Penalised Iterative Sparse Partial Correlation Estimation (Π-SPaCE)

- with an application to whole-brain graph estimation

Oliver Chén [1*], Junrui Di[1], Luo Xiao[2]

[1]Department of Biostatistics, the Johns Hopkins University Bloomberg School of Public Health
[2]Department of Statistics, North Carolina State University

[*]To whom correspondence should be addressed: E-mail: oliver@jhmi.edu

Abstract

Sparse matrix estimation is often used in network science including neuroscience, social network, and genomic study, where the networks are high-dimensional and sparse. Graph estimation is subsequently used to numerically and visually delineate the networks between different brain voxels, individuals, or genes. While there has been significant research on the topic in recent years, most existing methods require pre-selecting the non-zero support set of the correlation matrix, or entailing a time-consuming block-wise estimation fashion. As a motivating example, consider a functional magnetic resonance imaging (fMRI) study of thermal pain where, while we have little prior information of the non-zero support brain regions, we are interested in determining the whole-brain network (between hundreds of thousands of voxels) under thermal treatment. To address the problem of ultra-high-dimensional network estimation where little prior information is present, we propose a framework called the Penalized Iterative Sparse Partial Correlation Estimation (Π-SPaCE). This framework does not require prior information: it allows us to estimate the off-diagonal elements of the partial correlation matrix directly, and is faster than traditional methods in the high-dimensional sparse matrix setting. We study this method using simulation and an application to whole-brain graph estimation using data from an fMRI study.

Keywords Sparse Partial Correlation Estimation; Graph Estimation; Network Study; fMRI

1 Introduction

Denote an $n \times p$ matrix $\mathbf{X}$ consisting of of$p$ random variables $(\mathbf{X}_1, \cdots, \mathbf{X}_p)^T$ from n multivariate normal observations, with mean μ and covriance $\mathbf{\Sigma}$. The partial correlation between $\mathbf{X}_i$ and $\mathbf{X}_j$, the i^{th} and j^{th} variables is defined as

$$\rho_{ij|\mathbf{X}_{(-ij)}} := \text{corr}\big(R_{\mathbf{X}_i|\mathbf{X}_{(-ij)}}, R_{\mathbf{X}_j|\mathbf{X}_{(-ij)}}\big) \tag{1}$$

where $\mathbf{X}_{(-ij)}$ indicates the $\mathbf{X}$ after deleting the i^{th} and j^{th} variables, and $R_{\mathbf{X}_i|\mathbf{X}_{(-ij)}}$ defines the residual from regressing $\mathbf{X}_i$ on $\mathbf{X}_{(-ij)}$. For simplicity, define $\rho_{ij|\mathbf{X}_{(-ij)}} := \rho_{ij}$ henceforth.

Define $\mathbf{\Theta} = \mathbf{\Sigma}^{-1}$ and let $\mathbf{S} := \frac{1}{n}\sum_{i=1}^{n}(\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})^\intercal$ denote the sample covariance matrix. The multivariate Gaussian $\log$ likelihood is:

$$\log \mathcal{L}(\mathbf{\Theta}) = \log\det(\mathbf{\Theta}) - \text{tr}(\mathbf{S\Theta}) + \text{const} \tag{2}$$

where $\det(\cdot)$ defines determinant, tr denotes the trace.

Banerjee et al. (2007) define the partially maximal $\log$ likelihood approach with respect to μ as:

$$\mathbf{\Theta}^* = \arg\max_{\mathbf{\Theta} \succeq 0} \left\{ \log\det \mathbf{\Theta} - \text{tr}(\mathbf{S\Theta}) - \lambda \parallel \mathbf{\Theta} \parallel_{L_1} \right\} \tag{3}$$

where $\succeq$ indicates non-negative definite, and $\| \cdot \|_{L_1}$ denotes the L_1 norm.

Banerjee et al. (2007) show that (1) is convex and give an estimation of $\mathbf{\Sigma}$ using a block-wise interior-point procedure as follows.

Let $\hat{\mathbf{\Sigma}}$ be the estimator of $\mathbf{\Sigma}$, which is obtained in following steps.

Step 1: Partition $\hat{\mathbf{\Sigma}}$ and $\mathbf{S}$ such that $\hat{\mathbf{\Sigma}} = \begin{pmatrix} \hat{\mathbf{\Sigma}}_{11} & \hat{\sigma}_{12} \\ \hat{\sigma}_{12}^T & \hat{\sigma}_{22} \end{pmatrix}$ and $\mathbf{S} = \begin{pmatrix} \mathbf{S}_{11} & \mathbf{s}_{12} \\ \mathbf{s}_{12}^T & s_{22} \end{pmatrix}$;

Step 2:

$$\hat{\sigma}_{12} = \arg\min_{\mathbf{y}} \left\{ \mathbf{y}^T \hat{\mathbf{\Sigma}}_{11}^{-1} \mathbf{y} : \parallel \mathbf{y} - \mathbf{s}_{12} \parallel \leq \lambda \right\}; \tag{4}$$

Step 3: Permute the rows and columns so that the last column is the target column. Upate $\hat{\Sigma}$ every time after solving (4) for each column, until convergence.

Under convex duality, Banerjee et al (2007) show that solving (3) is equivalent to solving a lasso-like least square problem:

$$\min_{\boldsymbol{\beta}} \left\{ \frac{1}{2} \parallel \boldsymbol{\Sigma}_{11}^{1/2}\boldsymbol{\beta} - \boldsymbol{\Sigma}_{11}^{-1/2}\mathbf{s}_{12} \parallel^2 + \lambda \parallel \boldsymbol{\beta} \parallel_{L_1} \right\} \tag{5}$$

Friedman, Hastie, and Tibshirani (2007) state, if $\hat{\boldsymbol{\Sigma}}_{11} = \mathbf{S}_{11}$, the solution to (4) equals to the lasso estimates for the p^{th} variable on the others, which is related to the Meinshausen & Bühlmann (2006) proposal; if $\hat{\boldsymbol{\Sigma}}_{11} \neq \mathbf{S}_{11}$, Meinshausen & Bühlmann (2006) approach does not give the MLE. Banerjee et al (2007) show that their block-wise interior-point procedure is equivalent to recursively solve and update (4). Friedman, Hastie, and Tibshirani (2007) implement the approach and call it the graphical lasso (GLASSO) algorithm.

Equation (3) is straight-forward in estimating the precision matrix, but it does not allow directly estimating for partial correlation, whose only parameters to be estimated are off-diagonal elements. Peng et al. (2009) propose an active shooting algorithm for estimating the partial correlation in sparse case, but the complexity of the procedure can be avoid. We propose a system that can estimate partial correlation directly, is simple, and can achieve faster convergence than existing methods.

Ha & Sun, 2014 define the *scale* operator as follows:

for any square matrix $\mathbf{A}$,

$$\text{scale}(\mathbf{A}) = \text{diag}(\mathbf{A})^{-1/2}\mathbf{A}\text{diag}(\mathbf{A})^{-1/2} \tag{6}$$

where $\text{diag}(\mathbf{A})$ is the diagonal elements of square matrix A.

The the partial correlations can be obtained from the off diagonal elements of the negative definite matrix $-\text{scale}(\boldsymbol{\Theta})$, viz.:

$$\tilde{\mathbf{P}} = \{\boldsymbol{\Theta}_{ij}\}_{n\times n} = -\text{scale}\ (\boldsymbol{\Theta}) \tag{7}$$

where $\tilde{\mathbf{P}}$ is the partial correlation matrix.

Hence,

$$\boldsymbol{\Theta} = \mathbf{D}(-\tilde{\mathbf{P}})\mathbf{D} \tag{8}$$

where $\mathbf{D} = \text{diag}(\boldsymbol{\Theta})^{1/2}$.

Therefore, plug (8) into (2) we can define the $\log$ likelihood function as:

$$l(\mathbf{Y}; \mathbf{P}, \mathbf{D}) = \log\det[\mathbf{DQD}] - \text{tr}[\mathbf{SDQD}] \tag{9}$$

where $\mathbf{Q} = -\tilde{\mathbf{P}}$, $\mathbf{Q}_{\text{off}} = -\mathbf{P}_{\text{off}}$, and $\Psi = (\mathbf{P}, \mathbf{D}) \in \Gamma \times \Delta$.

Hence, we can estimate the partial correlation by solving for Ψ that maximizes the Gaussian $\log$ likelihood, with a constrain that the off-diagonal elements of the partial correlation matrix are L_1 penalized.

Banerjee et al. (2007) defined the penalized multivariate Gaussian log-likelihood and proposed to estimate the covariance matrix Σ using a block-wise interior-point procedure.Incorporating these concepts with coordinate descent procedure for the LASSO, Friedman, Hastie, and Tibshirani (2007) proposed Graphical LASSO as a simple and substantially faster approach which became the milestone in the field. However, this method directly gives the estiamte for the covariance matrix Σ itself. Since our main interest here is to estimate the partial correlation matrix $\mathbf{P}$ which can be achieved by $-\text{scale}(\boldsymbol{\Theta})$ where $\boldsymbol{\Theta} = \Sigma^{-1}$, it is ideal and computationally more efficient to directly estimate the inverse of covariance matrix in the ultra-high dimensional setting.

Ha & Sun, 2014 developed a three-step approach as followed (1) obtain a penalized estimate of partial correlation matrix using ridge penalty, (2)select the non-zero entries of the partial correlation matrix using hypothesis testing , and (3) re-estimate the partial correaltion coefficients at these non-zero entries. They showed that ridge estimate is desirable because it leads to parsimonious and more interpretable partial correlation matrix estimation, and further reduces estimations as well. In their application to the real genetic data where p is much larger than n, they cluster genes first and then estimate partial correlation matrix within each cluster under the assumption that partial correlation matrix of gene expression has a block diagonal; however, this assumption is not always valid in the neuro-imaging field, thus this approach may be computationally costly applying to all the nodes. Meanwhile, since their method does not require multivariate Gaussian distirbution assumption, estimated partial correlation being zero may not imply the two variables are conditionally independent with each other. Moreover, since their method borrows information across all the variables by utilizing a common tuning parameter across all variables, bias is increased even though the variance is reduced.

Peng et al. (2009) proposed a method (SPACE) for detecting paris of variables having nonzeor partial correlations among a large number of random variables. This methods assumes the overall sparsity of partial correlation matrix and employs sparse regression techniques for model fitting, and they use their active-shooting algorithm to efficiently solve for the LASSO regression problem. They showed that SPACE outperforms GLASSO in both edge detection and hub identification in speed and complexity. Nevertheless, their estimation relies on prior knowledge of the network structure by assigning different weights to different nodes, which in reality, may not always be available.

To the best of our knowledge, all existing methods for estimating partial correlation matrix

are limited to data with variables with magnitude of 10^3. In this paper, we introduce a framework that can estimate partial correlation matrix in ultra-high dimensional data setting, where for example, $200,000$ brain voxels are considered.

2 Data Description

Our proposed method is inspired by functional MRI research. Consider an resting stat fMRI study with n subjects and each subject has n_i trials. During each trial, fMRI signals are measured over p voxels, where p can be ultra-high dimensional, and stored in a row vector M_k of length p for $k \in \{1 \cdots, N\}$. where $N = \sum_{i=1}^{n} n_i$.

Hence we obtain a signal matrix $M_{N \times p}$, where N indicates the total number of trials for all subjects, and p represents the number of voxels studied.

3 Methods

3.1 Model

It is of great importance to first achieve the empirical covariance matrix.

Consider the following model to compute the empirical covrariance matrix

$$s_{j,k} = f(M[,j], M[,k]) \tag{10}$$

where $M[,j]$ and $M[,k]$ represents the jth and kth columns of M. In particular, they represents the fMRI signals over voxels j and k respectively, across subjects. $f : \mathbb{R}^n \times \mathbb{R}^n \mapsto \mathbb{R}^1$ is the correlation mapping. Finally, the empirical covariance matrix $S = \{s_{j,k}\}_{j,k=1}^{n}$

3.2 Framework

We define the penalized log likelihood as:

$$\hat{\mathbf{\Psi}} = \underset{\left\{ \begin{array}{c} \mathbf{Q}_{\text{off}}=-\mathbf{P}_{\text{off}} \\ d\in R_+^p:\mathbf{D}=diag(d) \end{array} \right\}}{\text{argmax}} \left\{ \log\det[\mathbf{DQD}] - \text{tr}[\mathbf{SDQD}] + \lambda \parallel \boldsymbol{Q} \parallel_{\text{off},1} \right\} \tag{11}$$

where $\parallel \boldsymbol{Q} \parallel_{\text{off},1}$ indicates the diagonal elements of $\boldsymbol{Q}$ are constrained to be 1's and the off-diagonal elements are L_1 penalized.

Notice in equation (10), **D** appears in quadratic term **DQD**. In order to find a closed form maximizer **D**, we consider conditioning on one of them, say, the first **D** in the quadratic term. Specifically, we denote $\mathbf{D}_1$ as the term on which we condition, and equation (10) can be expressed as:

$$\hat{\mathbf{\Psi}} = \underset{\left\{ \begin{array}{c} \mathbf{Q}_{\text{off}}=-\mathbf{P}_{\text{off}} \\ d\in R_+^p:\mathbf{D}_1,\mathbf{D}_2=diag(d) \end{array} \right\}}{\text{argmax}} \left\{ \log\det[\mathbf{D}_1\mathbf{QD}_2] - \text{tr}[\mathbf{SD}_1\mathbf{QD}_2] + \lambda \parallel \boldsymbol{Q} \parallel_{\text{off},1} \right\}. \tag{12}$$

For a pre-selected tuning parameter λ, conditioning on $\tilde{\mathbf{D}}$ and **Q**, by equation (11), we have [see Appendix for derivation]:

$$\hat{\mathbf{D}}_2|\mathbf{D}_1, \mathbf{Q} = \left(\text{diag}(\mathbf{SD}_1\mathbf{Q})\right)^{-1} \tag{13}$$

where **S** is the sample covariance.

Conditioning on $\tilde{\mathbf{D}}$ and **D**, by equation (11), we have [see Appendix for derivation]:

$$\hat{\mathbf{Q}}_{ij}|\mathbf{D}_1, \mathbf{D}_2 = \left\{\lambda\text{sgn}(\mathbf{Q}_{ij}) + (\mathbf{D}_1\mathbf{SD}_2)_{ij}\right\}^{-1} \tag{14}$$

where $(\cdot)_{ij}$ indicates the ij^{th} element of a matrix, $\forall i \neq j$. And $\forall i = j$, $\hat{\mathbf{Q}}_{ij}$=-1.

Or,

$$\hat{\mathbf{Q}}|\mathbf{D}_1, \mathbf{D}_2 = \left\{\mathbf{D}_1\mathbf{SD}_2 - \lambda \mid \mathbf{1} \mid_p^{-1} \right\} \tag{15}$$

where $|\ \mathbf{1}\ |_p^{-1} = \begin{cases} \text{sgn}(\mathbf{Q}_{ij}) & i \neq j; \\ 1 & i = j \end{cases}$.
The second way might give rise to faster speed.

Similarily,

$$\hat{\mathbf{D}}_1 | \mathbf{D}_2, \mathbf{Q} = \left(\text{diag}(\mathbf{S}\mathbf{D}_2\mathbf{Q})\right)^{-1}. \tag{16}$$

Start with initial $\mathbf{D}_1$ and $\mathbf{Q}$, defined as $\mathbf{D}_1^{(0)}$ and $\mathbf{Q}^{(0)}$, respectively.

Set $\hat{\mathbf{D}}_2^{(1)} | \mathbf{D}_1^{(0)}, \mathbf{Q}^{(0)} = \left(\text{diag}(\mathbf{S}\mathbf{D}_1^{(0)}\mathbf{Q}^{(0)})\right)^{-1}$.

Set $\hat{\mathbf{Q}}_{ij}^{(1)} | \mathbf{D}_1^{(0)}, \mathbf{D}_2^{(1)}, \mathbf{Q}^{(0)} = \left\{\lambda \text{sgn}(\mathbf{Q}_{ij}^{(0)}) + (\mathbf{D}_1^{(0)}\mathbf{S}\mathbf{D}_2^{(1)})_{ij}\right\}^{-1}$, $\forall i \neq j$, and -1 $\forall i = j$

Set $\hat{\mathbf{D}}_1^{(1)} | \mathbf{D}_2^{(1)}, \mathbf{Q}^{(1)} = \left(\text{diag}(\mathbf{S}\mathbf{D}_2^{(1)}\mathbf{Q}^{(1)})\right)^{-1}$.

Notice that intuitively we should expect $D_{ii} \leq 1$ when $\Sigma_{ii} \geq 1$ for sparse covariance matrix Σ, and vise versa. So that we consider the the following stopping criterion:

for some pre-defined thresholding value ϵ small, repeat until

$\min\{|\ \mathbf{D}_1^k - \mathbf{D}_2^k\ |, |\ (\mathbf{D}_1^k)^{-1} - (\mathbf{D}_2^k)^{-1}\} \leq \epsilon$ and $|\ \mathbf{Q}^k - \mathbf{Q}^{k-1}\ | \leq \epsilon$

$\{$

For each k, set

$$\hat{\mathbf{D}}_2^{(k)} | \mathbf{D}_1^{(k-1)}, \mathbf{Q}^{(k-1)} = \left(\text{diag}(\mathbf{S}\mathbf{D}_1^{(k-1)}\mathbf{Q}^{(k-1)})\right)^{-1} \tag{17}$$

$$\hat{\mathbf{Q}}_{ij}^{(k)} | \mathbf{D}_1^{(k-1)}, \mathbf{D}_2^{(k)}, \mathbf{Q}^{(k-1)} = \left\{\lambda \text{sgn}(\mathbf{Q}_{ij}^{(k-1)}) + (\mathbf{D}_1^{(k-1)}\mathbf{S}\mathbf{D}_2^{(k)})_{ij}\right\}^{-1}, \forall i \neq j \text{ and } 1\ \forall i = j \tag{18}$$

$$\hat{\mathbf{D}}_1^{(k)} | \mathbf{D}_2^{(k)}, \mathbf{Q}^{(k)} = \left(\text{diag}(\mathbf{S}\mathbf{D}_2^{(k)}\mathbf{Q}^{(k)})\right)^{-1}. \tag{19}$$

$\}$.

3.3 Ultra-high Dimensional Setting

4 Theory: Asymptotic Property of Estimates

5 Simulation

6 fMRI Graph Estimation

7 Conclusion

8 Appdendix

Lemma 1 (Matrix tricks):

(a)

$$\frac{\partial \log \det(\mathbf{X})}{\partial(\mathbf{X})} = \mathbf{X}^{-1};$$

(b)

$$\frac{\partial \det \mathbf{AXB}}{\partial \mathbf{X}} = \det \mathbf{AXB}(\mathbf{X}^T)^{-1};$$

(c)

$$\frac{\partial \text{tr}(\mathbf{XA})}{\partial \mathbf{X}} = \mathbf{A}^T;$$

(d)

$$\frac{\partial \sum_{i,j} \mid x_{ij} \mid}{\partial \mathbf{X}} = \left\{ \frac{x_{ij}}{\mid x_{ij} \mid} \right\}_{(i,j)=(1,1)}^{(i,j)=(p,p)} = \left\{ \text{sgn}(x_{ij}) \right\}_{(i,j)=(1,1)}^{(i,j)=(p,p)};$$

where $\left\{ \text{sgn}(x_{ij}) \right\}_{(i,j)=(1,1)}^{(i,j)=(p,p)}$ is a $p \times p$ symmentric matrix with entries 1 and -1, which we define as $\mid \mathbf{1} \mid_p$;

(e)

$$\frac{\partial \det(\mathbf{X}^T\mathbf{A}\mathbf{X})}{\partial \mathbf{X}} = 2\det(\mathbf{X}^T\mathbf{A}\mathbf{X})(\mathbf{X}^{-1})^T;$$

(f)

$$\frac{\partial \text{tr}(\mathbf{A}\mathbf{X}\mathbf{B}\mathbf{X}^T\mathbf{C})}{\partial \mathbf{X}} = \mathbf{A}^T\mathbf{C}^T\mathbf{X}\mathbf{B}^T + \mathbf{C}\mathbf{A}\mathbf{X}\mathbf{B}\,.$$

Proof of equation (12):

Define:

$$\mathcal{L}(\mathbf{D}, \tilde{\mathbf{D}}, \mathbf{Q}|\mathbf{Y}) = \log\det[\tilde{\mathbf{D}}\mathbf{Q}\mathbf{D}] - \text{tr}[\mathbf{S}\tilde{\mathbf{D}}\mathbf{Q}\mathbf{D}] + \lambda \parallel \boldsymbol{Q} \parallel_{\text{off},1}\,.$$

By Lemma (a) - (c), we have:

$$\frac{\partial \mathcal{L}(\mathbf{D}, \tilde{\mathbf{D}}, \mathbf{Q}|\mathbf{Y})}{\partial \mathbf{D}} | \tilde{\mathbf{D}}, \mathbf{Q} = (\mathbf{D}^T)^{-1} - (\mathbf{S}\tilde{\mathbf{D}}\mathbf{Q})^T.$$

Let the above equation equal to 0, the result follows. □

Proof of equation (13):

By Lemma (d) - (e), we have:

$$\frac{\partial \mathcal{L}(\mathbf{D}, \tilde{\mathbf{D}}, \mathbf{Q}|\mathbf{Y})}{\partial \mathbf{Q}} | \tilde{\mathbf{D}}, \mathbf{D} = (\mathbf{Q}^T)^{-1} - (\mathbf{D}\mathbf{S}\tilde{\mathbf{D}})^T + \lambda \mid 1 \mid_p.$$

Let the above equation equal to 0, the result follows. □

9 Reference

1. Peng, Wang, Zhou, and Zhu (2009). Partial Correlation Estimation by Joint Sparse Regression Models. Journal of the American Statistical Association, 104:486, 735-746.

2. Friedman, Hastie, and Tibshirani (2007). Sparse inverse covariance estimation with the graphical lasso. Biostatistics, 9 (3): 432-441.

3. Ha and Sun (2014). Partial Correlation Matrix Estimation Using Ridge Penalty Followed by Thresholding and Re-estimation. Biometrics, 70, 765773

4. Petersen and Pedersen (2006). The Matrix Cookbook.