Markov Network Estimation From Multi-attribute Data

Mladen Kolar MLADENK@CS.CMU.EDU
Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA 15217 USA

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

Eric P. Xing EPXING@CS.CMU.EDU
Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA 15217 USA

Abstract

Many real world network problems often concern multivariate nodal attributes such as image, textual, and multi-view feature vectors on nodes, rather than simple univariate nodal attributes. The existing graph estimation methods built on Gaussian graphical models and covariance selection algorithms can not handle such data, neither can the theories developed around such methods be directly applied. In this paper, we propose a new principled framework for estimating multi-attribute graphs. Instead of estimating the partial correlation as in current literature, our method estimates the *partial canonical correlations* that naturally accommodate complex nodal features. Computationally, we provide an efficient algorithm which utilizes the multi-attribute structure. Theoretically, we provide sufficient conditions which guarantee consistent graph recovery. Extensive simulation studies demonstrate performance of our method under various conditions.

1. Introduction

In many modern problems, we are interested in studying a network of entities with multiple attributes rather than a simple univariate attribute. For example, when an entity represents a person in a social network, it is widely accepted that the nodal attribute is most naturally a vector with many personal information including demographics, interests, and other features, rather than merely a single attribute, such as a binary vote as assumed in the current literature of social graph estimation based on Markov random fields (Banerjee et al., 2008; Kolar et al., 2010). In another example, when an entity represents a gene in a gene regulation network, modern data acquisition technologies allow researchers 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, rather than merely a single attribute such as an expression level, which is assumed in the current literature on gene graph estimation based on Gaussian graphical models (Peng et al., 2009). Indeed, it is somewhat surprising that existing research on graph estimation remains largely blinded to the analysis of multi-attribute data that are prevalent and widely studied in the network community. Existing algorithms and theoretical analysis relies heavily on covariance selection using graphical lasso, or penalized pseudo-likelihood. They can not be easily extended to graphs with multi-variate nodal attributes.

In this paper, we present a study on graph estimation from multi-attribute data, in an attempt to fill the gap between the practical needs and existing methodologies from the literature. Under a Gaussian graphical model, one assumes that a p-dimensional random vector $\mathbf{X} \in \mathbb{R}^p$ follows a multivariate Gaussian distribution with the mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, with each component of the vector corresponding to a node in the graph. Based on n independent and identically distributed observations, one can estimate an undirected graph $G = (V, E)$, where the node set V corresponds to the p variables, and the edge set E describes the conditional independence relationships among the variables, that is, variables X_a and X_b are conditionally independent given all the remaining vari-

Proceedings of the 30^{th} International Conference on Machine Learning, Atlanta, Georgia, USA, 2013. JMLR: W&CP volume 28.

ables if $(a, b) \notin E$. Given multi-attribute data, this approach is clearly invalid, because it naively translates to estimating one graph per attribute. A subsequent integration of all such graphs to a summary graph on the entire dataset may lead to unclear statistical interpretation.

We consider the following new setting for estimating a multi-attribute graph. Assume now a "stacked" long random vector $\mathbf{X} = (\mathbf{X}_1', ..., \mathbf{X}_p')'$ where $\mathbf{X}_1 \in \mathbb{R}^{k_1}, \ldots, \mathbf{X}_p \in \mathbb{R}^{k_p}$ are themselves random vectors that jointly follow the multivariate Normal distribution,

$$\left(\mathbf{X}_1', \ldots, \mathbf{X}_p'\right)' \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}^*) \tag{1}$$

where

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \vdots \\ \boldsymbol{\mu}_p \end{pmatrix} \text{ and } \boldsymbol{\Sigma}^* = \begin{pmatrix} \boldsymbol{\Sigma}^*_{11} & \boldsymbol{\Sigma}^*_{12} & \cdots & \boldsymbol{\Sigma}^*_{1p} \\ \boldsymbol{\Sigma}^*_{21} & \boldsymbol{\Sigma}^*_{22} & \cdots & \boldsymbol{\Sigma}^*_{2p} \\ \vdots & & \ddots & \vdots \\ \boldsymbol{\Sigma}^*_{p1} & \cdots & & \boldsymbol{\Sigma}^*_{pp} \end{pmatrix}.$$

Without loss of generality, we assume $\mu = 0$. Let $G = (V, E)$ be a graph with the vertex set $V = [p]$ ($[p]$ represents the set $\{1, \ldots, p\}$) and the set of edges $E \subseteq V \times V$ that encodes conditional independence relationships among $(\mathbf{X}_a)_{a \in V}$. That is, each node $a \in V$ of the graph G corresponds to the random vector $\mathbf{X}_a$ and there is no edge between nodes a and b in the graph if and only if $\mathbf{X}_a$ is conditionally independent of $\mathbf{X}_b$ given all the vectors corresponding to the remaining nodes, $\mathbf{X}_{\neg ab} = \{\mathbf{X}_c \ : \ c \in [p] \backslash \{a, b\}\}$. Such a graph is also known as a *Markov network* (of Markov graph), which we shall emphasize in this paper to contract 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, as the block corresponding to $\mathbf{X}_a$ and $\mathbf{X}_b$ will be equal to zero. Let $\mathcal{D}_n = \{\mathbf{x}_i\}_{i \in [n]}$ be a sample of n independent and identically distributed vectors drawn from $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$. For a vector $\mathbf{x}_i$, we denote $\mathbf{x}_{i,a} \in \mathbb{R}^{k_a}$ 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 may be useful in certain applications, e.g., when a node represents a gene pathway in a regulatory network.

Using the standard Gaussian graphical model for univariate nodal observations, one can estimate a Markov graph for each attribute individually, by estimating the sparsity pattern of the precision matrix $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$ of the GMM. This is also known as *covariance selection* (Dempster, 1972). For high dimensional problems, Meinshausen & 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 & Lin (2007), and Friedman et al. (2008) take a penalized likelihood approach to estimate the sparse precision matrix $\boldsymbol{\Omega}$. To reduce estimation bias, Lam & 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 have better theoretical properties than the penalized likelihood approach. Under certain regularity conditions, these methods have proven to estimate graph structure consistently (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). However, in the case of multi-attribute data, it is not clear how to combine estimated graphs to obtain a single Markov network reflecting the structure of the underlying complex system. This is especially the case when nodes in the graph contain different number of attributes.

Katenka & Kolaczyk (2011) proposed a method for estimating association networks from multi-attribute data using canonical correlation as a dependence measure between two groups of attributes. However, association networks are known to confound the direct interactions with indirect ones as they only represent marginal associations. In contrast, we develop a method based on *partial canonical correlation*, which give rise to a Markov network that is 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 & Samaras, 2010; Chiquet et al., 2011; Danaher et al., 2011), however, the model given in (1) is different from models considered in various multi-task settings and the optimization algorithms developed to handle the multi-task setting do not extend to handle the optimization problem given in (3) below.

Unlike the standard procedures for learning the structure of GGMs (e.g., neighborhood selection (Meinshausen & Bühlmann, 2006) or glasso (Friedman et al., 2008)), which infer the partial correlations between pairs of nodes, our proposed method estimates the *partial canonical correlations* between pairs of nodes. Under this new framework, the contri-

butions of this paper include: (i) computationally, an efficient algorithm is provided to estimate the multi-attribue graphs; (ii) theoretically, we provide sufficient conditions which guarantee consistent graph recovery; and (iii) empirically, a number of simulations are used to illustrate performance of the method. Additional results can be found in Kolar et al. (2012).

2. Preliminaries and Related Work

Canonical correlation, a classical tool in multivariate statistics, is defined between two multivariate random variables as

$$\rho_c(\mathbf{X}_a, \mathbf{X}_b) = \max_{\mathbf{u}\in\mathbb{R}^{k_a}, \mathbf{v}\in\mathbb{R}^{k_b}} \mathrm{Corr}(\mathbf{u}'\mathbf{X}_a, \mathbf{v}'\mathbf{X}_b),$$

that is, computing canonical correlation between $\mathbf{X}_a$ and $\mathbf{X}_b$ is equivalent to maximization of correlation between two linear combinations $\mathbf{u}'\mathbf{X}_a$ and $\mathbf{v}'\mathbf{X}_b$ with respect to vectors $\mathbf{u}$ and $\mathbf{v}$. Canonical correlation can be used to measure association strength between two nodes with multi-attribute observations. For example, in Katenka & Kolaczyk (2011), a graph is estimated from multi-attribute nodal observations by thresholding the canonical correlation between nodes, which may confound the direct interactions with indirect ones, as we describe later.

In this work, we will use 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 $\widehat{\mathbf{A}} = \mathrm{argmin}\ \mathbb{E}[||\mathbf{X}_a - \mathbf{A}\mathbf{X}_{\overline{ab}}||_2^2]$ and $\widehat{\mathbf{B}} = \mathrm{argmin}\ \mathbb{E}[||\mathbf{X}_b - \mathbf{B}\mathbf{X}_{\overline{ab}}||_2^2]$, then the partial canonical correlation between $\mathbf{X}_a$ and $\mathbf{X}_b$ is defined as

$$\rho_c(\mathbf{X}_a, \mathbf{X}_b; \mathbf{X}_{\overline{ab}}) = \max_{\mathbf{u}\in\mathbb{R}^{k_a}, \mathbf{v}\in\mathbb{R}^{k_b}} \mathrm{Corr}(\mathbf{u}'(\mathbf{X}_a - \widehat{\mathbf{A}}\mathbf{X}_{\overline{ab}}), \mathbf{v}'(\mathbf{X}_b - \widehat{\mathbf{B}}\mathbf{X}_{\overline{ab}})),$$

that is, the partial canonical correlation between $\mathbf{X}_a$ and $\mathbf{X}_b$ is equal to the canonical correlation between residual vectors of $\mathbf{X}_a$ and $\mathbf{X}_b$ after the effect of variables $\mathbf{X}_{\neg ab}$ is removed (Rao, 1969).

Let $\mathbf{\Omega}^*$ denote the precision matrix under the model in (1). Using standard results for the multivariate Normal distribution (see also equation (7) in Rao (1969)), a straight forward calculation shows that

$$\rho_c(X_a, X_b; X_{\neg ab}) \neq 0 \iff \max_{\mathbf{u}\in\mathbb{R}^{k_a}, \mathbf{v}\in\mathbb{R}^{k_b}} \mathbf{u}^T\Omega^*_{ab}\mathbf{v} \neq 0. \tag{2}$$

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, under model in (1), vectors X_a and X_b are conditionally independent given $X_{\neg ab}$ if and only if the partial canonical correlation is zero. A network 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 network even though the network is built based on pairwise conditional (in)dependence properties. In Section 3, we use the above observations to provide 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 (2), we can motivate an alternative method for estimating the non-zero partial canonical correlation. Let $\overline{a} = \{b \ : \ b \in [p]\backslash\{a\}\}$ denote the set of all nodes minus the node a. Then

$$\mathbb{E}[\mathbf{X}_a \mid \mathbf{X}_{\overline{a}} = \mathbf{x}_{\overline{a}}] = \mathbf{\Sigma}^*_{a,\overline{a}}\mathbf{\Sigma}^{*,-1}_{\overline{a},\overline{a}}\mathbf{x}_{\overline{a}}.$$

Since $\mathbf{\Omega}^*_{a,\overline{a}} = -(\mathbf{\Sigma}^*_{aa} - \mathbf{\Sigma}^*_{a,\overline{a}}\mathbf{\Sigma}^{*,-1}_{\overline{a},\overline{a}}\mathbf{\Sigma}^*_{\overline{a},a})^{-1}\mathbf{\Sigma}^*_{a,\overline{a}}\mathbf{\Sigma}^{*,-1}_{\overline{a},\overline{a}}$, we observe that a zero block $\mathbf{\Omega}_{ab}$ can be identified from the regression coefficients when each component of $\mathbf{X}_a$ is regressed on $\mathbf{X}_{\overline{a}}$. 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 & Kolaczyk (2011).

3. Estimation Procedure

3.1. Penalized Log-Likelihood Optimization

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

$$\min_{\mathbf{\Omega}\succ\mathbf{0}} \ \mathrm{tr}\,\mathbf{S}\mathbf{\Omega} - \log|\mathbf{\Omega}| + \lambda\sum_{a,b}||\mathbf{\Omega}_{ab}||_F \tag{3}$$

where $\mathbf{S} = n^{-1}\sum_{i\in[n]}\mathbf{x}_i\mathbf{x}_i^T$ is the sample covariance matrix and $||\mathbf{\Omega}_{ab}||_F$ denotes the Frobenius norm of $\mathbf{\Omega}_{ab}$. The Frobenius norm penalty 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 & Lin, 2006). Here we assume that the same number of samples is available per attribute. However, the same procedure 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 $\mathbf{S}$ from available samples, treating non-measured attributes as missing completely at random (see Kolar & Xing, 2012, for more details). The dual

problem to (3) is

$$\max_{\mathbf{\Sigma}} \sum_{j\in[p]} k_j + \log|\mathbf{\Sigma}| \quad \text{s.t.} \quad \max_{a,b} ||\mathbf{S}_{ab} - \mathbf{\Sigma}_{ab}||_F \leq \lambda. \tag{4}$$

where $\mathbf{\Sigma}$ is the dual variable to $\mathbf{\Omega}$. 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 estimate simultaneously the precision matrix and covariance matrix, without explicitly performing an expensive matrix inversion.

We propose to optimize the objective (3) using a block coordinate descent procedure, inspired by Mazumder & 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. Write

$$\mathbf{\Omega} = \begin{pmatrix} \mathbf{\Omega}_{aa} & \mathbf{\Omega}_{a,\bar{a}} \\ \mathbf{\Omega}_{\bar{a},a} & \mathbf{\Omega}_{\bar{a},\bar{a}} \end{pmatrix}, \quad \mathbf{S} = \begin{pmatrix} \mathbf{S}_{aa} & \mathbf{S}_{a,\bar{a}} \\ \mathbf{S}_{\bar{a},a} & \mathbf{S}_{\bar{a},\bar{a}} \end{pmatrix}$$

and suppose that $(\widetilde{\mathbf{\Omega}}, \widetilde{\mathbf{\Sigma}})$ are current iterates. With the block partition above, we have $\log|\mathbf{\Omega}| = \log(\mathbf{\Omega}_{\bar{a},\bar{a}}) + \log(\mathbf{\Omega}_{aa} - \mathbf{\Omega}_{a,\bar{a}}(\mathbf{\Omega}_{\bar{a},\bar{a}})^{-1}\mathbf{\Omega}_{\bar{a},a})$. The next iterate $\widehat{\mathbf{\Omega}}$ is of the form

$$\widehat{\mathbf{\Omega}} = \widetilde{\mathbf{\Omega}} + \begin{pmatrix} \mathbf{\Delta}_{aa} & \mathbf{\Delta}_{a,\bar{a}} \\ \mathbf{\Delta}_{\bar{a},a} & \mathbf{0} \end{pmatrix} = \begin{pmatrix} \widehat{\mathbf{\Omega}}_{aa} & \widehat{\mathbf{\Omega}}_{a,\bar{a}} \\ \widehat{\mathbf{\Omega}}_{\bar{a},a} & \widehat{\mathbf{\Omega}}_{\bar{a},\bar{a}} \end{pmatrix}$$

and is obtained by minimizing

$$\begin{aligned} &\operatorname{tr}\mathbf{S}_{aa}\mathbf{\Omega}_{aa} + 2\operatorname{tr}\mathbf{S}_{a,\bar{a}}\mathbf{\Omega}_{\bar{a},a} \\ &\quad - \log|\mathbf{\Omega}_{aa} - \mathbf{\Omega}_{a,\bar{a}}(\widetilde{\mathbf{\Omega}}_{\bar{a},\bar{a}})^{-1}\mathbf{\Omega}_{\bar{a},a}| \\ &\quad + \lambda||\mathbf{\Omega}_{aa}||_F + 2\lambda\sum_{b\neq a}||\mathbf{\Omega}_{ab}||_F. \end{aligned} \tag{5}$$

Complete minimization over the variables $\mathbf{\Omega}_{aa}$ and $\mathbf{\Omega}_{a,\bar{a}}$ at each iteration of the block coordinate descent can be computationally expensive. Therefore, we propose to update $\mathbf{\Omega}_{aa}$ and $\mathbf{\Omega}_{a,\bar{a}}$ using one generalized gradient step update (see Beck & Teboulle (2009)) in each iteration. Note that the objective in (5) is a sum of a smooth convex function and a non-smooth convex penalty, so that the gradient descent cannot be applied. Given a step size t, generalized gradient descent optimizes a quadratic approximation of the objective at the current iterate $\widetilde{\mathbf{\Omega}}$, which results in the following two updates

$$\widehat{\mathbf{\Omega}}_{aa} = \varphi_{t,\lambda}\left(\widetilde{\mathbf{\Omega}}_{aa} + t(\widetilde{\mathbf{\Sigma}}_{aa} - \mathbf{S}_{aa})\right) \tag{6}$$

and

$$\widehat{\mathbf{\Omega}}_{ab} = \varphi_{t,\lambda}\left(\widetilde{\mathbf{\Omega}}_{ab} + t(\widetilde{\mathbf{\Sigma}}_{ab} - \mathbf{S}_{ab})\right) \tag{7}$$

for all $b \in \bar{a}$, where $\varphi_{t,\lambda}(\mathbf{A}) = (1 - t\lambda/||\mathbf{A}||_F)_+\mathbf{A}$ and $(x)_+ = \max(0, x)$. If the resulting estimator $\widehat{\mathbf{\Omega}}$ is not positive definite or the update does not decrease the objective, we half the step size t and find new update. Once the update of the precision matrix, $\widehat{\mathbf{\Omega}}$, is found, we update the covariance matrix, $\widehat{\mathbf{\Sigma}}$. Updates to the covariance matrix can be found efficiently, without performing expensive matrix inversion as follows

$$\begin{aligned} \widehat{\mathbf{\Sigma}}_{\bar{a},\bar{a}} &= (\widetilde{\mathbf{\Omega}}_{\bar{a},\bar{a}})^{-1} + (\widetilde{\mathbf{\Omega}}_{\bar{a},\bar{a}})^{-1}\widehat{\mathbf{\Omega}}_{\bar{a},a}(\widehat{\mathbf{\Omega}}_{aa} \\ &\qquad - \widehat{\mathbf{\Omega}}_{a,\bar{a}}(\widetilde{\mathbf{\Omega}}_{\bar{a},\bar{a}})^{-1}\widehat{\mathbf{\Omega}}_{\bar{a},a})^{-1}\widehat{\mathbf{\Omega}}_{a,\bar{a}}(\widetilde{\mathbf{\Omega}}_{\bar{a},\bar{a}})^{-1}, \\ \widehat{\mathbf{\Sigma}}_{a,\bar{a}} &= -\widehat{\mathbf{\Omega}}_{aa}\widehat{\mathbf{\Omega}}_{a,\bar{a}}\widehat{\mathbf{\Sigma}}_{\bar{a},\bar{a}}, \\ \widehat{\mathbf{\Sigma}}_{aa} &= (\widehat{\mathbf{\Omega}}_{aa} - \widehat{\mathbf{\Omega}}_{a,\bar{a}}(\widetilde{\mathbf{\Omega}}_{\bar{a},\bar{a}})^{-1}\widehat{\mathbf{\Omega}}_{\bar{a},a})^{-1}, \end{aligned} \tag{8}$$

with $(\widetilde{\mathbf{\Omega}}_{\bar{a},\bar{a}})^{-1} = \widetilde{\mathbf{\Sigma}}_{\bar{a},\bar{a}} - \widetilde{\mathbf{\Sigma}}_{\bar{a},a}\widetilde{\mathbf{\Sigma}}_{aa}^{-1}\widetilde{\mathbf{\Sigma}}_{a,\bar{a}}$. Combining all the steps we arrive at the following algorithm:

1. Set the initial estimator $\widetilde{\mathbf{\Omega}} = \operatorname{diag}(\mathbf{S})$ and $\widetilde{\mathbf{\Sigma}} = \widetilde{\mathbf{\Omega}}^{-1}$. Set the step size $t = 1$.
2. For each $a \in [p]$ perform the following:
   - Update $\widehat{\mathbf{\Omega}}$. If $\widehat{\mathbf{\Omega}}$ is not positive definite, set $t \leftarrow t/2$ and repeat the update.
   - Update $\widehat{\mathbf{\Sigma}}$ using (8).
3. Repeat Step 2 until the duality gap

   $$\operatorname{tr}\mathbf{S}\widehat{\mathbf{\Omega}} - \log|\widehat{\mathbf{\Omega}}| + \lambda\sum_{a,b}||\widehat{\mathbf{\Omega}}_{ab}||_F - \sum_{j\in[p]} k_j - \log|\mathbf{\Sigma}| \leq \epsilon,$$

   where ϵ is a small, user defined parameter (for example, $\epsilon = 10^{-2}$).

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

Convergence of the above described procedure to the unique minimum of the objective in (3) does not simply follow from the standard results on the block coordinate descent (Tseng, 2001) for two reasons. First, the minimization problem in (5) is not solved to convergence at each iteration, since we only update $\mathbf{\Omega}_{aa}$ and $\mathbf{\Omega}_{a,\bar{a}}$ using one generalized gradient step update in each iteration. Second, 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.

Lemma 1. *For every value of $\lambda > 0$, proposed procedure produces a sequence of estimates $(\widehat{\mathbf{\Omega}}^{(t)})_{t\geq 1}$ of the precision matrix that monotonically decrease the objective value given in (3), are positive definite and converge to the unique minimizer $\widehat{\mathbf{\Omega}}$ of (3).*

3.2. Efficient Identification of Connected Components

When the target graph $\widehat{G}$ is composed of smaller, disconnected components, the solution to the problem in (3) is block diagonal (possibly after permuting the node indices) and can be obtained by solving smaller optimization problems. That is, the minimizer $\widehat{\mathbf{\Omega}}$ can be obtained by solving (3) for each connected component independently, resulting in massive computational gains. We give necessary and sufficient condition for the solution $\widehat{\mathbf{\Omega}}$ of (3) to be block-diagonal, which can be easily checked by inspecting the empirical covariance matrix $\mathbf{S}$.

Our first result follows immediately from the Karush-Kuhn-Tucker conditions for the optimization problem (3) and states that if $\widehat{\mathbf{\Omega}}$ is block-diagonal, then it can be obtained by solving a sequence of smaller optimization problems.

Lemma 2. *If the solution to (3) takes the form* $\widehat{\mathbf{\Omega}} = \mathrm{diag}(\widehat{\mathbf{\Omega}}_1, \widehat{\mathbf{\Omega}}_2, \ldots, \widehat{\mathbf{\Omega}}_l)$*, then it can be obtained by solving*

$$\min_{\mathbf{\Omega}_{l'} \succ \mathbf{0}} \; tr\mathbf{S}_{l'}\mathbf{\Omega}_{l'} - \log|\mathbf{\Omega}_{l'}| + \lambda \sum_{a,b} ||\mathbf{\Omega}_{ab}||_F$$

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

Next, we describe how to identify diagonal blocks of $\widehat{\mathbf{\Omega}}$. Let $\mathcal{P} = \{P_1, P_2, \ldots, P_l\}$ be a partition of the set $[p]$ 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 conditions for* $\widehat{\mathbf{\Omega}}$ *to be block diagonal with blocks* $P_1, P_2, \ldots, P_l$ *is that* $||\mathbf{S}_{ab}||_F \leq \lambda$ *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 $\mathbf{Q}$ with elements $q_{ab} = \mathbb{1}\{||\mathbf{S}_{ab}||_F > \lambda\}$ and computing connected components of the graph with adjacency matrix $\mathbf{Q}$. The lemma states also that given two penalty parameters λ_1, λ_2, $\lambda_1 < \lambda_2$ 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 networks on datasets with large number of nodes, if we are interested in networks with small number of edges. However, this is often the case when the networks are used for exploration and interpretation of complex systems.

4. Theoretical results

In this section, we provide theoretical analysis of the estimator described in §3. In particular, we provide sufficient conditions for the consistent graph structure recovery under the assumption that, for each[1] $a = 1, \ldots, kp$, $(\sigma^*_{aa})^{-1/2} X_a$ is a sub-Gaussian with parameter γ, where σ^*_{aa} is a diagonal element of $\mathbf{\Sigma}^*$. Recall that Z is a sub-Gaussian random variable if there exists a constant $\sigma \in (0, \infty)$ such that

$$\mathbb{E}[\exp(tZ)] \leq \exp(\sigma^2 t^2), \text{ for all } t \in \mathbb{R}.$$

A statement of a general result is given in (Kolar et al., 2012).

Our assumptions involve the Hessian of the function $f(\mathbf{A}) = \mathrm{tr}\,\mathbf{S}\mathbf{A} - \log|\mathbf{A}|$ evaluated at the true $\mathbf{\Omega}^*$, $\mathcal{H} = \mathcal{H}(\mathbf{\Omega}^*) = (\mathbf{\Omega}^*)^{-1} \otimes (\mathbf{\Omega}^*)^{-1} \in \mathbb{R}^{(pk)^2 \times (pk)^2}$, and the true covariance matrix $\mathbf{\Sigma}^*$. 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, $\mathcal{C}(\mathbf{\Sigma}^*) \in \mathbb{R}^{p \times p}$ with elements $\mathcal{C}(\mathbf{\Sigma}^*)_{ab} = ||\mathbf{\Sigma}^*_{ab}||_F$. Let $\mathcal{T} = \{(a,b) : ||\mathbf{\Omega}_{ab}||_F \neq 0\}$ and $\mathcal{N} = \{(a,b) : ||\mathbf{\Omega}_{ab}||_F = 0\}$. With this notation introduced, we assume that the following *irrepresentable* condition holds; there exists a constant $\alpha \in [0, 1)$ such that

$$|||\mathcal{C}\left(\mathcal{H}_{\mathcal{N}\mathcal{T}}(\mathcal{H}_{\mathcal{T}\mathcal{T}})^{-1}\right)|||_\infty \leq 1 - \alpha. \tag{9}$$

We will also need the following quantities to specify the results $\kappa_{\mathbf{\Sigma}^*} = |||\mathcal{C}(\mathbf{\Sigma}^*)|||_\infty$ and $\kappa_{\mathcal{H}} = |||\mathcal{C}(\mathcal{H}^{-1}_{\mathcal{T}\mathcal{T}})|||_\infty$. These conditions extend the conditions specified in Ravikumar et al. (2011) needed for estimation of networks from single attribute observations.

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

Proposition 4. *Set the penalty parameter* λ *in (3) as*

$$\lambda = 8k\alpha^{-1}\sqrt{128(1+4\gamma^2)^2(\max_a(\sigma^*_{aa})^2)} \times \sqrt{\frac{2\log(2k) + \tau\log(p)}{n}},$$

where $\tau > 2$*. 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*,*

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

[1] For simplicity of presentation, we assume that $k_a = k$, for all $a \in [p]$, that is, we assume that the same number of attributes is observed for each node.

and

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

then $\mathbb{P}\left(\widehat{G} = G\right) \geq 1 - p^{2-\tau}$.

5. Simulation studies

In this section, we perform a set of simulation studies to illustrate finite sample performance of our procedure. We demonstrate that the scalings predicted by the theory are sharp. Furthermore, we compare against three other procedures: 1) a procedure that uses the `glasso` first to estimate one network over each of the k individual attributes and then creates an edge in the resulting network if an edge appears in at least one of the single attribute networks, 2) that of Guo et al. (2011) and 3) that of Chiquet et al. (2011) (see also Danaher et al., 2011). 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 network if the corresponding block of the estimated precision matrix is non-zero. The results were worse compared to the first baseline, so we do not report them here. The tuning parameters are selected by minimizing the Bayesian information criterion, which balances the goodness of fit of the model and its complexity, over a grid of parameter values. For our multi-attribute method, it takes the following form

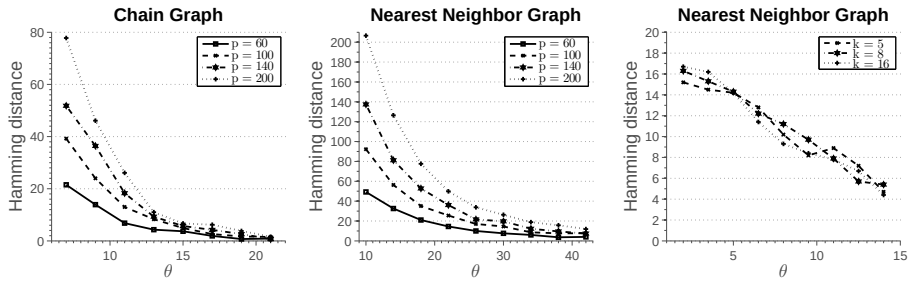
$$\text{BIC}(\lambda) = \text{tr}\,\mathbf{S}\widehat{\mathbf{\Omega}} - \log|\widehat{\mathbf{\Omega}}| + \sum_{a<b} \mathbb{I}\{\widehat{\mathbf{\Omega}}_{ab} \neq \mathbf{0}\} k_a k_b \log(n).$$

Theoretical results given in §4 predict 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 $\theta > 0$. Therefore, if we plot the hamming distance between the true and recovered graph structure 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.

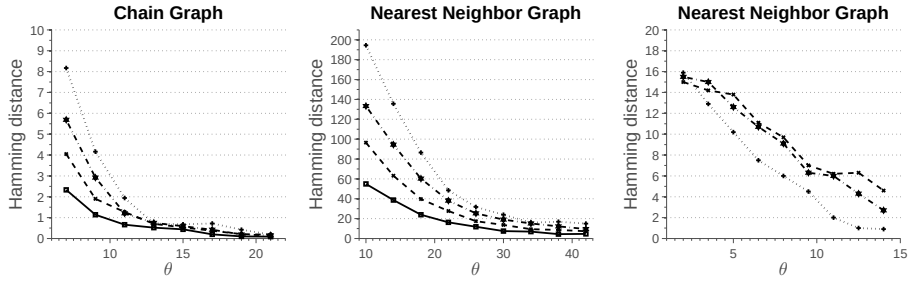
We generate data as follows. A random graph with p nodes is created by first partitioning nodes into $p/20$ connected components, each with 20 nodes, and then forming a random graph over these 20 nodes. A chain graph is formed by permuting the nodes and connecting them in succession, while a nearest-neighbor graph is constructed following the procedure outlined in Li & 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 remove randomly edges from nodes that have degree larger than 4 until the maximum degree of a node in a network is 4. Once the graph structure is created, we construct a precision matrix, with non-zero blocks corresponding to edges in the graph. Elements of the diagonal blocks take values as $0.5^{|a-b|}$, $0 \leq a, b \leq k$, while off-diagonal blocks have elements with the same value, 0.2 for chain graphs and $0.3/k$ for nearest-neighbor networks. Finally, we add $\rho\mathbf{I}$ 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 independent runs.

Figure 1 shows results of the simulations. Each row in the figure reports results for one procedure, while each column in the figure represents a different simulation setting. For the first two columns, we set $k = 3$ and vary the total number of nodes in the graph p. 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 method of (Chiquet et al., 2011) outperforms all the other procedures. We note that the multi-attribute method is estimating many more parameters, which require large sample size in order to be estimated consistently. However, as the sample size increases, we observe that multi-procedure starts to outperform the other procedures. In particular, for the sample size indexed by $\theta = 13$ all the graph are correctly recovered, while other procedures 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 procedure outperforms the other methods. Furthermore, as the sample size increases none of the other method recover the graph exactly. This suggests that the conditions for the consistent graph recovery may be weaker in the multi-attribute setting.

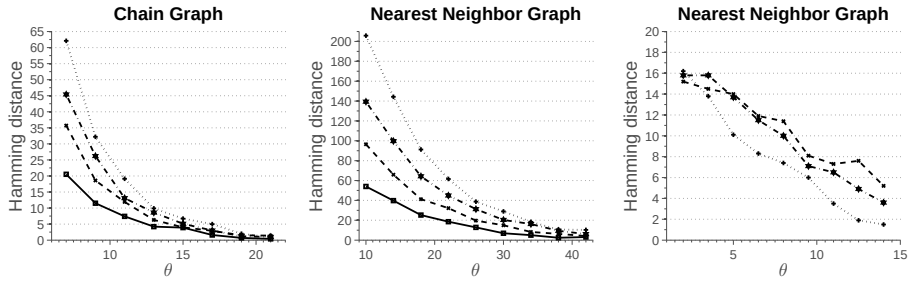
Next we investigate a situation where the multi-attribute procedure does not perform as well as the procedures that estimate multiple graphical models. One such situation arises when different attributes are conditionally independent. To simulate this situation, we follow the data generating approach as before, however, we make each block $\mathbf{\Omega}_{ab}$ of the precision matrix Ω a diagonal matrix. Figure 2 summarizes results of the simulation. We observe that methods of



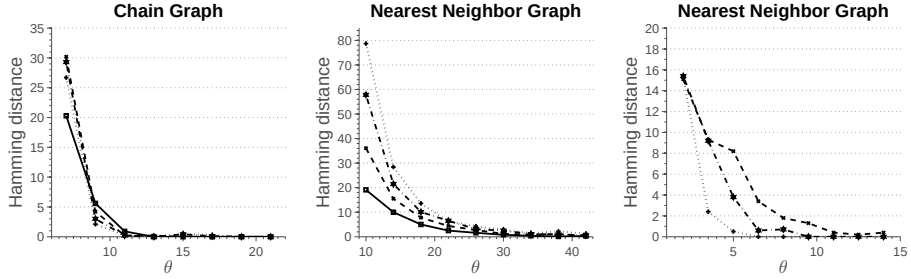
(a) glasso procedure



(b) Procedure of Chiquet et al. (2011)



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

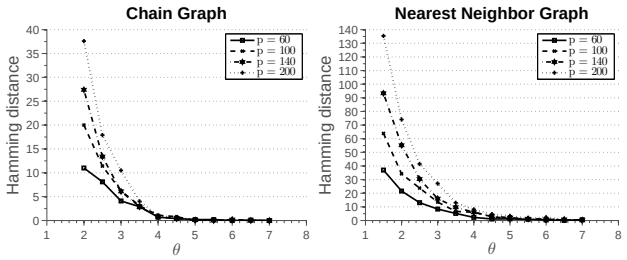


(d) Multi-attribute procedure

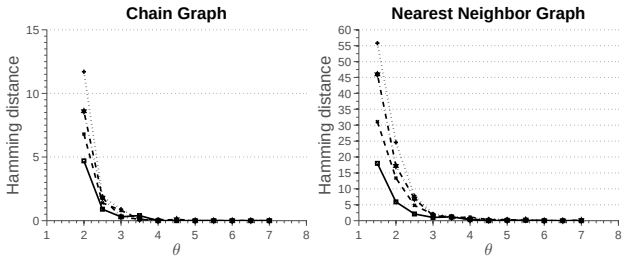
Figure 1. Average hamming distance plotted against the rescaled sample size. Results are averaged over 100 independent runs. Off-diagonal blocks are full matrices.

(Chiquet et al., 2011) and (Guo et al., 2011) perform better, since they are estimating much fewer parameters than the multi-attribute procedure. Glasso does not utilize any structural information underlying the estimation problem and requires larger sample size to estimate the graph correctly than other procedures.

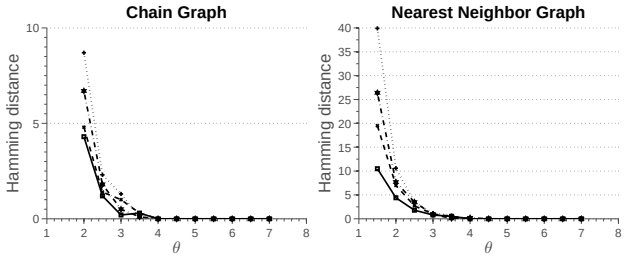
A completely different situation arises when the edges between nodes can be inferred only based on inter-attribute data, that is, when a graph based on any individual attribute is empty. To generate data under this situation, we follow the procedure as before, but with the diagonal elements of the off-diagonal blocks $\mathbf{\Omega}_{ab}$ set to zero. Figure 3 summarizes results of the simulation. In this setting, we clearly see the advantage of the multi-attribute procedure.



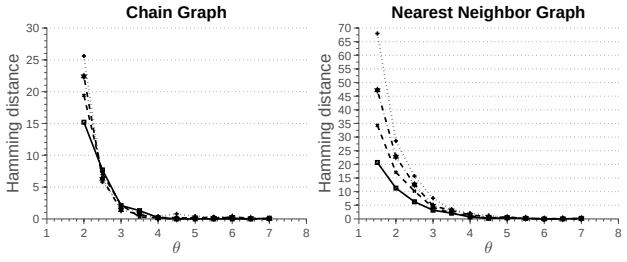
(a) glasso procedure



(b) Procedure of Chiquet et al. (2011)

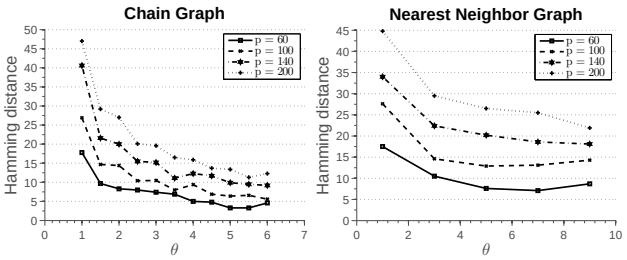


(c) Procedure of Guo et al. (2011)

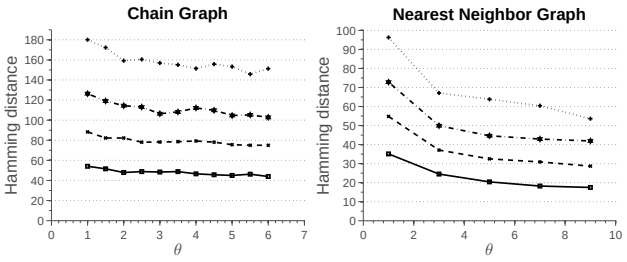


(d) Multi-attribute procedure

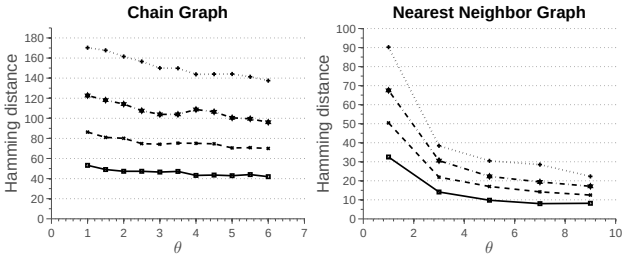
Figure 2. Average hamming distance plotted against the rescaled sample size. Results are averaged over 100 independent runs. Blocks $\mathbf{\Omega}_{ab}$ of the precision matrix $\mathbf{\Omega}$ are diagonal matrices.



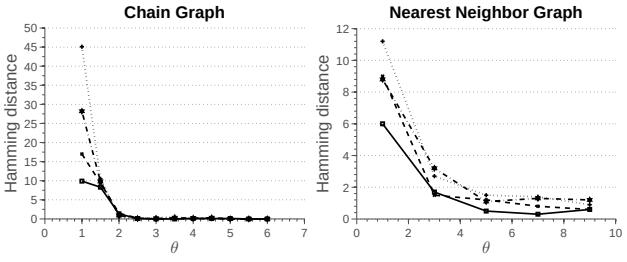
(a) glasso procedure



(b) Procedure of Chiquet et al. (2011)



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



(d) Multi-attribute procedure

Figure 3. Average hamming distance plotted against the rescaled sample size. Results are averaged over 100 independent runs. Off-diagonal blocks $\mathbf{\Omega}_{ab}$ of the precision matrix $\mathbf{\Omega}$ have zeros as diagonal elements.

6. Discussion and Extensions

In this paper, we have proposed a solution to the problem of learning networks from multivariate nodal attributes, which arises in a variety of domains. Our method is based on simultaneously estimating non-zero partial canonical correlations between nodes in a network. When all the attributes across all the nodes follow joint multivariate Normal distribution, our procedure is equivalent to estimating conditional independencies between nodes, which is revealed by relating the blocks of the precision matrix to partial canonical correlation. Although a penalized likelihood framework is adopted in the current paper for estimation of the non-zero blocks of the precision matrix, other approaches like neighborhood pursuit or greedy pursuit can also be developed. Thorough numerical evaluations and theoretical analysis of these methods is an interesting direction for future work.

References

Banerjee, O., El Ghaoui, L., and dAspremont, A. Model selection through sparse maximum likelihood estimation. *J. Mach. Learn. Res.*, 9:485–516, 2008.

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

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

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

Danaher, P., Wang, P., and Witten, D. M. The joint graphical lasso for inverse covariance estimation across multiple classes. Technical report, University of Washington, 2011.

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

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

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

Honorio, J. and Samaras, D. Multi-task learning of Gaussian graphical models. In Fürnkranz, Johannes and Joachims, Thorsten (eds.), *Proc. 27 Int. Conf. Mach. Learn.*, pp. 447–454. Omnipress, Haifa, Israel, June 2010.

Johnson, C., Jalali, A., and Ravikumar, P. High-dimensional sparse inverse covariance estimation using greedy methods. In Lawrence, Neil and Girolami, Mark (eds.), *Proc. 15 Int. Conf. Artif. Intel. Statist.*, pp. 574–582. 2012.

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

Kolar, M. and Xing, E. P. Consistent covariance selection from data with missing values. In Langford, John and Pineau, Joelle (eds.), *Proc. 29 Int. Conf. Mach. Learn.*, pp. 551–558, New York, NY, USA, July 2012. Omnipress.

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

Kolar, M., Liu, H., and Xing, E. P. Graph estimation from multi-attribute data. Technical report, Carnegie Mellon University (arXiv:1210.7665), 2012.

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

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

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

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

Peng, Jie, Wang, Pei, Zhou, Nengfeng, and Zhu, Ji. Partial correlation estimation by joint sparse regression models. *J. Am. Statist. Assoc.*, 104(486):735–746, 2009.

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

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

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

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

Varoquaux, G., Gramfort, A., Poline, J.-B., and Thirion, B. Brain covariance selection: better individual functional connectivity models using population prior. In Lafferty, J., Williams, C. K. I., Shawe-Taylor, J., Zemel, R.S., and Culotta, A. (eds.), *Adv. Neural Inf. Proc. Sys. 23*, pp. 2334–2342. 2010.

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

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

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

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