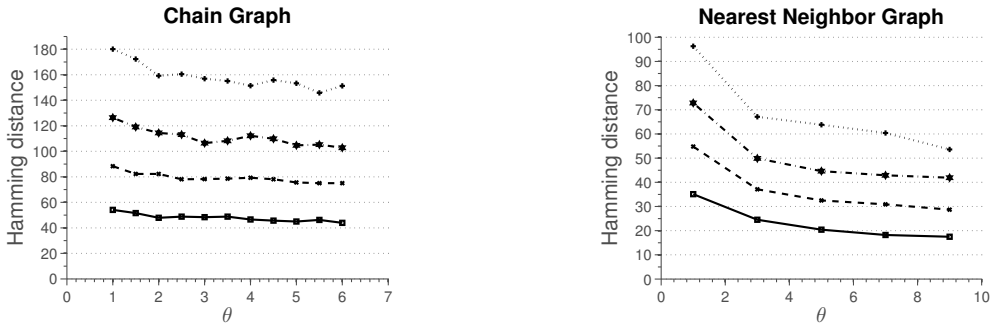


(d) Multi-attribute procedure

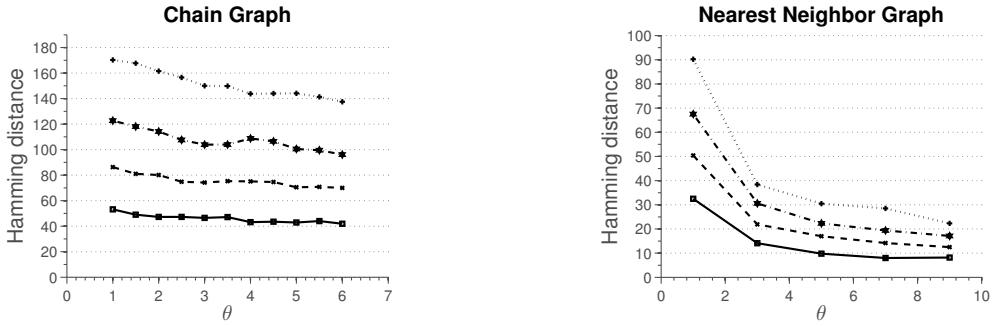
Figure 3: Results of Simulation 2 described in Section 5.1. Average hamming distance plotted against the rescaled sample size. Blocks Ω_{ab} of the precision matrix Ω are diagonal matrices.



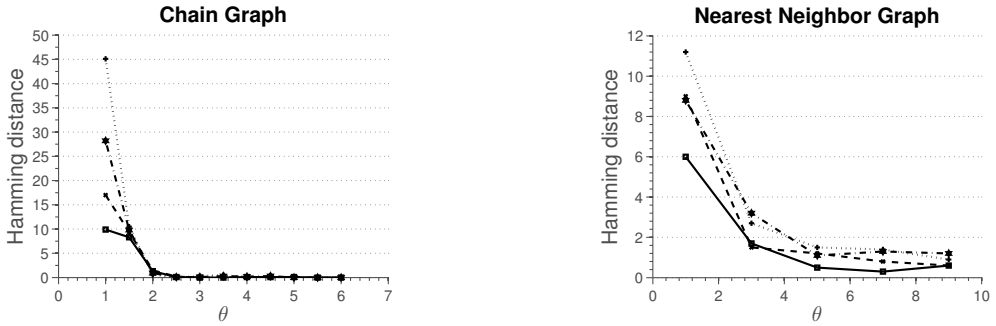
(a) glasso procedure



(b) Procedure of Danaher et al. (2014)

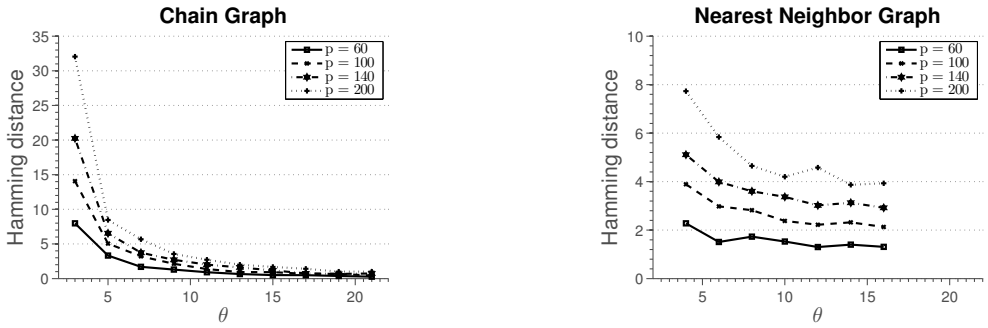


(c) Procedure of Guo et al. (2011)

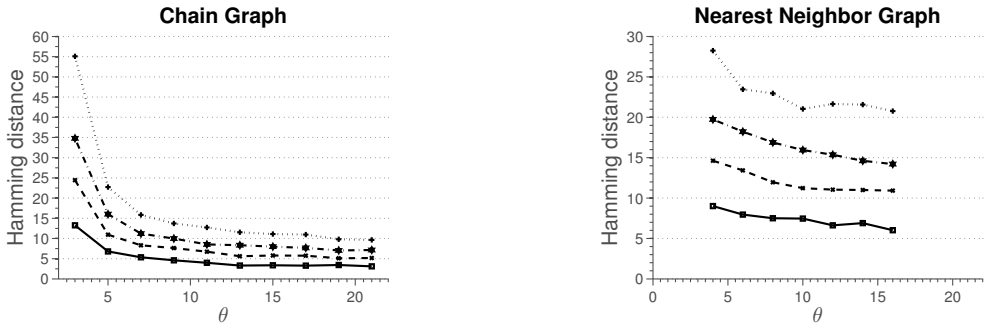


(d) Multi-attribute procedure

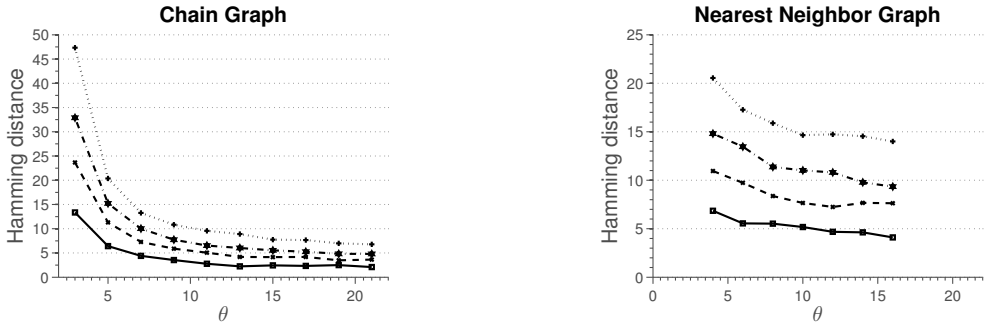
Figure 4: Results of Simulation 3 described in Section 5.1. Average hamming distance plotted against the rescaled sample size. Off-diagonal blocks Ω_{ab} of the precision matrix Ω have zeros as diagonal elements.



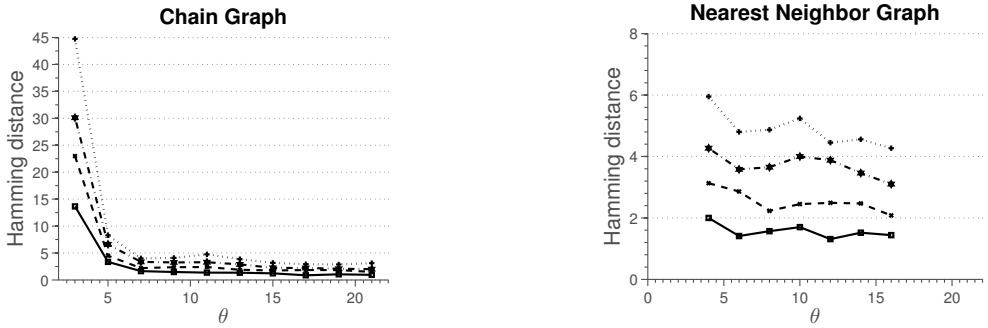
(a) glasso procedure



(b) Procedure of Danaher et al. (2014)



(c) Procedure of Guo et al. (2011)



(d) Multi-attribute procedure

Figure 5: Results of Simulation 4 described in Section 5.1. Average hamming distance plotted against the rescaled sample size. Off-diagonal blocks Ω_{ab} of the precision matrix Ω have elements uniformly sampled from $[-0.3, -0.1] \bigcup [0.1, 0.3]$.

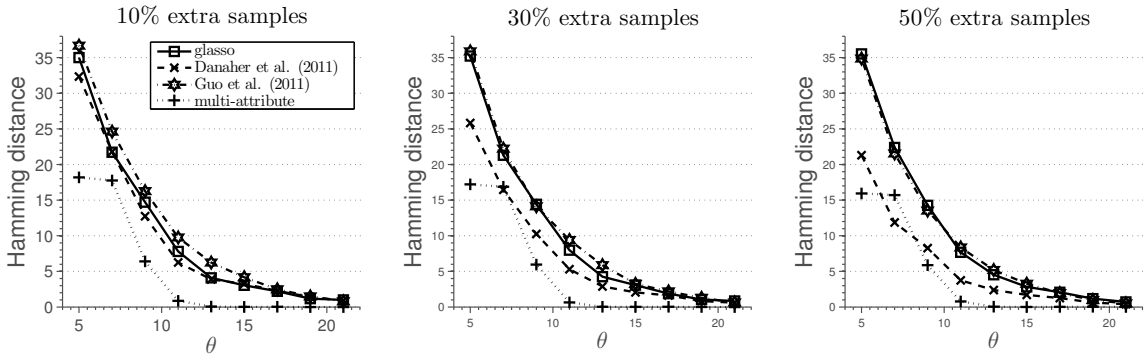


Figure 6: Results of Simulation 5 described in Section 5.2. Average hamming distance plotted against the rescaled sample size. Additional samples available for the first attribute.

Let $R = (r_{il})_{i\in\{1,\ldots,n\},l\in\{1,\ldots,pk\}} \in \mathbb{R}^{n\times pk}$ be an indicator matrix, which denotes for each sample point x_i the components that are observed. Then we can form an estimate of the sample covariance matrix $S = (\sigma_{lk}) \in \mathbb{R}^{pk\times pk}$ as

$$\sigma_{lk} = \frac{\sum_{i=1}^{n} r_{i,l} r_{i,k} x_{i,l} x_{i,k}}{\sum_{i=1}^{n} r_{i,l} r_{i,k}}.$$

This estimate is plugged into the objective in (4).

Simulation 5. We generate a chain graph with $p = 60$ nodes, construct a precision matrix associated with the graph and $k = 3$ attributes, and generate $n = \theta s^2 k^2 \log(pk)$ samples, $\theta > 0$. Next, we generate additional 10%, 30% and 50% samples from the same model, but record only the values for the first attribute. Results of the simulation are given in Figure 6. Qualitatively, the results are similar to those presented in Figure 1.

5.3 Scale-Free Graphs

In this section, we show simulation results when the methods are applied to estimate structure of scale-free graph, that is, graph whose degree distribution follows a power low. A prominent characteristic of these graphs is presence of hub nodes.[3] Such graphs commonly arise in studies of real world systems, such as gene or protein networks (Albert and Barabási, 2002).

Simulation 6. We generate a scale-free graph using the preferential attachment procedure described in Barabási and Albert (1999). The procedure starts with a 4-node cycle. New nodes are added to the graph, one at a time, and connected to nodes currently in the graph with probability proportional to their degree. Once the graph is generated, parameters in the model are set as in Simulation 1, with the number of attributes $k = 2$. Simulation results are summarized in Figure 7. These networks are harder to estimate using the ℓ_1-penalized procedures, due to the presence of high-degree hubs (Peng et al., 2009).

3. Hub nodes are nodes whose degree greatly exceeds average degree in a network.

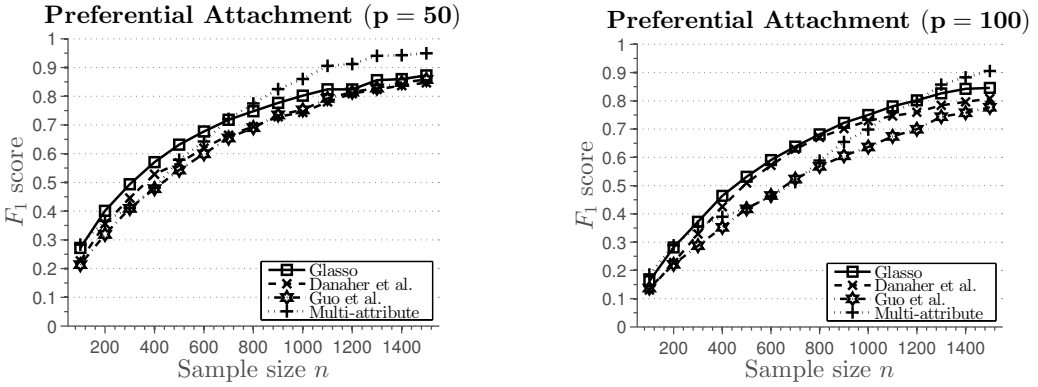


Figure 7: Results of Simulation 6 described in Section 5.3. Average F_1 score plotted against the sample size.

6. Illustrative Applications to Real Data

In this section, we illustrate how to apply our method to data arising in studies of biological regulatory networks and Alzheimer's disease.

6.1 Analysis of a Gene/Protein Regulatory Network

We provide illustrative, exploratory analysis of data from the well-known NCI-60 database, which contains different molecular profiles on a panel of 60 diverse human cancer cell lines. Data set consists of protein profiles (normalized reverse-phase lysate arrays for 92 antibodies) and gene profiles (normalized RNA microarray intensities from Human Genome U95 Affymetrix chip-set for > 9000 genes). We focus our analysis on a subset of 91 genes/proteins for which both types of profiles are available. These profiles are available across the same set of 60 cancer cells. More detailed description of the data set can be found in Katenka and Kolaczyk (2011).

We inferred three types of networks: a network based on protein measurements alone, a network based on gene expression profiles and a single gene/protein network. For protein and gene networks we use the `glasso`, while for the gene/protein network, we use our procedure outlined in Section 2.2. We use the stability selection (Meinshausen and Bühlmann, 2010) procedure to estimate stable networks. In particular, we first select the penalty parameter λ using cross-validation, which over-selects the number of edges in a network. Next, we use the selected λ to estimate 100 networks based on random subsamples containing 80% of the data-points. Final network is composed of stable edges that appear in at least 95 of the estimated networks. Table 1 provides a few summary statistics for the estimated networks. Furthermore, protein and gene/protein networks share 96 edges, while gene and gene/protein networks share 104 edges. Gene and protein network share only 17 edges. Finally, 66 edges are unique to gene/protein network. Figure 8 shows node degree distributions for the three networks. We observe that the estimated networks are much sparser than the association networks in Katenka and Kolaczyk (2011), as expected due to

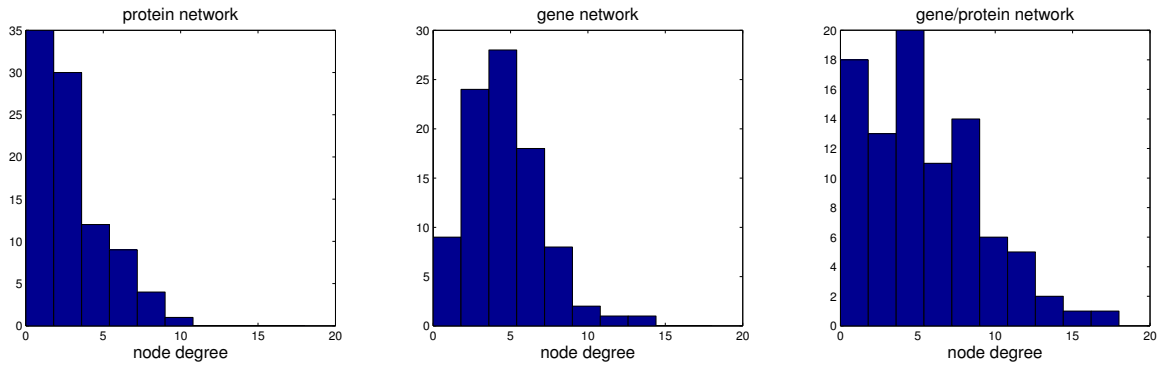


Figure 8: Node degree distributions for protein, gene and gene/protein networks.

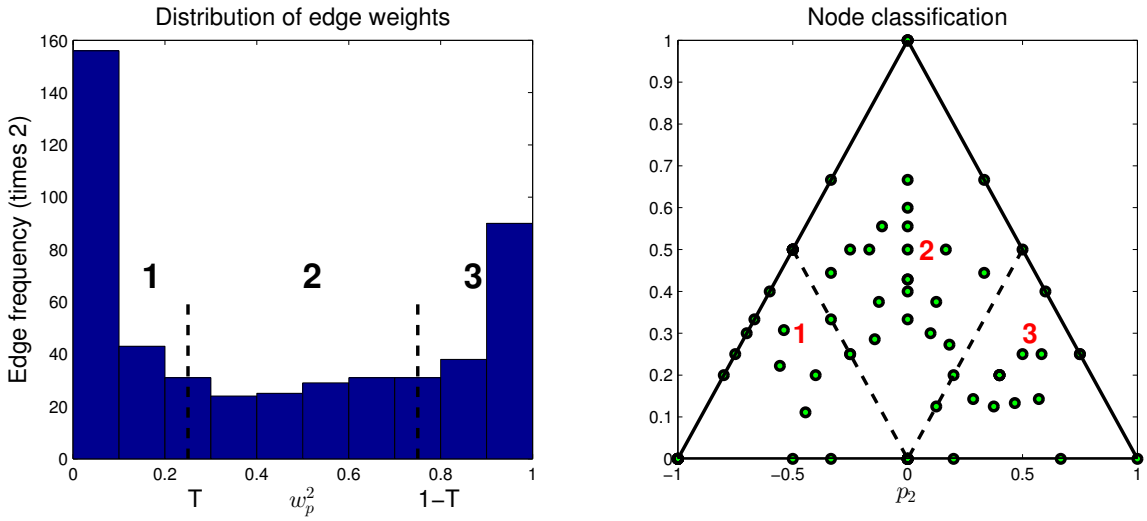


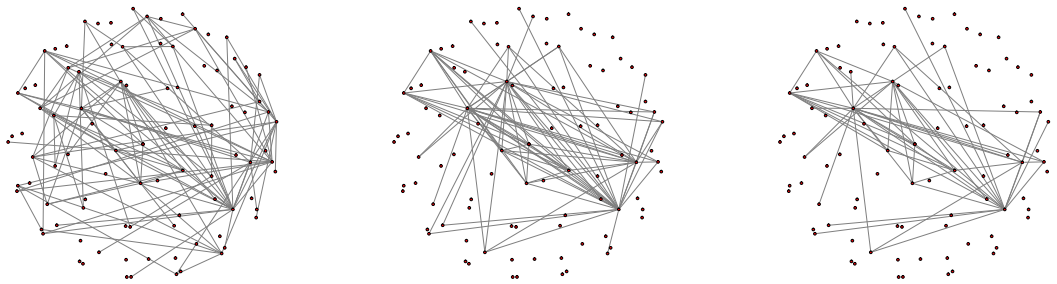
Figure 9: Edge and node classification based on w_p^2.

marginal correlations between a number of nodes. The differences in networks require a closer biological inspection by a domain scientist.

We proceed with a further exploratory analysis of the gene/protein network. We investigate the contribution of two nodal attributes to the existence of an edge between the nodes. Following Katenka and Kolaczyk (2011), we use a simple heuristic based on the

	protein network	gene network	gene/protein network
Number of edges	122	214	249
Density	0.03	0.05	0.06
Largest connected component	62	89	82
Avg Node Degree	2.68	4.70	5.47
Avg Clustering Coefficient	0.0008	0.001	0.003

Table 1: Summary statistics for protein, gene, and gene/protein networks ($p = 91$).



(a) Healthy subjects (b) Mild Cognitive Impairment (c) Alzheimer's & Dementia

Figure 10: Brain connectivity networks

weight vectors to classify the nodes and edges into three classes. For an edge between the nodes a and b, we take one weight vector, say w_a, and normalize it to have unit norm. Denote w_p the component corresponding to the protein attribute. Left plot in Figure 9 shows the values of w_p^2 over all edges. The edges can be classified into three classes based on the value of w_p^2. Given a threshold T, the edges for which $w_p^2 \in (0, T)$ are classified as gene-influenced, the edges for which $w_p^2 \in (1-T, 1)$ are classified as protein influenced, while the remainder of the edges are classified as mixed type. In the left plot of Figure 9, the threshold is set as $T = 0.25$ following Katenka and Kolaczyk (2011). Similar classification can be performed for nodes after computing the proportion of incident edges. Let p_1, p_2 and p_3 denote proportions of gene, protein and mixed edges, respectively, incident with a node. These proportions are represented in a simplex in the right subplot of Figure 9. Nodes with mostly gene edges are located in the lower left corner, while the nodes with mostly protein edges are located in the lower right corner. Mixed nodes are located in the center and towards the top corner of the simplex. Further biological enrichment analysis is possible (see Katenka and Kolaczyk, 2011), however, we do not pursue this here.

6.2 Uncovering Functional Brain Network

We apply our method to the Positron Emission Tomography data set, which contains 259 subjects, of whom 72 are healthy, 132 have mild cognitive Impairment and 55 are diagnosed as Alzheimer's & Dementia. Note that mild cognitive impairment is a transition stage from normal aging to Alzheimer's & Dementia. The data can be obtained from `http://adni.loni.ucla.edu/`. The preprocessing is done in the same way as in Huang et al. (2009).

Each Positron Emission Tomography image contains $91 \times 109 \times 91 = 902,629$ voxels. The effective brain region contains $180,502$ voxels, which are partitioned into 95 regions, ignoring the regions with fewer than 500 voxels. The largest region contains $5,014$ voxels and the smallest region contains 665 voxels. Our preprocessing stage extracts 948 representative voxels from these regions using the K-median clustering algorithm. The parameter K is chosen differently for each region, proportionally to the initial number of voxels in that

	Healthy subjects	Mild Cognitive Impairment	Alzheimer's & Dementia
Number of edges	116	84	59
Density	0.030	0.020	0.014
Largest connected component	48	27	25
Avg Node Degree	2.40	1.73	1.2
Avg Clustering Coefficient	0.001	0.0023	0.0007

Table 2: Summary statistics for brain connectivity networks

region. More specifically, for each category of subjects we have an $n \times (d_1 + \ldots + d_{95})$ matrix, where n is the number of subjects and $d_1 + \ldots + d_{95} = 902,629$ is the number of voxels. Next we set $K_i = \lceil d_i / \sum_j d_j \rceil$, the number of representative voxels in region i, $i = 1, \ldots, 95$. The representative voxels are identified by running the K-median clustering algorithm on a sub-matrix of size $n \times d_i$ with $K = K_i$.

We inferred three networks, one for each subtype of subjects using the procedure outlined in Section 2.2. Note that for different nodes we have different number of attributes, which correspond to medians found by the clustering algorithm. We use the stability selection (Meinshausen and Bühlmann, 2010) approach to estimate stable networks. The stability selection procedure is combined with our estimation procedure as follows. We first select the penalty parameter λ in (4) using cross-validation, which overselects the number of edges in a network. Next, we create 100 subsampled data sets, each of which contain 80% of the data points, and estimate one network for each data set using the selected λ. The final network is composed of stable edges that appear in at least 95 of the estimated networks.

We visualize the estimated networks in Figure 10. Table 2 provides a few summary statistics for the estimated networks. Appendix C contains names of different regions, as well as the adjacency matrices for networks. From the summary statistics, we can observe that in normal subjects there are many more connections between different regions of the brain. Loss of connectivity in Alzheimer's & Dementia has been widely reported in the literature (Greicius et al., 2004; Hedden et al., 2009; Andrews-Hanna et al., 2007; Wu et al., 2011).

Learning functional brain connectivity is potentially valuable for early identification of signs of Alzheimer's disease. Huang et al. (2009) approach this problem using exploratory data analysis. The framework of Gaussian graphical models is used to explore functional brain connectivity. Here we point out that our approach can be used for the same exploratory task, without the need to reduce the information in the whole brain to one number. For example, from our estimates, we observe the loss of connectivity in the cerebellum region of patients with Alzheimer's disease, which has been reported previously in Sjöbeck and Englund (2001). As another example, we note increased connectivity between the frontal lobe and other regions in the patients, which was linked to compensation for the lost connections in other regions (Stern, 2006; Gould et al., 2006).

7. Conclusion and Discussion

This paper extends the classical Gaussian graphical model to handle multi-attribute data. Multi-attribute data appear naturally in social media and scientific data analysis. For example, in a study of social networks, one may use personal information, including demographics, interests, and many other features, as nodal attributes. We proposed a new family of Gaussian graphical models for modeling such multi-attribute data. The main idea is to replace the notion of partial correlation in the existing graphical model literature by *partial canonical correlation*. Such a modification, though simple, has profound impact to both applications and theory. Practically, many challenging data, including brain imaging and gene expression profiles, can be naturally fitted using this model, which has been illustrated in the paper. Theoretically, we proved sufficient conditions that secure the correct recovery of the unknown population network structure.

The methods and theory of this paper can be naturally extended to handle non-Gaussian data by replacing the Gaussian model with the more general nonparanormal model (Liu et al., 2009) or the transelliptical model (Liu et al., 2012). Both of them can be viewed as semiparametric extensions of the Gaussian graphical model. Instead of assuming that data follows a Gaussian distribution, one assumes that there exists a set of strictly increasing univariate functions, so that after marginal transformation the data follows a Gaussian or Elliptical distribution. More details on model interpretation can be found in Liu et al. (2009) and Liu et al. (2012). To handle multi-attribute data in this semiparametric framework, we would replace the sample covariance matrix S in Eq. (4) by a rank-based correlation matrix estimator. We leave the formal analysis of this approach for future work.

Acknowledgments

We thank Eric D. Kolaczyk and Natallia V. Katenka for sharing preprocessed data used in their study with us. Eric P. Xing is partially supported through the grants NIH R01GM087694 and AFOSR FA9550010247. The research of Han Liu is supported by the grants NSF IIS-1116730, NSF III-1332109, NIH R01GM083084, NIH R01HG06841, and FDA HHSF223201000072C. This work was completed in part with resources provided by the University of Chicago Research Computing Center. We are grateful to the editor and three anonymous reviewers for helping us improve the manuscript.

Appendix A. Complexity Analysis of Multi-attribute Estimation

Step 2 of the estimation algorithm updates portions of the precision and covariance matrices corresponding to one node at a time. We observe that the computational complexity of updating the precision matrix is $\mathcal{O}\left(pk^2\right)$. Updating the covariance matrix requires computing $(\widetilde{\Omega}_{\overline{a},\overline{a}})^{-1}$, which can be efficiently done in $\mathcal{O}\left(p^2k^2 + pk^2 + k^3\right) = \mathcal{O}\left(p^2k^2\right)$ operations, assuming that $k \ll p$. With this, the covariance matrix can be updated in $\mathcal{O}\left(p^2k^2\right)$ operations. Therefore the total cost of updating the covariance and precision matrices is $\mathcal{O}\left(p^2k^2\right)$ operations. Since step 2 needs to be performed for each node $a \in V$, the total complexity is $\mathcal{O}\left(p^3k^2\right)$. Let T denote the total number of times step 2 is executed. This leads to the overall complexity of the algorithm as $\mathcal{O}\left(Tp^3k^2\right)$. In practice, we observe that $T \approx 10$ to 20 for

sparse graphs. Furthermore, when the whole solution path is computed, we can use warm starts to further speed up computation, leading to $T < 5$ for each λ.

Appendix B. Technical Proofs

In this appendix, we collect proofs of the results presented in the main part of the paper.

B.1 Proof of Lemma 1

We start the proof by giving to technical results needed later. The following lemma states that the minimizer of (4) is unique and has bounded minimum and maximum eigenvalues, denoted as $\Lambda_{\min}$ and $\Lambda_{\max}$.

Lemma 5 *For every value of $\lambda > 0$, the optimization problem in Eq. (4) has a unique minimizer $\widehat{\Omega}$, which satisfies $\Lambda_{\min}(\widehat{\Omega}) \geqslant (\Lambda_{\max}(S) + \lambda p)^{-1} > 0$ and $\Lambda_{\max}(\widehat{\Omega}) \leqslant \lambda^{-1} \sum_{j \in V} k_j$.*

Proof The optimization objective given in (4) can be written in the equivalent constrained form as

$$\min_{\Omega \succ 0} \ \operatorname{tr} S\Omega - \log |\Omega| \quad \text{subject to} \quad \sum_{a,b} ||\Omega_{ab}||_F \leqslant C(\lambda).$$

The procedure involves minimizing a continuous objective over a compact set, and so by Weierstrass theorem, the minimum is always achieved. Furthermore, the objective is strongly convex and therefore the minimum is unique.

The solution $\widehat{\Omega}$ to the optimization problem (4) satisfies

$$S - \widehat{\Omega}^{-1} + \lambda Z = 0, \tag{14}$$

where $Z \in \partial \sum_{a,b} ||\widehat{\Omega}_{ab}||_F$ is the element of the sub-differential and satisfies $||Z_{ab}||_F \leqslant 1$ for all $(a, b) \in V^2$. Therefore,

$$\Lambda_{\max}(\widehat{\Omega}^{-1}) \leqslant \Lambda_{\max}(S) + \lambda \Lambda_{\max}(Z) \leqslant \Lambda_{\max}(S) + \lambda p.$$

Next, we prove an upper bound on $\Lambda_{\max}(\widehat{\Omega})$. At optimum, the primal-dual gap is zero, which gives that

$$\sum_{a,b} ||\widehat{\Omega}_{ab}||_F \leqslant \lambda^{-1} (\sum_{j \in V} k_j - \operatorname{tr} S\widehat{\Omega}) \leqslant \lambda^{-1} \sum_{j \in V} k_j,$$

as $S \succeq 0$ and $\widehat{\Omega} \succ 0$. Since $\Lambda_{\max}(\widehat{\Omega}) \leqslant \sum_{a,b} ||\widehat{\Omega}_{ab}||_F$, the proof is done. ■

The next results states that the objective function has a Lipschitz continuous gradient, which will be used to show that the generalized gradient descent can be used to find $\widehat{\Omega}$.

Lemma 6 *The function $f(A) = \operatorname{tr} SA - \log |A|$ has a Lipschitz continuous gradient on the set $\{A \in \mathcal{S}^p \ : \ \Lambda_{\min}(A) \geqslant \gamma\}$, with the Lipschitz constant $L = \gamma^{-2}$.*

Proof We have that $\nabla f(A) = S - A^{-1}$. Then

$$\begin{aligned} ||\nabla f(A) - \nabla f(A')||_F &= ||A^{-1} - (A')^{-1}||_F \\ &\leqslant \Lambda_{\max} A^{-1} ||A - A'||_F \Lambda_{\max} A^{-1} \\ &\leqslant \gamma^{-2} ||A - A'||_F, \end{aligned}$$

which completes the proof. ■

Now, we provide the proof of Lemma 1.

By construction, the sequence of estimates $(\widetilde{\Omega}^{(t)})_{t\geqslant 1}$ decrease the objective value and are positive definite.

To prove the convergence, we first introduce some additional notation. Let $f(\Omega) = \operatorname{tr} S\Omega - \log|\Omega|$ and $F(\Omega) = f(\Omega) + \sum_{ab}||\Omega_{ab}||_F$. For any $L > 0$, let

$$Q_L(\Omega;\overline{\Omega}) := f(\overline{\Omega}) + \operatorname{tr}[(\Omega-\overline{\Omega})\nabla f(\overline{\Omega})] + \frac{L}{2}||\Omega-\overline{\Omega}||_F^2 + \sum_{ab}||\Omega_{ab}||_F$$

be a quadratic approximation of $F(\Omega)$ at a given point $\overline{\Omega}$, which has a unique minimizer

$$p_L(\overline{\Omega}) := \arg\min_{\Omega} Q_L(\Omega;\overline{\Omega}).$$

From Lemma 2.3. in Beck and Teboulle (2009), we have that

$$F(\overline{\Omega}) - F(p_L(\overline{\Omega})) \geqslant \frac{L}{2}||p_L(\overline{\Omega})-\overline{\Omega}||_F^2, \tag{15}$$

if $F(p_L(\overline{\Omega})) \leqslant Q_L(p_L(\overline{\Omega});\overline{\Omega})$. Note that $F(p_L(\overline{\Omega})) \leqslant Q_L(p_L(\overline{\Omega});\overline{\Omega})$ always holds if L is as large as the Lipschitz constant of ∇F.

Let $\widetilde{\Omega}^{(t-1)}$ and $\widetilde{\Omega}^{(t)}$ denote two successive iterates obtained by the procedure. Without loss of generality, we can assume that $\widetilde{\Omega}^{(t)}$ is obtained by updating the rows/columns corresponding to the node a. From (15), it follows that

$$\frac{2}{L_k}(F(\widetilde{\Omega}^{(t-1)}) - F(\widetilde{\Omega}^{(t)})) \geqslant ||\widetilde{\Omega}^{(t-1)}_{aa} - \widetilde{\Omega}^{(t)}_{aa}||_F + 2\sum_{b\neq a}||\widetilde{\Omega}^{(t-1)}_{ab} - \widetilde{\Omega}^{(t)}_{ab}||_F, \tag{16}$$

where L_k is a current estimate of the Lipschitz constant. Recall that in our procedure the scalar t serves as a local approximation of $1/L$. Since eigenvalues of $\widehat{\Omega}$ are bounded according to Lemma 5, we can conclude that the eigenvalues of $\widetilde{\Omega}^{(t-1)}$ are bounded as well. Therefore the current Lipschitz constant is bounded away from zero, using Lemma 6. Combining the results, we observe that the right hand side of (16) converges to zero as $t \to \infty$, since the optimization procedure produces iterates that decrease the objective value. This shows that $||\widetilde{\Omega}^{(t-1)}_{aa} - \widetilde{\Omega}^{(t)}_{aa}||_F + 2\sum_{b\neq a}||\widetilde{\Omega}^{(t-1)}_{ab} - \widetilde{\Omega}^{(t)}_{ab}||_F$ converges to zero, for any $a \in V$. Since $(\widetilde{\Omega}^{(t)}$ is a bounded sequence, it has a limit point, which we denote $\widehat{\Omega}$. It is easy to see, from the stationary conditions for the optimization problem given in (6), that the limit point $\widehat{\Omega}$ also satisfies the global KKT conditions to the optimization problem in (4).

B.2 Proof of Lemma 3

Suppose that the solution $\widehat{\Omega}$ to (4) is block diagonal with blocks $P_1, P_2, \ldots, P_l$. For two nodes a, b in different blocks, we have that $(\widehat{\Omega})^{-1}_{ab} = 0$ as the inverse of the block diagonal matrix is block diagonal. From the KKT conditions, it follows that $||S_{ab}||_F \leqslant \lambda$.

Now suppose that $||S_{ab}||_F \leqslant \lambda$ for all $a \in P_j, b \in P_{j'}, j \neq j'$. For every $l' = 1, \ldots, l$ construct

$$\widetilde{\Omega}_{l'} = \arg\min_{\Omega_{l'}>0} \operatorname{tr} S_{l'}\Omega_{l'} - \log|\Omega_{l'}| + \lambda\sum_{a,b}||\Omega_{ab}||_F.$$

Then $\widehat{\Omega} = \text{diag}(\widehat{\Omega}_1, \widehat{\Omega}_2, \ldots, \widehat{\Omega}_l)$ is the solution of (4) as it satisfies the KKT conditions.

B.3 Proof of Eq. (3)

First, we note that

$$\text{var}\left((X_a^T, X_b^T)^T \mid X_{\overline{ab}}\right) = \Sigma_{ab,ab} - \Sigma_{ab,\overline{ab}}\Sigma^{-1}_{\overline{ab},\overline{ab}}\Sigma_{\overline{ab},ab}$$

is the conditional covariance matrix of $(X_a^T, X_b^T)^T$ given the remaining nodes $X_{\overline{ab}}$ (see Proposition C.5 in Lauritzen (1996)). Define $\overline{\Sigma} = \Sigma_{ab,ab} - \Sigma_{ab,\overline{ab}}\Sigma^{-1}_{\overline{ab},\overline{ab}}\Sigma_{\overline{ab},ab}$. Partial canonical correlation between X_a and X_b is equal to zero if and only if $\overline{\Sigma}_{ab} = 0$. On the other hand, the matrix inversion lemma gives that $\Omega_{ab,ab} = \overline{\Sigma}^{-1}$. Now, $\Omega_{ab} = 0$ if and only if $\overline{\Sigma}_{ab} = 0$. This shows the equivalence relationship in Eq. (3).

B.4 Proof of Proposition 4

We provide sufficient conditions for consistent network estimation. Proposition 4 given in Section 3 is then a simple consequence. To provide sufficient conditions, we extend the work of Ravikumar et al. (2011) to our setting, where we observe multiple attributes for each node. In particular, we extend their Theorem 1.

For simplicity of presentation, we assume that $k_a = k$, for all $a \in V$, that is, we assume that the same number of attributes is observed for each node. Our assumptions involve the Hessian of the function $f(A) = \text{tr}\, SA - \log|A|$ evaluated at the true Ω^*,

$$\mathcal{H} = \mathcal{H}(\Omega^*) = (\Omega^*)^{-1} \otimes (\Omega^*)^{-1} \in \mathbb{R}^{(pk)^2 \times (pk)^2}, \tag{17}$$

and the true covariance matrix Σ^*. The Hessian and the covariance matrix can be thought of block matrices with blocks of size $k^2 \times k^2$ and $k \times k$, respectively. We will make use of the operator $\mathcal{C}(\cdot)$ that operates on these block matrices and outputs a smaller matrix with elements that equal to the Frobenius norm of the original blocks,

$$\begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1p} \\ A_{21} & A_{22} & \cdots & A_{2p} \\ \vdots & & \ddots & \vdots \\ A_{p1} & \cdots & & A_{pp} \end{pmatrix} \xrightarrow{\mathcal{C}(\cdot)} \begin{pmatrix} ||A_{11}||_F & ||A_{12}||_F & \cdots & ||A_{1p}||_F \\ ||A_{21}||_F & ||A_{22}||_F & \cdots & ||A_{2p}||_F \\ \vdots & & \ddots & \vdots \\ ||A_{p1}||_F & \cdots & & ||A_{pp}||_F \end{pmatrix}.$$

In particular, $\mathcal{C}(\Sigma^*) \in \mathbb{R}^{p \times p}$ and $\mathcal{C}(\mathcal{H}) \in \mathbb{R}^{p^2 \times p^2}$.

We denote the index set of the non-zero blocks of the precision matrix as

$$\mathcal{T} := \{(a,b) \in V \times V \ : \ ||\Omega^*_{ab}||_2 \neq 0\} \cup \{(a,a) \ : \ a \in V\}$$

and let $\mathcal{N}$ denote its complement in $V \times V$, that is,

$$\mathcal{N} = \{(a,b) \ : \ ||\Omega_{ab}||_F = 0\}.$$

As mentioned earlier, we need to make an assumption on the Hessian matrix, which takes the standard irrepresentable-like form. There exists a constant $\alpha \in [0,1)$ such that

$$|||\mathcal{C}\left(\mathcal{H}_{\mathcal{N}\mathcal{T}}(\mathcal{H}_{\mathcal{T}\mathcal{T}})^{-1}\right)|||_\infty \leqslant 1 - \alpha. \tag{18}$$

These condition extends the irrepresentable condition given in Ravikumar et al. (2011), which was needed for estimation of networks from single attribute observations. It is worth noting, that the condition given in Eq. (18) can be much weaker than the irrepresentable condition of Ravikumar et al. (2011) applied directly to the full Hessian matrix. This can be observed in simulations done in Section 5, where a chain network is not consistently estimated even with a large number of samples.

We will also need the following two quantities to specify the results

$$\kappa_{\Sigma^*} = |||\mathcal{C}(\Sigma^*)|||_\infty, \tag{19}$$

and

$$\kappa_{\mathcal{H}} = |||\mathcal{C}(\mathcal{H}_{\mathcal{T}\mathcal{T}}^{-1})|||_\infty. \tag{20}$$

Finally, the results are going to depend on the tail bounds for the elements of the matrix $\mathcal{C}(S - \Sigma^*)$. We will assume that there is a constant $v_* \in (0, \infty]$ and a function $f : \mathbb{N} \times (0, \infty) \mapsto (0, \infty)$ such that for any $(a, b) \in V \times V$

$$\mathrm{pr}\left(\mathcal{C}(S - \Sigma^*)_{ab} \geqslant \delta\right) \leqslant \frac{1}{f(n, \delta)} \qquad \delta \in (0, v_*^{-1}]. \tag{21}$$

The function $f(n, \delta)$ will be monotonically increasing in both n and δ. Therefore, we define the following two inverse functions

$$\overline{n}_f(\delta; r) = \arg\max\{n \ : \ f(n, \delta) \leqslant r\} \tag{22}$$

and

$$\overline{\delta}_f(r; n) = \arg\max\{\delta \ : \ f(n, \delta) \leqslant r\} \tag{23}$$

for $r \in [1, \infty)$.

With the notation introduced, we have the following result.

Theorem 7 *Assume that the irrepresentable condition in Eq.* (18) *is satisfied and that there exists a constant* $v_* \in (0, \infty]$ *and a function* $f(n, \delta)$ *so that Eq.* (21) *is satisfied for any* $(a, b) \in V \times V$. *Let*

$$\lambda = \frac{8}{\alpha}\overline{\delta}_f(n, p^\tau)$$

for some $\tau > 2$. *If*

$$n > \overline{n}_f\left(\frac{1}{\max(v_*, 6(1 + 8\alpha^{-1})s\max(\kappa_{\Sigma^*}\kappa_{\mathcal{H}}, \kappa_{\Sigma^*}^3\kappa_{\mathcal{H}}^2))}, p^\tau\right), \tag{24}$$

then

$$||\mathcal{C}(\widehat{\Omega} - \Omega)||_\infty \leqslant 2(1 + 8\alpha^{-1})\kappa_{\mathcal{H}}\overline{\delta}_f(n, p^\tau) \tag{25}$$

with probability at least $1 - p^{2-\tau}$.

Theorem 7 is of the same form as Theorem 1 in Ravikumar et al. (2011), but the ℓ_∞ element-wise convergence is established for $\mathcal{C}(\widehat{\Omega} - \Omega)$, which will guarantee successful recovery of non-zero partial canonical correlations if the blocks of the true precision matrix are sufficiently large.

Theorem 7 is proven as Theorem 1 in Ravikumar et al. (2011). We provide technical results in Lemma 8, Lemma 9 and Lemma 10, which can be used to substitute results of Lemma 4, Lemma 5 and Lemma 6 in Ravikumar et al. (2011) under our setting. The rest of the arguments then go through. Below we provide some more details.

First, let $\mathcal{Z} : \mathbb{R}^{pk\times pk} \mapsto \mathbb{R}^{pk\times pk}$ be the mapping defined as

$$\mathcal{Z}(A)_{ab} = \begin{cases} \frac{A_{ab}}{||A_{ab}||_F} & \text{if } ||A_{ab}||_F \neq 0, \\ Z \text{ with } ||Z||_F \leqslant 1 & \text{if } ||A_{ab}||_F = 0, \end{cases} \tag{26}$$

Next, define the function

$$G(\Omega) = \operatorname{tr}\Omega S - \log|\Omega| + \lambda||\mathcal{C}(\Omega)||_1, \quad \forall \Omega \succ 0 \tag{27}$$

and the following system of equations

$$\begin{cases} S_{ab} - (\Omega^{-1})_{ab} = -\lambda\mathcal{Z}(\Omega)_{ab}, & \text{if } \Omega_{ab} \neq 0 \\ ||S_{ab} - (\Omega^{-1})_{ab}||_F \leqslant \lambda, & \text{if } \Omega_{ab} = 0. \end{cases} \tag{28}$$

It is known that $\Omega \in \mathbb{R}^{\widetilde{p}\times\widetilde{p}}$ is the minimizer of optimization problem in Eq. (4) if and only if it satisfies the system of equations given in Eq. (28). We have already shown in Lemma 5 that the minimizer is unique.

Let $\widetilde{\Omega}$ be the solution to the following constrained optimization problem

$$\min_{\Omega\succ 0} \operatorname{tr} S\Omega - \log|\Omega| + \lambda||\mathcal{C}(\Omega)||_1 \text{ subject to } \mathcal{C}(\Omega)_{ab} = 0, \ \forall (a,b) \in \mathcal{N}. \tag{29}$$

Observe that one cannot find $\widetilde{\Omega}$ in practice, as it depends on the unknown set $\mathcal{N}$. However, it is a useful construction in the proof. We will prove that $\widetilde{\Omega}$ is solution to the optimization problem given in Eq. (4), that is, we will show that $\widetilde{\Omega}$ satisfies the system of equations (28).

Using the first-order Taylor expansion we have that

$$\widetilde{\Omega}^{-1} = (\Omega^*)^{-1} - (\Omega^*)^{-1}\Delta(\Omega^*)^{-1} + R(\Delta), \tag{30}$$

where $\Delta = \Omega - \Omega^*$ and $R(\Delta)$ denotes the remainder term. With this, we state and prove Lemma 8, Lemma 9 and Lemma 10. They can be combined as in Ravikumar et al. (2011) to complete the proof of Theorem 7.

Lemma 8 *Assume that*

$$\max_{ab}||\Delta_{ab}||_F \leqslant \frac{\alpha\lambda}{8} \quad \text{and} \quad \max_{ab}||\Sigma^*_{ab} - S_{ab}||_F \leqslant \frac{\alpha\lambda}{8}. \tag{31}$$

Then $\widetilde{\Omega}$ is the solution to the optimization problem in Eq. (4)*.*

Proof We use R to denote $R(\Delta)$. Recall that $\Delta_{\mathcal{N}} = 0$ by construction. Using (30) we can rewrite (28) as

$$\mathcal{H}_{ab,\mathcal{T}}\overline{\Delta}_{\mathcal{T}} - \overline{R}_{ab} + \overline{S}_{ab} - \overline{\Sigma}^*_{ab} + \lambda\overline{\mathcal{Z}}(\widetilde{\Omega})_{ab} = 0 \qquad \text{if } (a,b) \in \mathcal{T} \tag{32}$$

$$||\mathcal{H}_{ab,\mathcal{T}}\overline{\Delta}_{\mathcal{T}} - \overline{R}_{ab} + \overline{S}_{ab} - \overline{\Sigma}^*_{ab}||_2 \leqslant \lambda \qquad \text{if } (a,b) \in \mathcal{N}. \tag{33}$$

By construction, the solution $\widetilde{\Omega}$ satisfy (32). Under the assumptions, we show that (33) is also satisfied with inequality.

From (32), we can solve for $\Delta_{\mathcal{T}}$,

$$\Delta_{\mathcal{T}} = \mathcal{H}^{-1}_{\mathcal{T},\mathcal{T}}[\overline{R}_{\mathcal{T}} - \overline{\Sigma}_{\mathcal{T}} + \overline{S}_{\mathcal{T}} - \lambda\overline{\mathcal{Z}}(\widetilde{\Omega})_{\mathcal{T}}].$$

Then

$$\begin{aligned}
&||\mathcal{H}_{ab,\mathcal{T}}\mathcal{H}^{-1}_{\mathcal{T},\mathcal{T}}[\overline{R}_{\mathcal{T}} - \overline{\Sigma}_{\mathcal{T}} + \overline{S}_{\mathcal{T}} - \lambda\overline{\mathcal{Z}}(\widetilde{\Omega})_{\mathcal{T}}] - \overline{R}_{ab} + \overline{S}_{ab} - \overline{\Sigma}^*_{ab}||_2 \\
&\quad \leqslant \lambda||\mathcal{H}_{ab,\mathcal{T}}\mathcal{H}^{-1}_{\mathcal{T},\mathcal{T}}\overline{\mathcal{Z}}(\widetilde{\Omega})_{\mathcal{T}}||_2 + ||\mathcal{H}_{ab,\mathcal{T}}\mathcal{H}^{-1}_{\mathcal{T},\mathcal{T}}[\overline{R}_{\mathcal{T}} - \overline{\Sigma}_{\mathcal{T}} + \overline{S}_{\mathcal{T}}]||_2 + ||\overline{R}_{ab} + \overline{S}_{ab} - \overline{\Sigma}^*_{ab}||_2 \\
&\quad \leqslant \lambda(1-\alpha) + (2-\alpha)\frac{\alpha\lambda}{4} \\
&\quad < \lambda
\end{aligned}$$

using assumption on $\mathcal{H}$ in (18) and (31). This shows that $\widetilde{\Omega}$ satisfies (28). ■

Lemma 9 *Assume that*

$$||\mathcal{C}(\Delta)||_\infty \leqslant \frac{1}{3\kappa_{\Sigma^*}s}. \tag{34}$$

Then

$$||\mathcal{C}(R(\Delta))||_\infty \leqslant \frac{3s}{2}\kappa^3_{\Sigma^*}||\mathcal{C}(\Delta)||^2_\infty. \tag{35}$$

Proof Remainder term can be written as

$$R(\Delta) = (\Omega^* + \Delta)^{-1} - (\Omega^*)^{-1} + (\Omega^*)^{-1}\Delta(\Omega^*)^{-1}.$$

Using (40), we have that

$$\begin{aligned}
|||\mathcal{C}((\Omega^*)^{-1}\Delta)|||_\infty &\leqslant |||\mathcal{C}((\Omega^*)^{-1})|||_\infty|||\mathcal{C}(\Delta)|||_\infty \\
&\leqslant s|||\mathcal{C}((\Omega^*)^{-1})|||_\infty||\mathcal{C}(\Delta)||_\infty \\
&\leqslant \frac{1}{3},
\end{aligned}$$

which gives us the following expansion

$$(\Omega^* + \Delta)^{-1} = (\Omega^*)^{-1} - (\Omega^*)^{-1}\Delta(\Omega^*)^{-1} + (\Omega^*)^{-1}\Delta(\Omega^*)^{-1}\Delta J(\Omega^*)^{-1},$$

with $J = \sum_{k\geqslant 0}(-1)^k((\Omega^*)^{-1}\Delta)^k$. Using (41) and (40), we have that

$$\begin{aligned}
||\mathcal{C}(R)||_\infty &\leqslant ||\mathcal{C}((\Omega^*)^{-1}\Delta)||_\infty|||\mathcal{C}((\Omega^*)^{-1}\Delta J(\Omega^*)^{-1})^T|||_\infty \\
&\leqslant |||\mathcal{C}((\Omega^*)^{-1})|||^3_\infty||\mathcal{C}(\Delta)||_\infty|||\mathcal{C}(J^T)|||_\infty|||\mathcal{C}(\Delta)|||_\infty \\
&\leqslant s|||\mathcal{C}((\Omega^*)^{-1})|||^3_\infty||\mathcal{C}(\Delta)||^2_\infty|||\mathcal{C}(J^T)|||_\infty.
\end{aligned}$$

Next, we have that

$$\begin{aligned}\|\mathcal{C}(J^T)\|_\infty &\leqslant \sum_{k>0} \|\mathcal{C}(\Delta(\Omega^*)^{-1})\|_\infty^k \\ &\leqslant \frac{1}{1-\|\mathcal{C}(\Delta(\Omega^*)^{-1})\|_\infty} \\ &\leqslant \frac{3}{2},\end{aligned}$$

which gives us

$$||\mathcal{C}(R)||_\infty \leqslant \frac{3s}{2}\kappa_{\Sigma^*}^3 ||\mathcal{C}(\Delta)||_\infty^2$$

as claimed. ■

Lemma 10 *Assume that*

$$r := 2\kappa_{\mathcal{H}}(||\mathcal{C}(S-\Sigma^*)||_\infty + \lambda) \leqslant \min\left(\frac{1}{3\kappa_{\Sigma^*}s}, \frac{1}{3\kappa_{\mathcal{H}}\kappa_{\Sigma^*}^3 s}\right). \tag{36}$$

Then

$$||\mathcal{C}(\Delta)||_\infty \leqslant r. \tag{37}$$

Proof The proof follows the proof of Lemma 6 in Ravikumar et al. (2011). Define the ball

$$\mathcal{B}(r) := \{A \ : \ \mathcal{C}(A)_{ab} \leqslant r, \forall (a,b) \in \mathcal{T}\},$$

the gradient mapping

$$G(\Omega_{\mathcal{T}}) = -(\Omega^{-1})_{\mathcal{T}} + S_{\mathcal{T}} + \lambda \mathcal{Z}(\Omega)_{\mathcal{T}}$$

and

$$F(\overline{\Delta}_{\mathcal{T}}) = -\mathcal{H}_{\mathcal{T}\mathcal{T}}^{-1}\overline{G}(\Omega_{\mathcal{T}}^* + \Delta_{\mathcal{T}}) + \overline{\Delta}_{\mathcal{T}}.$$

We need to show that $F(\mathcal{B}(r)) \subseteq \mathcal{B}(r)$, which implies that $||\mathcal{C}(\Delta_{\mathcal{T}})||_\infty \leqslant r$.

Under the assumptions of the lemma, for any $\Delta_S \in \mathcal{B}(r)$, we have the following decomposition

$$F(\overline{\Delta}_{\mathcal{T}}) = \mathcal{H}_{\mathcal{T}\mathcal{T}}^{-1}\overline{R}(\Delta)_{\mathcal{T}} + \mathcal{H}_{\mathcal{T}\mathcal{T}}^{-1}(\overline{S}_{\mathcal{T}} - \overline{\Sigma}_{\mathcal{T}}^* + \lambda\overline{\mathcal{Z}}(\Omega^* + \Delta)_{\mathcal{T}}).$$

Using Lemma 9, the first term can be bounded as

$$\begin{aligned}||\mathcal{C}(\mathcal{H}_{\mathcal{T}\mathcal{T}}^{-1}\overline{R}(\Delta)_{\mathcal{T}})||_\infty &\leqslant \|\mathcal{C}(\mathcal{H}_{\mathcal{T}\mathcal{T}}^{-1})\|_\infty ||\mathcal{C}(R(\Delta)||_\infty \\ &\leqslant \frac{3s}{2}\kappa_{\mathcal{H}}\kappa_{\Sigma^*}^3 ||\mathcal{C}(\Delta)||_\infty^2 \\ &\leqslant \frac{3s}{2}\kappa_{\mathcal{H}}\kappa_{\Sigma^*}^3 r^2 \\ &\leqslant r/2\end{aligned}$$

where the last inequality follows under the assumptions. Similarly

$$\begin{aligned}&||\mathcal{C}(\mathcal{H}_{\mathcal{T}\mathcal{T}}^{-1}(\overline{S}_{\mathcal{T}} - \overline{\Sigma}_{\mathcal{T}}^* + \lambda\overline{\mathcal{Z}}(\Omega^* + \Delta)_{\mathcal{T}})||_\infty \\ &\quad\leqslant \|\mathcal{C}(\mathcal{H}_{\mathcal{T}\mathcal{T}}^{-1})\|_\infty(||\mathcal{C}(S-\Sigma^*)||_\infty + \lambda||\mathcal{C}(\mathcal{Z}(\Omega^* + \Delta))||_\infty) \\ &\quad\leqslant \kappa_{\mathcal{H}}(||\mathcal{C}(S-\overline{\Sigma}^*)||_\infty + \lambda) \\ &\quad\leqslant r/2.\end{aligned}$$

This shows that $F(\mathcal{B}(r)) \subseteq \mathcal{B}(r)$. ■

The following result is a corollary of Theorem 7, which shows that the graph structure can be estimated consistently under some assumptions.

Corollary 11 *Assume that the conditions of Theorem 7 are satisfied. Furthermore, suppose that*

$$\min_{(a,b)\in\mathcal{T},\ a\neq b} ||\Omega||_F > 2(1+8\alpha^{-1})\kappa_{\mathcal{H}}\overline{\delta}_f(n,p^\tau),$$

then Algorithm 1 estimates a graph $\widehat{G}$ which satisfies

$$\text{pr}\left(\widehat{G} \neq G\right) \geqslant 1 - p^{2-\tau}.$$

Next, we specialize the result of Theorem 7 to a case where X has sub-Gaussian tails. That is, the random vector $X = (X_1, \ldots, X_{pk})^T$ is zero-mean with covariance Σ^*. Each $(\sigma^*_{aa})^{-1/2}X_a$ is sub-Gaussian with parameter γ.

Proposition 12 *Set the penalty parameter in λ in Eq. (4) as*

$$\lambda = 8k\alpha^{-1}\left(128(1+4\gamma^2)^2(\max_a(\sigma^*_{aa})^2)n^{-1}(2\log(2k)+\tau\log(p))\right)^{1/2}.$$

If

$$n > C_1 s^2k^2(1+8\alpha^{-1})^2(\tau\log p + \log 4 + 2\log k),$$

where $C_1 = (48\sqrt{2}(1+4\gamma^2)(\max_a \sigma^*_{aa})\max(\kappa_{\Sigma^*}\kappa_{\mathcal{H}}, \kappa^3_{\Sigma^*}\kappa^2_{\mathcal{H}}))^2$, then

$$||\mathcal{C}(\widehat{\Omega}-\Omega)||_\infty \leqslant 16\sqrt{2}(1+4\gamma^2)\max_i \sigma^*_{ii}(1+8\alpha^{-1})\kappa_{\mathcal{H}}k\left(\frac{\tau\log p+\log 4+2\log k}{n}\right)^{1/2}$$

with probability $1-p^{2-\tau}$.

The proof simply follows by observing that, for any (a,b),

$$\begin{aligned}\text{pr}\left(\mathcal{C}(S-\Sigma^*)_{ab} > \delta\right) &\leqslant \text{pr}\left(\max_{(c,d)\in(a,b)}(\sigma_{cd}-\sigma^*_{cd})^2 > \delta^2/k^2\right)\\ &\leqslant k^2\text{pr}\left(|\sigma_{cd}-\sigma^*_{cd}| > \delta/k\right)\\ &\leqslant 4k^2\exp\left(-\frac{n\delta^2}{c_*k^2}\right)\end{aligned} \tag{38}$$

for all $\delta \in (0, 8(1+4\gamma^2)(\max_a \sigma^*_{aa}))$ with $c_* = 128(1+4\gamma^2)^2(\max_a(\sigma^*_{aa})^2)$. Therefore,

$$\begin{aligned}f(n,\delta) &= \frac{1}{4k^2}\exp(c_*\frac{n\delta^2}{k^2}),\\ \overline{n}_f(\delta;r) &= \frac{k^2\log(4k^2r)}{c_*\delta^2},\\ \overline{\delta}_f(r;n) &= \left(\frac{k^2\log(4k^2r)}{c_*n}\right)^{1/2}.\end{aligned}$$

Theorem 7 and some simple algebra complete the proof.

Proposition 4 is a simple consequence of Proposition 12.

B.5 Some Results on Norms of Block Matrices

Let $\mathcal{T}$ be a partition of V. Throughout this section, we assume that matrices $A, B \in \mathbb{R}^{p\times p}$ and a vector $b \in \mathbb{R}^p$ are partitioned into blocks according to $\mathcal{T}$.

Lemma 13

$$\max_{a\in\mathcal{T}} ||A_{a\cdot}b||_2 \leqslant \max_{a\in\mathcal{T}} \sum_{b\in\mathcal{T}} ||A_{ab}||_F \max_{c\in\mathcal{T}} ||b_c||_2. \tag{39}$$

Proof For any $a \in \mathcal{T}$,

$$\begin{aligned}
||A_{a\cdot}b||_2 &\leqslant \sum_{b\in\mathcal{T}} ||A_{ab}b_b||_2 \\
&= \sum_{b\in\mathcal{T}} \left(\sum_{i\in a}(A_{ib}b_b)^2\right)^{1/2} \\
&\leqslant \sum_{b\in\mathcal{T}} \left(\sum_{i\in a}||A_{ib}||_2^2||b_b||_2^2\right)^{1/2} \\
&\leqslant \sum_{b\in\mathcal{T}} \left(\sum_{i\in a}||A_{ib}||_2^2\right)^{1/2} \max_{c\in\mathcal{T}} ||b_c||_2 \\
&= \sum_{b\in\mathcal{T}} ||A_{ab}||_F \max_{c\in\mathcal{T}} ||b_c||_2.
\end{aligned}$$

■

Lemma 14

$$|||\mathcal{C}(AB)|||_\infty \leqslant |||\mathcal{C}(B)|||_\infty |||\mathcal{C}(A)|||_\infty. \tag{40}$$

Proof Let $\mathbf{C} = AB$ and let $\mathcal{T}$ be a partition of V.

$$\begin{aligned}
|||\mathcal{C}(AB)|||_\infty &= \max_{a\in\mathcal{T}} \sum_{b\in\mathcal{T}} ||\mathbf{C}_{ab}||_F \\
&\leqslant \max_{a\in\mathcal{T}} \sum_b \sum_c ||A_{ac}||_F ||B_{cb}||_F \\
&\leqslant \{\max_{a\in\mathcal{T}} \sum_c ||A_{ac}||_F\}\{\max_{c\in\mathcal{T}} \sum_b ||B_{cb}||_F\} \\
&= |||\mathcal{C}(A)|||_\infty |||\mathcal{C}(B)|||_\infty.
\end{aligned}$$

■

Lemma 15

$$||\mathcal{C}(AB)||_\infty \leqslant ||\mathcal{C}(A)||_\infty |||\mathcal{C}(B)^T|||_\infty. \tag{41}$$

Proof For a fixed a and b,

$$\begin{aligned}\mathcal{C}(AB)_{ab} &= ||\sum_c A_{ac}B_{cb}||_F \\ &\leqslant \sum_c ||A_{ac}||_F||B_c b||_F \\ &\leqslant \max_c ||A_{ac}|| \sum_c ||B_{cb}||_F.\end{aligned}$$

Maximizing over a and b gives the result. ■

Appendix C. Additional Information About Functional Brain Networks

Table 3 contains list of the names of the brain regions. The number before each region is used to index the node in the connectivity models. Figures 11, 12 and 13 contain adjacency matrices for the estimated graph structures.

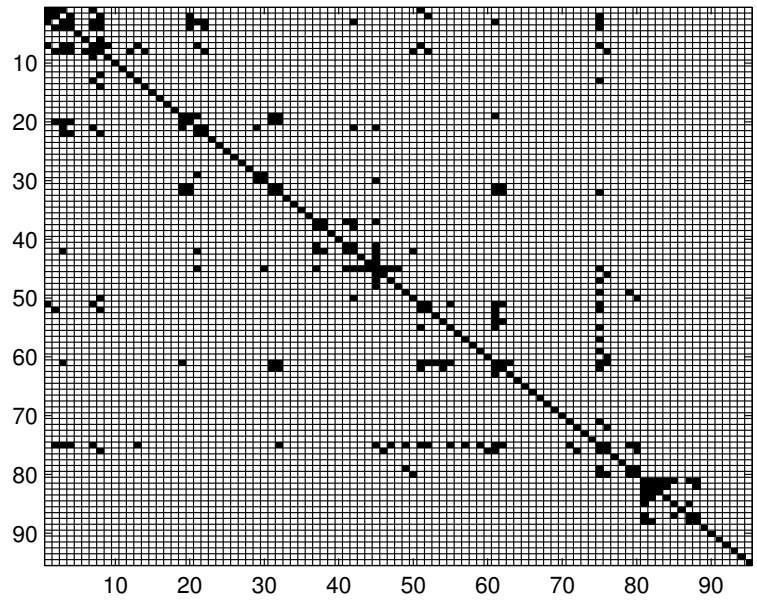


Figure 11: Adjacency matrix for the brain connectivity network: healthy subjects

References

H. Akaike. A new look at the statistical model identification. *IEEE Trans. Automat. Contr.*, 19(6):716–723, Dec 1974.

R. Albert and A.-L. Barabási. Statistical mechanics of complex networks. *Rev. Mod. Phys.*, 74:47–97, Jan 2002.

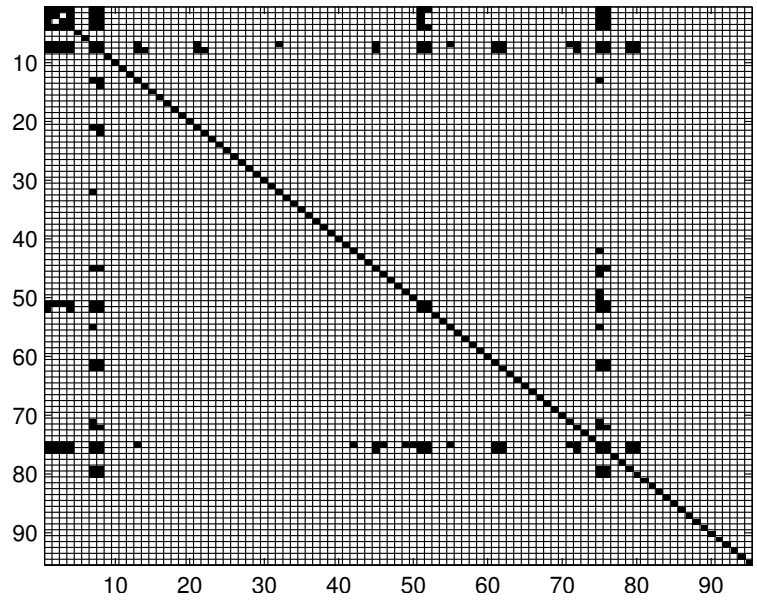


Figure 12: Adjacency matrix for the brain connectivity network: Mild Cognitive Impairment

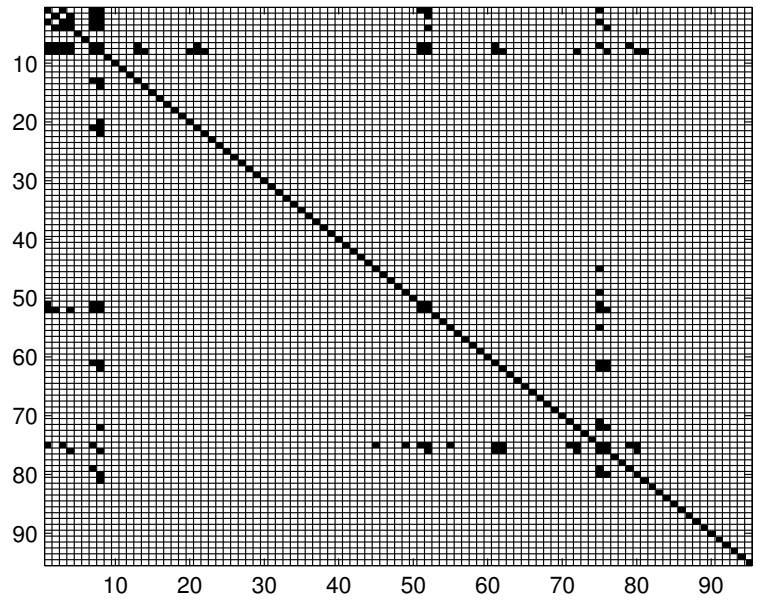


Figure 13: Adjacency matrix for the brain connectivity network: Alzheimer's & Dementia

1	Precentral_L	49	Fusiform_L
2	Precentral_R	50	Fusiform_R
3	Frontal_Sup_L	51	Postcentral_L
4	Frontal_Sup_R	52	Postcentral_R
5	Frontal_Sup_Orb_L	53	Parietal_Sup_L
6	Frontal_Sup_Orb_R	54	Parietal_Sup_R
7	Frontal_Mid_L	55	Parietal_Inf_L
8	Frontal_Mid_R	56	Parietal_Inf_R
9	Frontal_Mid_Orb_L	57	SupraMarginal_L
10	Frontal_Mid_Orb_R	58	SupraMarginal_R
11	Frontal_Inf_Oper_L	59	Angular_L
12	Frontal_Inf_Oper_R	60	Angular_R
13	Frontal_Inf_Tri_L	61	Precuneus_L
14	Frontal_Inf_Tri_R	62	Precuneus_R
15	Frontal_Inf_Orb_L	63	Paracentral_Lobule_L
16	Frontal_Inf_Orb_R	64	Paracentral_Lobule_R
17	Rolandic_Oper_L	65	Caudate_L
18	Rolandic_Oper_R	66	Caudate_R
19	Supp_Motor_Area_L	67	Putamen_L
20	Supp_Motor_Area_R	68	Putamen_R
21	Frontal_Sup_Medial_L	69	Thalamus_L
22	Frontal_Sup_Medial_R	70	Thalamus_R
23	Frontal_Med_Orb_L	71	Temporal_Sup_L
24	Frontal_Med_Orb_R	72	Temporal_Sup_R
25	Rectus_L	73	Temporal_Pole_Sup_L
26	Rectus_R	74	Temporal_Pole_Sup_R
27	Insula_L	75	Temporal_Mid_L
28	Insula_R	76	Temporal_Mid_R
29	Cingulum_Ant_L	77	Temporal_Pole_Mid_L
30	Cingulum_Ant_R	78	Temporal_Pole_Mid_R
31	Cingulum_Mid_L	79	Temporal_Inf_L
32	Cingulum_Mid_R	80	Temporal_Inf_R
33	Hippocampus_L	81	Cerebelum_Crus1_L
34	Hippocampus_R	82	Cerebelum_Crus1_R
35	ParaHippocampal_L	83	Cerebelum_Crus2_L
36	ParaHippocampal_R	84	Cerebelum_Crus2_R
37	Calcarine_L	85	Cerebelum_4_5_L
38	Calcarine_R	86	Cerebelum_4_5_R
39	Cuneus_L	87	Cerebelum_6_L
40	Cuneus_R	88	Cerebelum_6_R
41	Lingual_L	89	Cerebelum_7b_L
42	Lingual_R	90	Cerebelum_7b_R
43	Occipital_Sup_L	91	Cerebelum_8_L
44	Occipital_Sup_R	92	Cerebelum_8_R
45	Occipital_Mid_L	93	Cerebelum_9_L
46	Occipital_Mid_R	94	Cerebelum_9_R
47	Occipital_Inf_L	95	Vermis_4_5
48	Occipital_Inf_R		

Table 3: Names of the brain regions. L means that the brain region is located at the left hemisphere; R means right hemisphere.

J. R. Andrews-Hanna, A. Z. Snyder, J. L. Vincent, C. Lustig, D. Head, M. E. Raichle, and R. L. Buckner. Disruption of large-scale brain systems in advanced aging. *Neuron*, 56(5): 924–935, 2007.

O. Banerjee, L. El Ghaoui, and A. d'Aspremont. Model selection through sparse maximum likelihood estimation. *J. Mach. Learn. Res.*, 9(3):485–516, 2008.

A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286 (5439):509–512, 1999.

A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imag. Sci.*, 2:183–202, 2009.

T. T. Cai, W. Liu, and X. Luo. A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *J. Am. Stat. Assoc.*, 106(494):594–607, 2011.

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

P. Danaher, P. Wang, and D. M. Witten. The joint graphical lasso for inverse covariance estimation across multiple classes. *J. R. Stat. Soc. B*, 76(2):373–397, 2014.

A. P. Dempster. Covariance selection. *Biometrics*, 28:157–175, 1972.

J. H. Friedman, T. J. Hastie, and R. J. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

R. L. Gould, B. Arroyo, R. G. Brown, A. M. Owen, E. T. Bullmore, and R. J. Howard. Brain mechanisms of successful compensation during learning in alzheimer disease. *Neurology*, 67(6):1011–1017, 2006.

M. D. Greicius, G. Srivastava, A. L. Reiss, and V. Menon. Default-mode network activity distinguishes alzheimer's disease from healthy aging: Evidence from functional MRI. *Proc. Natl. Acad. Sci. U.S.A.*, 101(13):4637–4642, 2004.

J. Guo, E. Levina, G. Michailidis, and J. Zhu. Joint estimation of multiple graphical models. *Biometrika*, 98(1):1–15, 2011.

T. Hedden, K. R. A. V. Dijk, J. A. Becker, A. Mehta, R. A. Sperling, K. A. Johnson, and R. L. Buckner. Disruption of functional connectivity in clinically normal older adults harboring amyloid burden. *J. Neurosci.*, 29(40):12686–12694, 2009.

J. Honorio and D. Samaras. Multi-task learning of gaussian graphical models. In J. Fürnkranz and T. Joachims, editors, *Proc. of ICML*, pages 447–454, Haifa, Israel, June 2010. Omnipress.

S. Huang, J. Li, L. Sun, J. Liu, T. Wu, K. Chen, A. Fleisher, E. Reiman, and J. Ye. Learning brain connectivity of alzheimer's disease from neuroimaging data. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Proc. of NIPS*, pages 808–816, 2009.

C. Johnson, A. Jalali, and P. Ravikumar. High-dimensional sparse inverse covariance estimation using greedy methods. In N. Lawrence and M. Girolami, editors, *Proc. of AISTATS*, pages 574–582, 2012.

N. Katenka and E. D. Kolaczyk. Multi-attribute networks and the impact of partial information on inference and characterization. *Ann. Appl. Stat.*, 6(3):1068–1094, 2011.

M. Kolar and E. P. Xing. Consistent covariance selection from data with missing values. In J. Langford and J. Pineau, editors, *Proc. of ICML*, pages 551–558, Edinburgh, Scotland, GB, July 2012. Omnipress.

M. Kolar, L. Song, A. Ahmed, and E. P. Xing. Estimating Time-varying networks. *Ann. Appl. Stat.*, 4(1):94–123, 2010.

C. Lam and J. Fan. Sparsistency and rates of convergence in large covariance matrix estimation. *Ann. Stat.*, 37:4254–4278, 2009.

S. L. Lauritzen. *Graphical Models*, volume 17 of *Oxford Statistical Science Series*. The Clarendon Press Oxford University Press, New York, 1996. Oxford Science Publications.

H. Li and J. Gui. Gradient directed regularization for sparse gaussian concentration graphs, with applications to inference of genetic networks. *Biostatistics*, 7(2):302–317, 2006.

H. Liu, J. D. Lafferty, and L. A. Wasserman. The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *J. Mach. Learn. Res.*, 10:2295–2328, 2009.

H. Liu, F. Han, and C.-H. Zhang. Transelliptical graphical models. In P. Bartlett, F. Pereira, C. Burges, L. Bottou, and K. Weinberger, editors, *Proc. of NIPS*, pages 809–817. 2012.

K. Lounici, M. Pontil, A. B. Tsybakov, and S. A. van de Geer. Oracle inequalities and optimal inference under group sparsity. *Ann. Stat.*, 39:2164–204, 2011.

R. Mazumder and D. K. Agarwal. A flexible, scalable and efficient algorithmic framework for primal graphical lasso. Technical report, Stanford University, 2011.

R. Mazumder and T. J. Hastie. Exact covariance thresholding into connected components for large-scale graphical lasso. *J. Mach. Learn. Res.*, 13:781–794, 2012.

N. Meinshausen and P. Bühlmann. High dimensional graphs and variable selection with the lasso. *Ann. Stat.*, 34(3):1436–1462, 2006.

N. Meinshausen and P. Bühlmann. Stability selection. *J. R. Stat. Soc. B*, 72(4):417–473, 2010.

J. Peng, P. Wang, N. Zhou, and J. Zhu. Partial correlation estimation by joint sparse regression models. *J. Am. Stat. Assoc.*, 104(486):735–746, 2009.

M. Pourahmadi. Covariance estimation: The glm and regularization perspectives. *Stat. Sci.*, 26(3):369–387, 08 2011.

B. Rao. Partial canonical correlations. *Trabajos de Estadstica y de Investigacin Operativa*, 20(2):211–219, 1969.

P. Ravikumar, M. J. Wainwright, G. Raskutti, and B. Yu. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electron. J. Stat.*, 5: 935–980, 2011.

G. Schwarz. Estimating the dimension of a model. *Ann. Stat.*, 6(2):461–464, 03 1978.

X. Shen, W. Pan, and Y. Zhu. Likelihood-based selection and sharp parameter estimation. *J. Am. Stat. Assoc.*, 107:223–232, 2012.

M. Sjöbeck and E. Englund. Alzheimer's disease and the cerebellum: A morphologic study on neuronal and glial changes. *Dementia and Geriatric Cognitive Disorders*, 12(3):211–218, 2001.

Y. Stern. Cognitive reserve and Alzheimer disease. *Alzheimer Disease & Associated Disorders*, 20(2):112–117, 2006.

P. Tseng. Convergence of a block coordinate descent method for nondifferentiable minimization. *J. Optim. Theory Appl.*, 109(3):475–494, 2001.

G. Varoquaux, A. Gramfort, J.-B. Poline, and B. Thirion. Brain covariance selection: Better individual functional connectivity models using population prior. In J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, editors, *Proc. of NIPS*, pages 2334–2342, 2010.

D. M. Witten, J. H. Friedman, and N. Simon. New insights and faster computations for the graphical lasso. *J. Comput. Graph. Stat.*, 20(4):892–900, 2011.

X. Wu, R. Li, A. S. Fleisher, E. M. Reiman, X. Guan, Y. Zhang, K. Chen, and L. Yao. Altered default mode network connectivity in Alzheimer's diseasea resting functional MRI and Bayesian network study. *Human Brain Mapping*, 32(11):1868–1881, 2011.

M. Yuan. High dimensional inverse covariance matrix estimation via linear programming. *J. Mach. Learn. Res.*, 11:2261–2286, 2010.

M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. B*, 68:49–67, 2006.

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

T. Zhao, H. Liu, K. E. Roeder, J. D. Lafferty, and L. A. Wasserman. The huge package for high-dimensional undirected graph estimation in R. *J. Mach. Learn. Res.*, 13:1059–1062, 2012.