Sparse Network Modeling

Moo K. Chung
mchung@stat.wisc.edu

September 29, 2012

Many brain connectivity studies are usually performed following the standard analysis frameworks such as structural equation modeling (SEM), dynamic causal modeling (DCM) or partial correlation modeling [10, 12, 14, 15, 27, 26, 31, 41]. From whole brain images, \( n \) regions of interest (ROI) are identified and serve as the nodes of the brain network. Measurements at ROIs are then correlated in a pair-wise fashion to produce the connectivity matrix of size \( n \times n \).

The connectivity matrix is then thresholded to produce the adjacency matrix consisting of zeros and ones that define the link between two nodes. The adjacency matrix is then used to construct the brain network visually and to detect focal regions of correlated nodes [6, 23]. However, for a large number of nodes, this brute force approach has a serious computational bottleneck. Even if we solve the computational problem, biomedical interpretation of network will be difficult with the huge number of links. For example, for \( 10^6 \) voxels in an image, we can possibly have a total of \( 10^{12} \) links in the graph. Therefore, it is necessary to thin out the number of edges somehow [24, 40].

[40] proposed to use the singular value decomposition (SVD) in showing that SVD is better at detecting extensive regions of correlated voxels compared to the traditional method of simple correlation thresholding. Let \( X_{n \times p} = (x_{ij}) \) be the matrix of \( p \) regions and \( n \) subjects. In the whole brain studies, it is likely that the number of regions \( p \) is significantly smaller than the number of subjects \( n \). So we expect the large-\( p \) small-\( n \) problem to occur. We assume \( X \) to be centered by subtracting their mean value. We further assume that each column of \( X \) is normalized by dividing by its root sum of squares, so that the diagonal elements of the cross-correlation matrix \( \Sigma_{p \times p} = X' X \) is 1. The SVD of \( \Sigma \) is then

\[
\Sigma = U W U',
\]

1
where \( U \) is an orthonormal matrix and \( W \) is a diagonal matrix of component weights. Then \( \Sigma \) is estimated by setting the smaller weights in \( W \) to be zero [40]. This is related to the principal component analysis or partial least squares (PLS) [29, 30]. PSL is similar to PCA but the solutions of PSL are constrained to be the part of the covariance structure. Since \( p \) can possibly reach upward of few million voxels, the computational burden of finding SVD of \( \Sigma \) can be prohibitive in small computers. [40] proposed to bypass the problem by matrix decompositions. Afterward, the statistical inference is done either using permutation tests [32] or the random field theory [6, 38]. Since the brain networks are known to be sparse and highly clustered [1, 19], it is reasonable to incorporate the sparsity of network structures into PCA further. There have been various attempts in incorporating spasticity in PCA using LASSO (least absolute shrinkage and selection operator) in statistics [22, 42]. LASSO is a widely used variable selection technique that produces sparse models [37]. It is based on the observation that PCA can be reformulated as the optimal solution of a regression so that LASSO can be integrated into the regression.

The main limitation of the connectivity studies based on correlation is that it fails to explicitly factor out the confounding effect of other regions. To remedy this limitation, partial correlation has been introduced in factoring out the dependencies of other regions [19, 27] or eliminating the effect of the experimental design [29]. Since the partial correlation corresponds to the off-diagonal entries of the inverse covariance matrix, sparse PCA can be used to invert the covariance matrix. A similar frameworks found applications in image classification [4], gene expression [7], flow cytometry data [11] and functional brain network models [20, 21].

In this chapter, we introduce sparse partial correlation modeling technique for brain connectivity studies. This chapter follows the derivations given in [24].

1 Correlations

In the whole brain connectivity studies, the number of nodes \( p \) are expected to be larger than the number of subjects \( n \), which gives an underdetermined system. Consider measurement matrix \( X_{n \times p} = (x_{ij}) \) of \( n \) subjects and \( p \) regions. At node \( j \), we have the random variable \( x_j \), which is realized by the observations \( x_{1j}, \ldots, x_{nj} \). We will denote this realization as

\[
x_j = (x_{1j}, \ldots, x_{nj})'.
\]
So at the node $j$, the $i$-th measurement is given by $x_{ij}$. The collection of random variables $(x_1, \cdots, x_p)'$ are assumed to be distributed with mean zero and covariance $\Sigma = (\sigma_{jk})$ i.e.

$$E x_j = 0, \ E(x_jx_k) = \sigma_{jk}.$$ 

If $E x_j \neq 0$, we can always center the data by translation. We may further assume that the diagonal terms in the covariance matrix is 1, i.e. $E x_jx_j = 1$. By scaling the data by the standard deviation, we can always make the diagonal term equals to 1.

The correlation $\gamma_{jk}$ between the two nodes $j$ and $k$ is given by

$$\gamma_{jk} = \frac{\sigma_{jk}}{\sqrt{\sigma_{jj}\sigma_{kk}}}. $$

This is the population correlation, which can be estimated as

$$\hat{\gamma}_{ik} = \frac{x'_jx_k}{\sqrt{x'_jx_jx'_kx_k}}.$$ 

Here we abused the notation and did not use the hat notation $\hat{}$ for denoting the estimator. Since $E x_jx_j = 1$, it is reasonable to assume $x'_jx = 1$ so that

$$\gamma_{ik} = x'_jx_k.$$ 

The correlation $\gamma_{jk}$ can be computed from a linear regression \[24\]. Consider the linear model

$$x_j = \alpha_{jk}x_k + \epsilon_i$$

that links the measurements at the node $j$ to $k$. The parameters $\alpha_{jk}$ are estimated by minimizing the sum of squared residuals:

$$\| x_j - \alpha_{jk}x_k \|^2 = \alpha_{jk}^2 x'_jx_k - 2\alpha_{jk}x'_jx_k + x'_jx_j = \alpha_{jk}^2 - 2\alpha_{jk}x'_jx_k + 1$$

Here we used the notation $\| x_j \|^2 = x'_jx_j$.

By differentiating with respect to $\alpha_{jk}$ and letting the derivative equals to zero, the minimum is obtained when

$$\alpha_{jk} = \frac{x'_jx_k}{x'_kx_k} = \sqrt{\frac{x'_jx_j}{x'_kx_k}} \gamma_{ij} = \gamma_{ij} = x'_jx_k.$$ 

Hence, for centered scaled data matrix $X$, the correlation is simply the parameter in the linear model.
2 Sparse Correlations

There is a serious problem with the least squares estimation framework in the large-\(p\) and small-\(n\) setting. Since \(n \ll p\), this is a significantly underdetermined system. This is also related to the covariance matrix \(\Sigma\) being singular so we cannot just invert the covariance matrix in (4). One way of making the estimated covariance nonsingular is to sparsify the off-diagonal entries in the covariance matrix using LASSO. Following [24], we will formulate a sparse correlation estimation method. The sparse correlation is estimated by solving

\[
\hat{\alpha}_{jk} = \frac{1}{2} \arg \min_{\alpha_{jk}} \sum_{j=1}^{p} \sum_{k \neq j} \| x_j - \alpha_{jk} x_k \|^2 + \lambda \sum_{j=1}^{p} |\alpha_{jk}|. \quad (1)
\]

The coordinate descent learning and the active-set algorithm can be used to solve these types of LASSO [34, 11]. However, it can be shown that sparse correlation in (1) is equivalent to simply thresholding correlation at a certain level. Let

\[
L(\alpha) = \frac{1}{2} \sum_{j=1}^{p} \sum_{k \neq j} \| x_j - \alpha_{jk} x_k \|^2 = \frac{1}{2} \sum_{j=1}^{p} \sum_{k \neq j} (\alpha_{jk}^2 - 2\alpha_{jk} x_j^t x_k + 1).
\]

Note that

\[
\frac{\partial L}{\partial \alpha_{jk}} = \alpha_{jk} - x_j^t x_k.
\]

Let

\[
F = L(\alpha) + \lambda \sum_{j=1}^{p} |\alpha_{jk}|.
\]

The minimum of \(F\) is then achieved when

\[
0 = \frac{\partial F}{\partial \alpha_{jk}} = \alpha_{jk} - x_j^t x_k \pm \lambda.
\]

The sign of \(\lambda\) depends on the sign of \(\alpha_{jk}\). Then for \(\lambda \geq 0\), the sparse correlation estimation is given by

\[
\hat{\gamma}_{jk}(\lambda) = \begin{cases} 
   x_j^t x_k - \lambda & \text{if } x_j^t x_k > \lambda \\
   x_j^t x_k + \lambda & \text{if } x_j^t x_k < -\lambda \\
   0 & \text{otherwise}
\end{cases}
\] \quad (2)
Due to this simple expression, there is no need to optimize (1) numerically using the coordinate descent learning or the active-set algorithm often used in compressed sensing [34, 11]. Based on the expression (2), it can be shown that sparse correlation obtained in (1) is equivalently obtained by simple correlation thresholding.

To simplify the argument, let just assume $\lambda \geq 0$. Let $A(\lambda) = (a_{ij})$ be the adjacency matrix given by

$$a_{jk}(\lambda) = \begin{cases} 1 & \text{if } \hat{\gamma}_{jk} \neq 0; \\ 0 & \text{otherwise}. \end{cases}$$

This is equivalent to the adjacency matrix $B = (b_{jk})$ defined as

$$b_{jk}(\lambda) = \begin{cases} 1 & \text{if } |x_j^* x_k| > \lambda; \\ 0 & \text{otherwise}. \end{cases}$$

The adjacency matrix $B$ is simply obtained by thresholding the sample correlations. Then the adjacency matrices $A$ and $B$ induce a graph $G(\lambda)$ consisting of $\kappa(\lambda)$ disconnected components

$$G(\lambda) = \bigcup_{t=1}^{\kappa(\lambda)} G_t \text{ with } G_t = (V_t(\lambda), E_t(\lambda)),$$

where the whole node and edge sets are partitioned into $V_t$ and $E_t$ respectively. No two nodes between the different partitions are connected. From (3), it is trivial to show the partitioned graphs are nested within each partition such that

$$G_t(\lambda_1) \supset G_t(\lambda_2) \supset G_t(\lambda_3) \supset \cdots$$

for $\lambda_1 \leq \lambda_2 \leq \lambda_3$.

Hence we have the persistent homological structure induced from the compressed sensing type of the form (1). Any computation involving (1) can be done without optimization in $O(p^2 \ln p)$ time by thresholding the sorted sample correlations sequentially. A similar but more restricted persistent homological structure can be also found in the sparse-likelihood method.

### 3 Partial Correlations

By thresholding the correlation, we can establish a link between two nodes. However, there is a problem with this simplistic approach in that it fails to explicitly factor out the confounding effect of other nodes. To remedy the problem,
partial correlations have been used in factoring out the dependency of other nodes [19, 27, 20, 21, 34].

If we denote the inverse covariance matrix as $\Sigma^{-1} = (\sigma^{-1})$, the partial correlation between the nodes $j$ and $k$ while factoring out the effect of all other nodes is given by

$$
\rho_{jk} = -\frac{\sigma_{jk}}{\sqrt{\sigma_{jj}\sigma_{kk}}}
$$

Equivalently, we can compute the partial correlation via a linear model as follows. Consider a linear model of correlating measurement at node $j$ to all other nodes:

$$
x_j = \sum_{k \neq j} \beta_{jk} x_k + \epsilon_k.
$$

(5)

The parameters $\beta_{jk}$ are estimated by minimizing the sum of squared residual of (5)

$$
L(\beta) = \sum_{j=1}^{p} \|x_j - \sum_{k \neq j} \hat{\beta}_{jk} x_k\|^2
$$

(6)

in a least squares fashion. If we denote the least squares estimator by $\hat{\beta}_{jk}$, the residuals are given by

$$
r_j = x_j - \sum_{k \neq j} \hat{\beta}_{jk} x_k.
$$

(7)

The partial correlation $\rho_{jk}$ is then obtained by computing the correlation between the residuals of the model fit (7) [19, 25, 34]:

$$
\rho_{jk} = \text{corr} (r_j, r_k).
$$

Then from [34], it can be shown that

$$
\beta_{jk} = -\frac{\sigma_{jk}}{\sigma_{jj}} = \sqrt{\frac{\sigma_{kk}}{\sigma_{jj}}} \rho_{jk}.
$$

Moreover we have

$$
\text{Var} \epsilon_k = \frac{1}{\sigma_{kk}},
$$

(8)

$$
\text{Cov}(\epsilon_j, \epsilon_k) = \frac{\sigma_{jk}}{\sigma_{jj} \sigma_{kk}}.
$$

(9)
The diagonal terms in $\Sigma^{-1}$ tend to be stable regardless of if $\Sigma^{-1}$ is singular and we can simply estimate them directly from (8).

The sparse version of partial correlation estimation is obtained through

$$\hat{\beta}_{jk} = \arg \min_{\beta_{jk}} \sum_{j=1}^{p} \| x_j - \sum_{k \neq j} \beta_{jk} x_k \|^2 + \lambda \sum_{j,k=1}^{p} |\beta_{jk}|. \quad (10)$$

For the optimization detail, see [24].

From (10), sparse network recovery can be done by estimating the regression parameters $\beta_{jk}$. This is the well known LASSO framework [37, 34, 24]. Define $L_1$-penalty as

$$J(\beta) = \sum_{j,k} |\beta_{j,k}|.$$  

The sparse estimation of $\beta$ is then given by

$$\min_{\beta} L(\beta) + \lambda J(\beta). \quad (11)$$

The larger the value of $\lambda$, more sparsity constraint we are enforcing. Note that (11) can be equivalently formulated as

$$\min_{\beta} L(\beta) \text{ subject to } J(\beta) \leq \epsilon$$

for some $\epsilon \geq 0$ [9]. This formulation is often used in compressed sensing [5, 17].

Since the response $x_j$ is correlated with $\sum_{k \neq j} \beta_{jk} x_j$, this is not a standard compressed sensing problem. Nevertheless, as a first order exploratory analysis, we can still treat the problem as a compressed sensing problem [34, 24].

### 3.1 Limitation of Sparse Partial Correlations

However, the compressed sensing framework has a serious computational bottleneck. For $n$ measurements over $p$ nodes, it is required that we solve a linear system with an extremely large $A$ matrix of size $np \times p^2$, so that the complexity of the problem increases by a factor of $p^3$! Consequently, for a large number of nodes, the problem immediately becomes almost intractable for a small computer. For example, for 1 million nodes, we have to compute 1 trillion possible pairwise relationships between nodes. One practical solution is to modify (5) so that the
measurement at node $j$ is represented more sparsely over some possible index set $S_j$:

$$x_j = \sum_{k \in S_j} \beta_{jk} x_j + \epsilon_i,$$

making the problem substantially smaller.

An alternate approach is to simply follow the *homotopy path*, which enables to add network links one by one with a very limited increase of computational complexity so we do not need to compute $\beta$ repeatedly from scratch [8, 35, 33]. The trajectory of the optimal solution $\beta$ in LASSO follows a piecewise linear path as we change $\lambda$. By tracing the linear path, we can substantially reduce the computational burden of reestimating $\beta$ when $\lambda$ changes.

### 4 Sparse Likelihood

Beyond sparse regression, others have proposed the likelihood methods. Suppose $y_i = (x_{i1}, \cdots, x_{ip})'$ follows multivariate normal with mean $\mu$ and covariance $\Sigma = (\sigma_{kl})$. Neglecting constant terms, the log-likelihood function $L$ of $y_i$ is given by

$$L(\mu, \Sigma) = \log \det \Sigma^{-1} - \frac{1}{n} \sum_{i=1}^{n} (y_i - \mu)' \Sigma^{-1} (y_i - \mu)$$

(12)

By maximizing the log-likelihood, we obtain the maximum likelihood estimate (MLE) of $\mu$ and $\Sigma$:

$$\hat{\mu} = \bar{y}_i = \frac{1}{n} \sum_{i=1}^{n} y_i$$

(13)

$$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y}_i)(y_i - \bar{y}_i)'.$$

(14)

Note MLE is biased and the unbiased estimator need to be normalized by $n - 1$ instead of $n$. But it will properly cancel out in computing correlation, which we are eventually after, so the issue of proper normalization is not important. For a notational convenience, suppose we center the Jacobian determinant such that

$$x_{ij} \leftarrow x_{ij} - \bar{y}_i.$$
We are basically subtracting the group mean from individual measurements. Then MLE (14) can be written in a more compact form

$$\hat{\Sigma} = \frac{1}{n} X'_{p \times n} X_{n \times p}. \quad (15)$$

However, there is a serious defect with MLE (15) here; namely the estimated covariance matrix $\hat{\Sigma}$ is positive definite only for $n \geq p$ [13, 36]. $X'_{p \times n} X_{n \times p}$ becomes rank deficient for $n < p$. In most imaging studies, there are more voxels than the number of subjects, i.e. $n < p$. Further, when $\hat{\Sigma}$ is singular, we can not compute the inverse of $\Sigma$, which is called the precision matrix and related to partial correlation coefficients [24].

This is the main reason the covariance matrix is rarely estimated over the whole brain region and instead researchers have been using massive univariate approaches in most anatomical studies. As done in all previous cortical thickness based connectivity studies [39, 40, 25, 19, 18], one may tempted to compute the pairwise covariance and correlation between two voxels, but the resulting covariance and correlation matrix estimation is not optimal and suffers the small-$n$ large-$p$ problem.

To remedy the small-$n$ and large-$p$ value problem, we propose to incorporate the sparsity constraint in the log-likelihood and regularizes the covariance matrix. The log-likelihood (12) can be written as a matrix form as

$$L(\Sigma) = \log \det \Sigma^{-1} - \text{tr} (\Sigma^{-1} \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y}_i)(y_i - \bar{y}_i)) \quad (16)$$

$$= \log \det \Sigma^{-1} - \text{tr} (\Sigma^{-1} S), \quad (17)$$

where

$$S = \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y}_i)'(y_i - \bar{y}_i)$$

is the sample covariance, which is unbiased. To make it unbiased, it needs to be normalized by $n - 1$ instead of $n$. We used the fact that the trace of a scalar value is equivalent to the scalar value itself and $\text{tr}(AB) = \text{tr}(BA)$ for matrices $A$ and $B$. If we incorporate the $L_1$-norm penalty to the log-likelihood, we have the penalized log-likelihood give by

$$L(\Sigma^{-1}) = \log \det \Sigma^{-1} - \text{tr} (\Sigma^{-1} S) - \lambda \|\Sigma^{-1}\|_1,$$
Figure 1: Left: Adjacency matrices obtained through GLASSO with increasing $\lambda$ values. The persistent homological structure is self-evident. As $\lambda$ increases, the matrices become more sparse and eventually will become an identity matrix. Right: A block diagonal matrix $D$ with $\kappa$ number of blocks such that $D = PAP^{-1}$ with a permutation matrix $P$. $D$ also shows the persistent structure. The persistent structure is permutation invariant. The block diagonal structure shows the nested persistent homological structure.

where $\|\cdot\|$ is the sum of the absolute values of the elements. The tuning parameter $\lambda > 0$ controls the sparsity of the off diagonal elements of the covariance matrix. Then we maximize $L$ over the space of all possible symmetric definite matrices:

$$\hat{\Sigma}^{-1} = \arg \max_{\Sigma > 0} \log \det \Sigma^{-1} - \text{tr}(\Sigma^{-1}S) - \lambda \|\Sigma^{-1}\|_1$$  \hspace{1cm} (18)

The relationship of (18) to LASSO framework is explored in [11]. This is a convex problem and we solved it using the graphical lasso (GLASSO) algorithm [3, 2, 11, 21]. We first estimate $\Sigma^{-1}$ and invert it to obtain the sparse version of the covariance matrix. Since the optimization is done over the space of positive definite symmetric matrices, the GLASSO estimate is expected to be invertible.

The construction of the persistent homological structure out of $\hat{\Sigma}^{-1}(\lambda)$ is similar to the sparse correlation case. Let $A(\lambda) = (a_{ij})$ be given by the adjacency matrix out of the representation (18) such that

$$a_{ij}(\lambda) = \begin{cases} 1 & \text{if } \hat{\sigma}_{ij} \neq 0; \\ 0 & \text{otherwise.} \end{cases}$$
Then the adjacency matrix $A$ induces a graph $G(\lambda)$ consisting of $\kappa(\lambda)$ disconnected components $G(\lambda) = \bigcup_{l=1}^{\kappa(\lambda)} (V_l(\lambda), A_l(\lambda))$. The node set $\{1, 2, \cdots, p\}$ is partitioned into the node set $V_l$ such that

$$\{1, 2, \cdots, p\} = \bigcup_{l=1}^{\kappa} V_l(\lambda).$$

No two nodes between the different partitions are connected. Then it can be shown that the partitioned node sets are nested within each partition such that

$$V_l(\lambda_1) \supset V_l(\lambda_2) \supset V_l(\lambda_3) \supset \cdots$$

(19) for $\lambda_1 \leq \lambda_2 \leq \lambda_3$ [21, 28]. The construction of the persistent structure (19) is fairly time consuming since we have to solve the sequence of GLASSO. However, it can be shown that the partitioned node sets $V_l$ can be obtained by simply thresholding of the sample covariance $S = (s_{ij})$ [28]. Let us define a new adjacency matrix $B(\lambda) = (b_{ij})$ as

$$b_{ij}(\lambda) = \begin{cases} 1 & \text{if } |\hat{s}_{ij}| > \lambda; \\ 0 & \text{otherwise}. \end{cases}$$

The adjacency matrix $B$ similarly induces the partitioned graph

$$G(\lambda) = \bigcup_{l=1}^{\kappa(\lambda)} (W_l(\lambda), B_l(\lambda)).$$

It is trivial to see

$$W_l(\lambda_1) \supset W_l(\lambda_2) \supset W_l(\lambda_3) \supset \cdots$$

for $\lambda_1 \leq \lambda_2 \leq \lambda_3$. It can be further shown that $V_l = W_l$. By the simple thresholding the sample covariance matrix, it is possible to construct the the persistent structure (19) in $O(p^2 \ln p)$ time again without the computational bottleneck associated with GLASSO. However, in this construction, the edge sets may not exhibit nestedness.

5 Experimental Results

We illustrate our methods with two examples. In Figure 1, we randomly simulated the data matrix $X_{5 \times 10}$ from the standard normal distribution. The sample
Figure 2: Jacobian determinant of deformation field are measured at 548 nodes along the human brain surface. The barcodes of the filtrations on the sample correlations and covariances show huge group separation between normal controls and post-institutionalized (PI) children.

The covariance matrix is then feed into GLASSO with different filtration values and the induced adjacency matrices show the nestedness. Theoretically only the partitioned node sets are expected to exhibit the nestedness.

Figure 2 shows the real world example based on the 3 Tesla magnetic resonance images (MRI) of 23 maltreated children who have been post-institutionalized (PI) in orphanages but later adopted to the families in US and age-matched 31 normal control subjects. The Jacobian determinants of warping individual images to the template were computed at 548 positions along the white matter brain surface [16]. The nested node sets $V_i$ are constructed by simple thresholding on correlations and covariances. For 547 level of filtrations, the sequence of GLASSO would take more than 54 hours in a laptop (6min. per GLASSO). But it took less than 1 min using the simple thresholding method to obtain the barcodes.

**References**


