Reprinted from

Chung, M.K. 2019. Brain Network Analysis.
Cambridge University Press

4
Correlation Networks

Correlation based networks are probably the most often used network model in brain imaging. In this chapter, we study Pearson's product-moment correlation (Fisher, 1915), in short *Pearson correlation*. In sciences, Person correlation has been widely used as a simple index for measuring dependency and the linear relationship between two variables. In human brain mapping research, it has been mainly used to map out functional or anatomical connectivity (Friston et al., 1993a; Worsley et al., 2005b; Chung et al., 2015a).

4.1 Pearson correlations

Definition 4.1 *Consider two data vectors* $\mathbf{x} = (x_1, x_2, \cdots, x_n)^\top$ *and* $\mathbf{y} = (y_1, y_2, \cdots, y_n)^\top$. *The Pearson* correlation *coefficient* ρ *between two vectors* $\mathbf{x}$ *and* $\mathbf{y}$ *is defined as*

$$\rho(\mathbf{x}, \mathbf{y}) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}, \tag{4.1}$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$ *and* $\bar{y} = \frac{1}{n}\sum_{i=1}^n y_i$ *are the sample means.*

Then algebraic manipulation can show that

$$\rho(\mathbf{x}, \mathbf{y}) = \rho(a\mathbf{x} + b, c\mathbf{y} + d) \tag{4.2}$$

for any nonzero $a, c \in \mathbb{R}$ and any $b, d \in \mathbb{R}$. Thus, the correlation is scale and translation invariant. The correlation (4.1) can be factored as a vector product:

$$\rho(\mathbf{x}, \mathbf{y}) = [\mathbf{x}']^\top \mathbf{y}', \tag{4.3}$$

where $\mathbf{x}' = \alpha\mathbf{x} + \beta$ and $\mathbf{y}' = \gamma\mathbf{y} + \delta$ such that

$$\alpha = \frac{1}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}, \quad \beta = -\frac{\bar{x}}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}},$$

$$\gamma = \frac{1}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}, \quad \delta = -\frac{\bar{y}}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}.$$

Then trivially, correlations under *any* nontrivial linear transformation, i.e., $a, c \neq 0$, can be represented as vector products due to the invariance (4.2).

Among all linear transformations, following linear transformation is the most often used in relation to sparse models (Chung et al., 2015a, 2017a). Consider a transformation that center and scale $\mathbf{x}$ and $\mathbf{y}$ such that

$$\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i = 0,$$

$$\|\mathbf{x}\|^2 = \mathbf{x}^\top\mathbf{x} = \|\mathbf{y}\|^2 = \mathbf{y}^\top\mathbf{y} = 1. \tag{4.4}$$

This projects n-dimensional vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ onto a n-dimensional unit sphere S^{n-1}. Thus the centering and scaling operations can be viewed as data embedding from higher Euclidean space $\mathbb{R}^n$ onto unit sphere S^{n-1}.

Then the correlation between $\mathbf{x}$ and $\mathbf{y}$ is simply the *cosine similarity* often used in engineering and computer science literature:

$$\cos\theta = \mathbf{x}^\top\mathbf{y},$$

where θ is the angle between vectors $\mathbf{x}$ and $\mathbf{y}$.

Assuming $\mathbf{x}$ and $\mathbf{y}$ are centered and scaled, consider the following linear model

$$\mathbf{y} = \beta\mathbf{x} + \mathbf{e},$$

where $\mathbf{e}$ is a mean zero error vector and β is unknown scalar parameter we need to estimate. The least squares estimation (LSE) of β is given by minimizing the sum of the squared residuals:

$$\mathbf{e}^\top\mathbf{e} = (\mathbf{y} - \beta\mathbf{x})^\top(\mathbf{y} - \beta\mathbf{x}).$$

Trivially we can show that LSE of β is

$$\widehat{\beta} = \mathbf{x}^\top\mathbf{y},$$

the Pearson correlation. Even if $\mathbf{x}$ and $\mathbf{y}$ are not centered and scaled, there is a relationship between correlations and the regression coefficients.

4.2 Partial correlations

Let $Y = (Y_1, Y_2)$ be two variables of interests and $X = (X_1, \cdots, X_p)$ be a row vector of variables that should be removed in a data analysis. For instance, we may let Y_1 and Y_2 be functional activity at two different voxels, and X_1 and X_2 be the age and gender. The covariance matrix of $(Y, X)^\top$ is denoted by

$$\mathbb{V}(Y, X)^\top = \begin{pmatrix} \Sigma_{YY} & \Sigma_{YX} \\ \Sigma_{XY} & \Sigma_{XX} \end{pmatrix} \tag{4.5}$$

Note Σ_{XY} is the cross-covariance matrix of X and Y. Σ_{YX}, Σ_{XX} and Σ_{YY} are defined similarly. Then the partial covariance of Y given X is

$$(\sigma_{ij}) = \Sigma_{YY} - \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}.$$

The *partial correlation* $\rho_{Y_i,Y_j|X}$ is the correlation between variables Y_i and Y_j while removing the effect of variables X and it is defined as

$$\rho_{Y_i,Y_j|X} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}}.$$

The *conditional* notation $|$ is used in defining the partial correlation since the partial correlation is equivalent to *conditional correlation* if

$$\mathbb{E}(Y|X) = a + BX$$

for some vector a and matrix B, which is true under the normality of data. This is the formulation we used to compute the partial correlation. If vector X consists of a single measurement, i.e., $X = X_1$, the partial correlation can be computed from the simple correlation via

$$\rho_{Y_1,Y_2|X} = \frac{\rho_{Y_1,Y_2} - \rho_{Y_1,X}\rho_{Y_2,X}}{\sqrt{(1-\rho_{Y_1,X}^2)(1-\rho_{Y_2,X}^2)}}.$$

The definition above assumes $\rho_{Y_1,X}$ and $\rho_{Y_2,X}$ are not exactly -1 or 1. The *sample partial correlation* $r_{Y_1,Y_2|x}$ is defined similarly by replacing the covariance with the sample covariance in (4.5).

The Matlab codes for computing the partial correlation is as follows. Let `rho` be the sample partial correlation between time series fMRI1 and fMRI2 while removing the effect of age (`age`) and gender (`gender`) effects. For n subjects in the group, all variables are row vectors of size $1 \times n$. Then we compute the partial correlation `rho` as

```
x=[age; gender];
y=[fMRI1; fMRI2];
```

```
a=cov([x;y]');
b=a(1:2,1:2)-a(1:2,3:4)*inv(a(3:4,3:4))*a(3:4,1:2);
rho=b(1,2)/sqrt(b(1,1)*b(2,2));
```

Here `x` and `y` are $2 \times n$ matrices, and the covariance matrix `a` is the size 4×4.

4.3 Averaging correlations

In brain imaging, it frequently happens that the addition is not a well defined concept. Correlation is such an example. Given two correlations r_1 and r_2, it is possible to have $r_1 + r_2 < 1$ or $r_1 + r_2 > 1$, Thus, the sum of correlations may not be correlation. Subsequent, the average of correlation may not be a well defined concept. In this section, we show how to properly average correlations and correlation matrices. We start with vector space.

4.3.1 Vector spaces

Definition 4.2 *A vector space $\mathcal{S}$ is a collection of objects satisfying various axioms of algebraic rules. Given $x, y, z \in \mathcal{S}$, we need to have*

$$\begin{aligned} x + y = y + x &\in \mathcal{S} \ \textit{(commutative)}, \\ x + (y + z) = (x + y) + z &\in \mathcal{S} \ \textit{(associative)}. \end{aligned}$$

It also requires to have identity $0 \in \mathcal{S}$ such that

$$0 + x = x$$

and inverse $-x \in \mathcal{S}$ such that

$$x + (-x) = 0.$$

Given $a, b \in F$, a field,

$$\begin{aligned} a(bx) &= (ab)x \ \ \textit{(compatibility)}, \\ 1x &= x \ \ \textit{for some } 1 \in F \textit{ (identity)} \\ a(x + y) &= ax + ay \ \ \textit{(distributivity w.r.t. vector addition)} \\ (a + b)x &= ax + bx \ \ \textit{(distributivity w.r.t. field addition)} \end{aligned}$$

Given any $x, y \in \mathcal{S}$ and $a, b \in \mathbb{R}$, if $ax + by \in \mathcal{S}$, $\mathcal{S}$ is most likely a vector space in practice. Given two correlations r_1 and r_2, it is possible to have $r_1 + r_2 < -1$ or $r_1 + r_2 > 1$. Thus, correlations do not form a vector space.

Definition 4.3 *Given objects* $f_1, \cdots, f_n \in \mathcal{S}$ *for some space* S*, the (sample) mean* $\bar{f}$ *of the objects is defined as*

$$\bar{f} = \frac{1}{n} \sum_{j=1}^{n} f_j. \tag{4.6}$$

Thus, the sample mean is only defined in a vector space. A space that is not a vector space may not have the properly defined sample mean. To properly do a statistical inference, it is a necessity to have a vector space at least.

In the Euclidean space $\mathcal{S}$, the sample mean is the minimizer of the following cost function.

$$\bar{f} = \arg\min_{g \in \mathcal{S}} \sum_{j=1}^{n} \|g - f_j\|_2^2,$$

where $\| \cdot \|_2$ is the L_2-norm. This is easily proved by noting that the cost function is quadratic in norm $\|g\|_2$. By differentiating the cost function with respect to $\|g\|_2$ and setting the differentiation equals zero, we obtain the minimum.

4.3.2 Averaging by back projection

The average $\bar{f}$ may not belong to $\mathcal{S}$ if $\mathcal{S}$ is not a vector space. Thus, to define average, $\mathcal{S}$ must be a vector space. If $\mathcal{S}$ is not a vector space, we transform $\mathcal{S}$ to vector space $\mathcal{T}$ using some nonlinear transform

$$\mathcal{F} : \mathcal{S} \to \mathcal{T}.$$

We assume the inverse of $\mathcal{F}$ is well defined and easy to compute. $\mathcal{F}$ has to be one-to-one to make any sense in the operations below.

Given any $f_1, \cdots, f_n \in \mathcal{S}$, we have

$$g_j = \mathcal{F}(f_j) \in \mathcal{S}.$$

Then the average in $\mathcal{T}$ is defined as

$$\bar{g} = \frac{1}{n} \sum_{j=1}^{n} g_j = \frac{1}{n} \sum_{j=1}^{n} \mathcal{F}(f_i).$$

This new point $\bar{g}$ can be back projected into $\mathcal{F}$ via $\mathcal{F}^{-1}(\bar{g})$. Thus the reasonable average in $\mathcal{S}$ is defined as

$$\bar{f} = \mathcal{F}^{-1}\Big[\frac{1}{n} \sum_{j=1}^{n} \mathcal{F}(f_j)\Big].$$

The back projection method can be used to average correlations. The sample mean of correlations is not a well defined concept. Over the years, a number of different methods for averaging correlations have been proposed. Silver and Hollingsworth (1989) proposed to transform correlations into z-scores by Fisher's transform (Fisher, 1915):

$$F(\rho) = \text{arctanh}(\rho) = \frac{1}{2} \ln \frac{1+\rho}{1-\rho}.$$

Note the inverse of the Fisher's transform is given by

$$F^{-1}(z) = \frac{e^{2z}-1}{e^{2z}+1}.$$

Given a collection of correlations $\rho_1, \rho_2, \cdots, \rho_n$, the average of Fisher's transforms is

$$z = \frac{1}{n} \sum_{i=1}^{n} F(\rho_i).$$

Then $\mathcal{F}^{-1}(z)$ is the average of correlations via the backprojection. This approach is known to introduce bias (Corey et al., 1998). Another slightly less obvious approach is to use the log-Euclidean distance.

4.3.3 Metric spaces

Diffusion tensor images can produce 3D unit vectors that measures the principle direction of of diffusion of water molecules at each voxel. Given n unit vectors $\mathbf{v}_1, \cdots, \mathbf{v}_n$ with $\|\mathbf{v}_j\| = 1$, the average diffusion direction is not well defined in the following sense. Consider collection $S^2$of all the unit vectors. Obvious, $\mathbf{v}_1, \cdots, \mathbf{v}_n \in S^2$. However, $\bar{\mathbf{v}} = \sum_{j=1}^{n} \mathbf{v}_j/n \notin S^2$. $\mathbf{v}_1, \cdots, \mathbf{v}_n$ are points along the 3D unit sphere. $\bar{\mathbf{v}}$ is inside the the 3D unit solid ball. To properly define averaging operation in this example, we need a concept called Fréchet mean, which is defined in metric spaces. The metric spaces were introduced by Fréchet in 1906 as a part of his PhD thesis. Fréchet axiomatized the notion of distance and showed that many abstract spaces are metric spaces simplifying various difficult problems to that of metric properties.

Definition 4.4 *A* metric space *is a collection of objects for which the pairwise distance between objects is well defined. The distances are called* metric. *Formally,* $(\mathcal{M}, d)$ *is a metric space if* d *satisfies the condition*

$$d(x,y) = 0 \leftrightarrow x = y. \textit{ (identity)}$$
$$d(x,y) = d(y,x) \textit{ (symmetry)}$$
$$d(x,z) \le d(x,y) + d(y,z) \textit{ (triangle inequality)}$$

Then from the three axioms of metric, it can be shown that metric is always nonnegative, i.e.,

$$d(x, y) \geq 0.$$

This is left as an exercise.

There are numerous metric spaces. The usual Euclidean space $\mathbb{R}^n$ is a metric space with L_2-norm

$$d(x, y) = \|x - y\|_2 = \Big[\sum_{i=1}^{n} (x_i - y_i)^2\Big]^{1/2},$$

where $x = (x_1, \cdots, x_n)^\top$ and $y = (y_1, \cdots, y_n)^\top$

In another example, let $\mathcal{M}$ be a collection of $n \times n$ matrices that represent a graph with n nodes. Given $W^1 = (w^1_{ij})$, $W^2 = (w^2_{ij}) \in \mathcal{M}$, define L_l-distance as

$$D_l(W^1, W^2) = \| w^1 - w^2 \|_l = \Big(\sum_{i,j} |w^1_{ij} - w^2_{ij}|^l\Big)^{1/l}.$$

When $l = \infty$, L_∞-distance is written as

$$D_\infty(W^1, W^2) = \| w^1 - w^2 \|_\infty = \max_{\forall i,j} |w^1_{ij} - w^2_{ij}|.$$

Then we can show that $(\mathcal{M}, D_l)$ and $(\mathcal{M}, D_\infty)$ are metric spaces. This is left as an exercise.

The element-wise matrix distances differences (10.1) and (10.1) may not necessarly the best distance for matrices. L_1 and L_2-distances usually surfer the problem of outliers. Few outlying extreme edge weights may severely affect the distance. Further, these distances ignore the underlying topological structures. There exists a more topologically sensitive network distances (Chung et al., 2017a,d), which we will study later.

4.3.4 Fréchet mean

Generalizing the idea in (4.7), we define the Fréchet mean of $f_1, \cdots, f_n \in \mathcal{M}$ with metric d as

$$\bar{f} = \arg\min_{f \in \mathcal{M}} \sum_{j=1}^{n} d(f, f_j)^2.$$

In the Euclidean space, the sample mean and the Fréchet mean are identical. If $\mathcal{M}$ is not a vector space, the Fréchet mean may not be the sample mean.

On the sphere, the shortest distance between any two points $\mathbf{v}$ and $\mathbf{v}_j$ is the shortest arc in of the greatest circle passing through the two points. The arclength is the angle

$$\theta_j = \cos^{-1}\left(\mathbf{v}_j^\top \mathbf{v}\right).$$

We can show that the arclength is a metric. Then, the Fréchet mean of diffusion direction is given by

$$\bar{\theta} = \min_{\mathbf{v} \in S^2} \sum_{j=1}^{n} \Big[\cos^{-1}\left(\mathbf{v}_j^\top \mathbf{v}\right)\Big]^2.$$

The numerical implementation of computing the Fréchet mean on a sphere is left as an exercise. This is not a trivial problem on the sphere since it is mathematically not possible to have uniform grids on the sphere. Also the Fréchet mean may not be unique for some pathological example on S^2.

Fréchet mean can be used to average the covariance and correlation matrices for brain network modeling. (Qiu et al., 2015) is the first paper that averaged the collection of brain networks using the Fréchet mean. This requires defining a metric in the space of symmetric positive definite matrices, which is not a trivial problem. The metric is given by the log-Euclidean distance (Arsigny et al., 2005, 2006, 2007) which we will study later. The log-Euclidean framework is often used in averaging diffusion tensor images.

4.3.5 Log-Euclidean distance

The concept of Fréchet mean can be used to average correlation or covariance matrices.

Definition 4.5 *Given symmetric and positive definite matrix $C = (c_{ij})$, its* matrix exponential *is defined as a Taylor expansion:*

$$e^C = I + C + \frac{1}{2!}C^2 + \frac{1}{3!}C^3 + \cdots,$$

where I is the identity matrix. If there exists matrix A satisfying $e^A = C$, then A is called the matrix logrithm *of C and denoted as* $\log C$.

For symmetric positive definite $p \times p$ matrix C, the logarithm of C can be computed as follows. Note C has p positive eigenvalues $\lambda_1, \cdots, \lambda_p$. There exists an orthogonal matrix Q such that

$$C = Q^\top D Q$$

with $D = diag(\lambda_1, \cdots, \lambda_p)$, the diagonal matrix consisting of entries $\lambda_1, \cdots, \lambda_p$. Then using the fact $(Q^\top DQ)^k = Q^\top D^k Q$,

$$e^C = I + Q^\top DQ + \frac{1}{2!}Q^\top D^2 Q + \frac{1}{3!}Q^\top D^3 Q + \cdots = Q^\top e^D Q.$$

Thus the exponential of C is $Q^\top e^D Q$. Similarly, the logarithm of C is $Q^\top \log DQ$.

If matrix C is nonnegative definite with zero eigenvalues, the matrix logarithm is *not* defined since $\log 0$ is not defined. Thus, we cannot apply logarithm directly to rank-deficient large correlation and covariance matrices obtained from small number of samples. One way of applying logarithm to nonnegative definite matrices is to make matrix C diagonally dominant by adding a diagonal matrix αI with suitable choice of relatively large α (Chan and Wood, 1997). Alternately, we can perform a graphical LASSO-type of sparse model and obtain the closet positive definite matrices (Qiu et al., 2015; Mazumder and Hastie, 2012)

Definition 4.6 *Given two symmetric and positive definite matrices C_1 and C_2, the* log-Euclidean *distance between C_1 and C_2 is given by*

$$d(C_1, C_2) = \| \log C_1 - \log C_2 \|_F,$$

where the Frobenius norm is defined as (Arsigny et al., 2005, 2006, 2007)

$$\|A\|_F = \sqrt{tr(A^\top A)}.$$

The log-Euclidean distance can be written differently as

$$d(C_1, C_2) = \Big[tr\,(\log C_1 - \log C_2)^2\Big]^{1/2}.$$

The log-Euclidean distance can be viewed as the generalized manifold version of Frobenius norm distance. Given a collection of correlation matrices $C_1, C_2, \cdots, C_n$, *log-Euclidean Fréchet mean* $\bar{C}$ is given by (Arsigny et al., 2007)

$$\bar{C} = \exp\Big(\frac{1}{n}\sum_{i=1}^{n} \log C_i\Big). \tag{4.7}$$

Since correlation is the off-diagonal entries of correlation matrices, the diagonal entries are the average of correlations in log-Euclidean sense.

Limitation of log-Euclidean distance. If a matrix is nonnegative definite with zero eigenvalues, the matrix logarithm is *not* defined since $\log 0$ is not defined. Thus, we cannot apply the logarithm directly to rank-deficient large correlation and covariance matrices obtained from data with small sample sizes relative to the number of nodes. One way of applying the logarithm to nonnegative definite matrices is to make the matrix diagonally dominant by adding a diagonal matrix αI with suitable choice of relatively large α (Chan and Wood, 1997). Alternately, we can perform a graphical LASSO-type of sparse model and obtain the closest positive definite matrices (Qiu et al., 2015; Mazumder and Hastie, 2012). Developing the log-Euclidean for general correlation matrices including nonnegative ones is beyond the scope of this paper. Note topological distances are applicable to nonnegative definite connectivity matrices.

4.4 Correlation as metric

Consider a node set $V = \{1, \dots, p\}$ and edge weights $\rho = (\rho_{ij})$, where ρ_{ij} is the weight between nodes i and j. The edge weights measure similarity or dissimilarity between nodes. The edge weights in most brain networks are usually given by some similarity measure between nodes (Lee et al., 2011a; Li et al., 2009; Mclntosh and Gonzalez-Lima, 1994; Newman and Watts, 1999; Song et al., 2005). Weighted network $X = (V, \rho)$ is formed by the pair of node set V and edge weights ρ. If X is a metric, network interpretation is straightforward.

We will show how to construct a metric using correlations.

Consider $n \times 1$ measurement vector $\mathbf{x}_j = (x_{1j}, \cdots, x_{nj})^\top$ on node j. Suppose we center and rescale the measurement $\mathbf{x}_j$ such that

$$\| \mathbf{x}_j \|^2 = \mathbf{x}_j' \mathbf{x}_j = \sum_{i=1}^{n} x_{ij}^2 = 1$$

and

$$\sum_{i=1}^{n} x_{ij} = 0.$$

Naturally, we are interested in using correlations or their simple functions as edge wights, i.e.,

$$\rho_{ij} = \mathbf{x}_i^\top \mathbf{x}_j \quad \text{or} \quad \rho_{ij} = 1 - \mathbf{x}_i^\top \mathbf{x}_j.$$

However, not every functions of correlations are metric.

Example 4.1 [1] *$\rho_{ij} = 1 - \mathbf{x}_i^\top \mathbf{x}_j$ is* not *a metric. Consider the following 3-node*

[1] The counter example is provided by Zhiwei Ma of University of Chicago.

counter example:

$$\mathbf{x}_i = (0, \frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}})^\top,$$
$$\mathbf{x}_j = (\frac{1}{\sqrt{2}}, 0, -\frac{1}{\sqrt{2}})^\top,$$
$$\mathbf{x}_k = (\frac{1}{\sqrt{6}}, \frac{1}{\sqrt{6}}, -\frac{2}{\sqrt{6}})^\top.$$

Then we have $\rho_{ij} > \rho_{ik} + \rho_{jk}$.

Then interesting methodological question is to identify minimum conditions that make a function of correlations a metric.

Theorem 4.1 *For centered and scaled data* $\mathbf{x}_1, \cdots, \mathbf{x}_p$*, let*

$$\rho_{ij} = cos^{-1}(\mathbf{x}_i^\top \mathbf{x}_j).$$

Then ρ_{ij} *is metric.*

Proof. On unit sphere S^{n-1}, the correlation between $\mathbf{x}_i$ and $\mathbf{x}_j$ is the cosine angle θ_{ij} between the two vectors, i.e.,

$$\mathbf{x}_i^\top \mathbf{x}_j = \cos \theta_{ij}.$$

The geodesic distance ρ between nodes $\mathbf{x}_i$ and $\mathbf{x}_j$ on the unit sphere is given by angle θ_{ij}:

$$\rho(\mathbf{x}_i, \mathbf{x}_j) = \cos^{-1}(\mathbf{x}_i^\top \mathbf{x}_j).$$

For nodes $\mathbf{x}_i, \mathbf{x}_j \in S^{n-1}$, there are two possible angles θ_{ij} and $2\pi - \theta_{ij}$ depending on if we measure the angles along the shortest arc or longest arc. We take the convention of using the smallest angle in defining θ_{ij}. With this convention,

$$\rho(\mathbf{x}_i, \mathbf{x}_j) \leq \pi.$$

Given three nodes $\mathbf{x}_i, \mathbf{x}_j$ and $\mathbf{x}_k$, which forms a spherical triangle, we then have spherical triangle inequality

$$\rho(\mathbf{x}_i, \mathbf{x}_j) \leq \rho(\mathbf{x}_i, \mathbf{x}_k) + \rho(\mathbf{x}_k, \mathbf{x}_j).$$

The proof to (4.8) is given in Reid and Szendròi (2005). Thus we proved ρ is a metric. □

Theorem 4.2 *For any metric* ρ_{ij}*,* $f(\rho_{ij})$ *is also a metric if* $f(0) = 0$ *and* $f(x)$ *is increasing and concave for* $x > 0$

The proof is given in Van Dijk et al. (2012). Such function f is called the *metric preserving function*.

Any power $[cos^{-1}(\mathbf{x}_i^\top \mathbf{x}_j)]^{1/m}$ for $m \geq 1$ is metric. When $m = 1$, we have the simplest possible metric $\rho(\mathbf{x}_i, \mathbf{x}_j) = \cos^{-1}(\mathbf{x}_i^\top \mathbf{x}_j)$, which obtains minimum 0 when $\mathbf{x}_i^\top \mathbf{x}_j = 1$ and maximum π when $\mathbf{x}_i^\top \mathbf{x}_j = -1$.

Theorem 4.3 *For any* $\mathbf{x}_1, \cdots, \mathbf{x}_p \in \mathbb{R}^n$,

$$\rho_{ij} = \big[1 - corr(\mathbf{x}_i, \mathbf{y}_j)\big]^{1/2}$$

is a metric, where $corr(\mathbf{x}_i, \mathbf{y}_j)$ *is the Pearson correlation.*

4.5 Statistical inference on correlations

Regardless of if we have Pearson correlations or partial correlations, the statistical inference can be done similarly.

4.5.1 Inference on one sample

Let $\rho(p)$ be correlation or partial correlation for each voxel p, we are interested in testing

$$H_0 : \rho(p) = \rho \quad \text{vs.} \quad H_1 : \rho(p) \neq \rho \tag{4.8}$$

for some fixed ρ. In the usual one sample inference, we simply test if correlation $\rho(p)$ each voxel is 0 or not. Inference type (4.8) is useful if only one sample is available or determining high correlation regions within the brain. There are many different ways for testing the above hypotheses. One widely used technique is to use the Fisher transform (Fisher, 1915) that transforms the sample correlation r into

$$F(r) = \text{arctanh}(r) = \frac{1}{2} \ln \frac{1+r}{1-r}.$$

Then for moderately large samples, $F(r)$ shows asymptotic normality:

$$F(r) \sim N\Big(\frac{1}{2} \ln \Big(\frac{1+\rho_k}{1-\rho_k}\Big), \frac{1}{n_k - 3}\Big).$$

The transform can be viewed as a variance stabilizing normalization process. Then the test statistic under null is

$$Z = \sqrt{n-3}[F(r) - F(\rho)] \sim N(0,1),$$

which is a standard normal distribution (Chung, 2007)

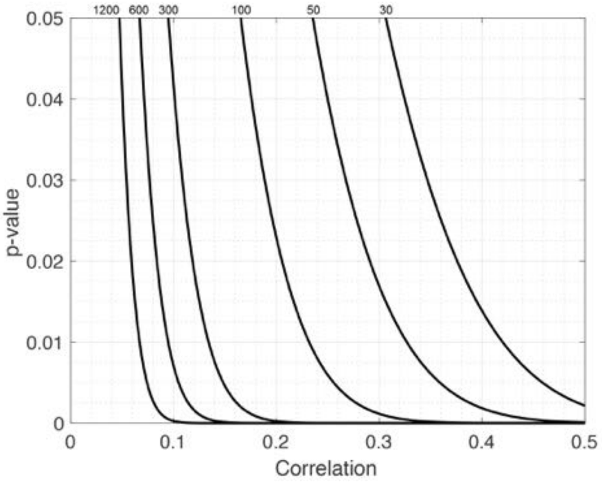

Figure 4.1 p-value over correlation in one sample. The numbers at the top are sample sizes. For large sample size of 1200, even correlation of 0.1 gives the statistical significance below 0.01. However, for small sample size of 30, correlation 0.5 does not give the statistical significance 0.01.

Another way of testing the significance is to use the t-statistic. Assuming the normality of data, the sample correlation or partial correlation r can be transformed to be distributed as:

$$T = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t_{n-2},$$

the t distribution with $n-2$ degrees of freedom. This test statistic can be used for testing hypothesis (4.8).

Here n is the sample size. If we are correlating time series with n time points, the the number of time points is the sample size.

4.5.2 Inference on two samples

Let $\rho_1(p)$ and $\rho_2(p)$ be two independent correlations at position p. We are interested in testing the equality of correlations. At each fixed point p, we are interested in testing

$$H_0 : \rho_1(p) = \rho_2(p) \text{ vs. } H_1 : \rho_1(p) \neq \rho_2(p). \quad (4.9)$$

For two sample inference type (4.9), the test statistic under H_0 is given by:

$$W(p) = \frac{\ln\left(\frac{1+r_1}{1-r_1} \cdot \frac{1-r_2}{1+r_2}\right)}{2\sqrt{\frac{1}{n_1-3} + \frac{1}{n_2-3}}} \sim N(0,1), \tag{4.10}$$

where r_1 and r_2 are the sample means that estimate $\rho_1(p)$ and $\rho_2(p)$ at fixed point p.

4.6 Cosine series representation

EEG at a channel and fMRI at a voxel are time series. Many time series analysis techniques have been applied in correlating two functional signals at different spatial locations. Since functional signals are highly noise, often Fourier transform type of signal filtering is applied before. Here we briefly explain the cosine series representation which is the analytic version of the often used cosine Fourier transform (Chung et al., 2010a). The cosine series representation can be viewed as a type of Fourier descriptor. Fourier descriptors have been around for many decades for modeling noise functional data and planar curves (Persoon and Fu, 1977; Staib and Duncan, 1992). They have been previously used to classify curves (Batchelor et al., 2006), where the Fourier coefficients are computed by the Fourier transform that involves both the sine and cosine series expansion. Then the sum of the squared coefficients are obtained up to certain degree for each functional data and the k-means clustering is used to classify the data. The cosine series representation differs from (Batchelor et al., 2006) in that we represent functional data employing cosine series only, without using both the cosine and sine series making the representation more compact. Matlab implementation for the cosine series representation is available [2].

4.6.1 Eigenfunctions of Laplacian in unit interval

There are infinitely many possible orthonormal basis in interval $[0,1]$. Here we explain the spectral approach for obtaining orthonormal basis.

Consider the space of square integrable functions in $[0,1]$ denoted by $\mathcal{L}^2[0,1]$. Let us solve the eigenequation

$$\Delta\psi + \lambda\psi = 0$$

[2] `brainimaging.waisman.wisc.edu/~chung/tracts`

in $\mathcal{L}^2[0,1]$ with 1D Laplacian $\Delta = \frac{d^2}{dt^2}$. Then it can be shown that the eigenfunctions $\psi_0, \psi_1, \cdots$ form an orthonormal basis in $\mathcal{L}^2[0,1]$. Note that if ψ_j is an eigenfunction, any multiple of ψ_j is also eigenfunction. Thus, it is expected the eigenfunctions are properly normalized. The eigenfunctions satisfying (8.2) is then given by the usual Fourier sine and cosine basis

$$\psi_0(t) = 1, \psi_l = \sqrt{2}\sin(l\pi t), \sqrt{2}\cos(l\pi t)$$

with the corresponding eigenvalues $\lambda_l = l^2\pi^2$.

Note that there are two eigenfunctions corresponding to the same eigenvalue. Note that the multiplicity of eigenfunctions only happen if there is a symmetry in the domain of the eigenvalue problem. The constant $\sqrt{2}$ is introduced to make the eigenfunctions orthonormal in $[0,1]$ with respect to the inner product

$$\langle f, g \rangle = \int_0^1 f(t)g(t)\, dt. \tag{4.11}$$

With respect to the inner product, the norm $\|\cdot\|$ is then defined as

$$\|f\| = \langle f, f \rangle^{1/2}.$$

Using both sine and cosine basis is not algebraically efficient. Instead of solving (8.2) in the domain $[0,1]$, consider solving the problem in the larger unbounded domain $\mathbb{R}$ with the periodic constraint

$$\psi(t+2) = \psi(t).$$

The period 2 constraint forces the basis function expansion to be only valid in the intervals $\cdots, [-2,-1], [0,1], [2,3], \cdots$ while there are gaps in $\cdots, (-1,0)$, $(1,2), (3,4), \cdots$. We can fill the gap by padding with some arbitrary function. However, if we pad the gaps with any function, it may result in the Gibbs phenomenon (ringing artifacts) at the boundary of the intervals $\cdots, 2, 1, 0, 1, 2, \cdots$ (Chung et al., 2007). To avoid the Gibbs phenomenon, we force the function to be continuous at the boundary by putting the constraint of evenness, i.e.,

$$\psi(t) = \psi(-t)$$

If $\psi(t)$ is the eigenfunction well defined in $[0,1]$, in the intervals, $\cdots, [-2,-1]$, $(-1,0)[0,1]$, $(1,2)$, $[2,3], \cdots$ we must have

$$\cdots, \psi(t-2), \psi(-t), \psi(t), \psi(-t+2), \psi(t+2), \cdots$$

The only eigenfunctions satisfying the two constraints (4.12) and (4.12) are the cosine basis

$$\psi_0(t) = 1, \psi_l(t) = \sqrt{2}\cos(l\pi t) \tag{4.12}$$

with the corresponding eigenvalues $\lambda_l = l^2\pi^2$ for integers $l > 0$. Then using the cosine basis only, any $f \in \mathcal{L}^2[0,1]$ can be represented as

$$f(t) = \sum_{l=0}^{k} c_l \psi_l(t) + \epsilon(t),$$

where c_l is the Fourier coefficients and ϵ is the residual error for using only k-th degree expansion.

Note, it is possible to put the constraint of oddness, i.e.,

$$\psi(t) = -\psi(-t).$$

Then we have sine basis

$$\psi_l(t) = \sqrt{2}\sin(l\pi t). \tag{4.13}$$

4.6.2 Fourier series

Consider i-th functional time series data

$$\zeta_i(t) = \mu_i(t) + \epsilon_i(t),$$

where t is time variable. We assume the functional data is scaled in such a way that they are defined in $[0,1]$. Formulating the Fourier analysis in a unit interval makes the numerical implementation more convenient. ϵ_i is a zero mean noise at each fixed t, i.e.,

$$\mathbb{E}\epsilon_i(t) = 0.$$

μ_i is an unknown smooth function to be estimated. It is reasonable to assume that

$$\zeta_i, \mu_i \in \mathcal{L}^2[0,1],$$

the space of square integrable functions. Any function $f \in \mathcal{L}^2[0,1]$ satisfies the condition

$$\int_0^1 f^2(t)\, dt < \infty.$$

This condition is needed to guarantee the convergence in the Fourier series.

Instead of estimating μ_i in $\mathcal{L}^2[0,1]$, we estimate it in a smaller subspace $\mathcal{H}_k$, which is spanned by up to the k orthonormal basis functions:

$$\mathcal{H}_k = \Big\{ \sum_{l=0}^{k} c_l \psi_l(t) : c_l \in \mathbb{R} \Big\} \subset \mathcal{L}^2[0,1].$$

Then the least squares estimation (LSE) of μ_i in $\mathcal{H}_k$ is given by

$$\widehat{\mu_i} = \arg\min_{f\in\mathcal{H}_k} \left\|f - \zeta_i(t)\right\|^2. \tag{4.14}$$

Theorem 4.4 *The minimization of (4.14) is given by*

$$\widehat{\mu_i} = \sum_{l=0}^{k} \langle \zeta_i, \psi_l\rangle \psi_l,$$

where the l-th degree Fourier coefficient $\langle \zeta_i, \psi_l\rangle$ *is given by the inner product*

$$\langle \zeta_i, \psi_l\rangle = \int_0^1 \zeta_i(t)\psi_l(t)\,dt. \tag{4.15}$$

Proof. Heuristically, we need to find function

$$f(t) = \sum_{l=0}^{k} c_l\psi_l(t)$$

that is the closest to ζ_i. The distance between f and ζ_i is given by

$$I(c_0, c_1, \cdots, c_k) = \int_0^1 \Big|\sum_{l=0}^{k} c_l\psi_l(t) - \zeta_i(t)\Big|^2\,dt,$$

which is a $k{+}1$ dimensional function in unknown parameter space $(c_0, c_1, \cdots, c_k) \in \mathbb{R}^{k+1}$. Since I is a quadratic function in $(c_0, c_1, \cdots, c_k)$, it has the global minimum at

$$\frac{\partial I}{\partial c_0} = \frac{\partial I}{\partial c_1} = \cdots = \frac{\partial I}{\partial c_k} = 0.$$

The algebraic derivation is left as an exercise but we have

$$c_l = \langle \zeta_i, \psi_l\rangle$$

for all $l = 0, 1, \cdots, k$. □

The expansion (4.15) is called the k-th degree *Fourier series*. As $k \to \infty$, the expansion converges to ζ_i, i.e.,

$$\zeta_i(t) = \sum_{l=0}^{\infty} \langle \zeta_i, \psi_l\rangle \psi_l.$$

It is also possible to have a slightly different but equivalent model that is easier to use in statistical inference. Assuming Gaussianness of data, $\epsilon_i(t)$ is a Gaussian stochastic process, which is simply a collection of random variables. Then $\epsilon_i(t)$ can be expanded using the given basis ψ_l as follows.

$$\epsilon_i(t) = \sum_{l=0}^{k} Z_l\psi_l(t) + e_i(t),$$

where $Z_l \sim N(0, \tau_l^2)$ are possibly *correlated* Gaussian random variables and e_i is the residual error that can be neglected in practice if enough number of basis are used. This is the consequence of the Karhunen-Loeve expansion (Adler, 1990; Dougherty, 1999; Kwapien and Woyczynski, 1992; Yaglom, 1987).

Karhunen-Loeve expansion states that $\epsilon_i(t)$ can be decomposed as

$$\epsilon_i(t) = \sum_{l=0}^{m} Z_l \phi_l(t)$$

for uncorrelated Gaussian random variables Z_l and some orthonormal basis $\phi_l(t)$. The algebraic determination of Z_l and $\phi_l(t)$ are left as an exercise. Since ϕ_l and ψ_l are different basis, ϕ_l can be represented as a linear combination of ψ_l. Rewriting ϕ_l in terms of ψ_l will make the Gaussian random variables Z_l correlated.

At the end, we can write model (4.14) as

$$\zeta_i(t) = \sum_{l=0}^{k} X_l \psi_l(t) + e_i(t), \tag{4.16}$$

where X_l are correlated Gaussian random variables.

4.6.3 Parameter estimation

In practice, functional time series are observed at discrete time points $t_1, t_2, \cdots, t_n$:

$$\zeta_i(t_j) = \mu_i(t_j) + \epsilon_i(t_j), \quad j = 1, \cdots, n.$$

The underlying mean functions $\mu_i(t)$ are estimated as

$$\widehat{\mu}_i(t) = \sum_{l=0}^{k} c_{li} \psi_l(t),$$

where the Fourier coefficients $(c_{0i}, c_{1i}, \cdots, c_{ki})$ for the i-th time series is estimated using LSE. If we have p number of time seriese, it requires p number of SLE separately for each time series. For really large LSE problems, this is computationally very inefficient. A more efficient way is to estimate the all the coefficients using a single LSE by solving the following large normal equation:

$$Y_{n\times p} = \Psi_{n\times k} C_{k\times p},$$

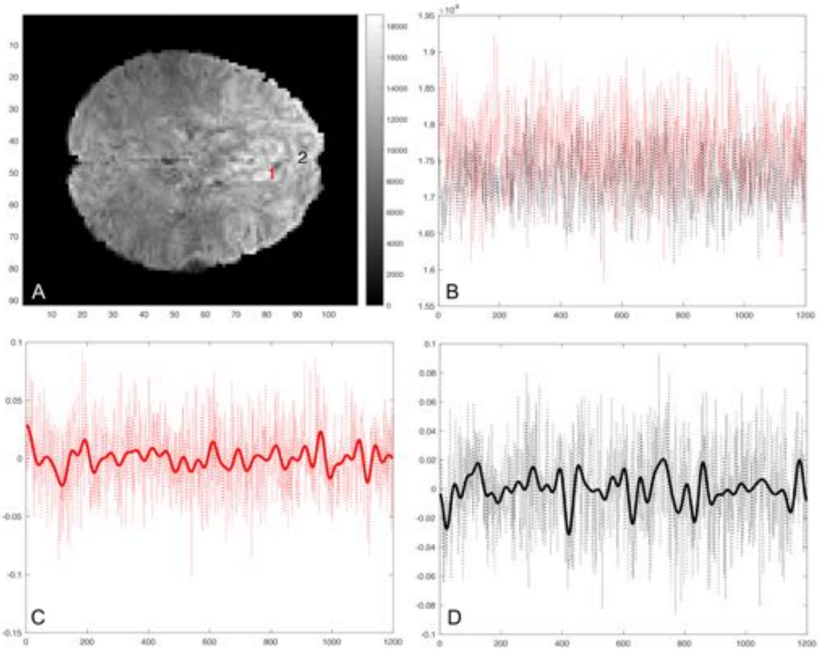

Figure 4.2 A. Resting-state fMRI at two different voxels at the first time point. B. Resting-state fMRI at voxel 1 (red) and 2 (black) shown for all 1200 time points. C. Normalized and scaled time series at voxel 1 and its cosine series representation with degree $k = 59$. D. Normalized and scaled time series at voxel 2 and its cosine series representation with degree $k = 59$. It is unclear what optimal degree we should use.

where

$$Y_{n\times p} = \begin{pmatrix} \zeta_1(t_1) & \zeta_2(t_1) & \cdots & \zeta_p(t_1) \\ \zeta_1(t_2) & \zeta_2(t_2) & \cdots & \zeta_p(t_2) \\ \vdots & \vdots & \ddots & \vdots \\ \zeta_1(t_n) & \zeta_2(t_n) & \cdots & \zeta_p(t_n) \end{pmatrix}, \tag{4.17}$$

$$\Psi_{n\times k} = \begin{pmatrix} \psi_0(t_1) & \psi_1(t_1) & \cdots & \psi_k(t_1) \\ \psi_0(t_2) & \psi_1(t_2) & \cdots & \psi_k(t_2) \\ \vdots & \vdots & \ddots & \vdots \\ \psi_0(t_n) & \psi_1(t_n) & \cdots & \psi_k(t_n) \end{pmatrix}, \tag{4.18}$$

$$C_{k\times p} = \begin{pmatrix} c_{01} & c_{02} & \cdots & c_{0p} \\ c_{11} & c_{12} & \cdots & c_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ c_{k1} & c_{k2} & \cdots & c_{kp} \end{pmatrix}. \tag{4.19}$$

Subsequently, the coefficients are simultaneously estimated in the least squares fashion as

$$\widehat{C} = (\Psi^\top \Psi)^{-1}\Psi^\top Y.$$

It is possible to discretize the basis ψ_l in such a way that $\Psi^\top \Psi = I_k$. This will make the numerical implementation of the Fourier series expansion computationally much more efficient for big data. The proposed least squares estimation technique avoids using the often used implicit Fourier transform (FT) (Batchelor et al., 2006; Bulow, 2004; Gu et al., 2004). The advantage of the cosine representation is that, instead of recording all the values of time series data, we only need to record $p \cdot (k+1)$ number of parameters. This is a substantial data reduction and we may be able to compress fMRI into 5% of the original data while suppressing high frequency noise (Figure 5.11).

Cosine series representation of fiber tracts. Cosine series representation can be also used in modeling white matter fiber tracts (Chung et al., 2010a). Unlike the nonparametric way of representing white matter fiber connectivity probabilistically, parametric methods can be used to model white matter fibers explicitly. Splines have also been often used for modeling and matching 3D curves (Clayden et al., 2007; Gruen and Akca, 2005; Kishon et al., 1990). Unfortunately, splines are not easy to model and to manipulate explicitly compared to Fourier descriptors, due to the introduction of internal knots. In Clayden et al. (2007), the cubic-B spline is used to parameterize the median of a set of tracts for tract dispersion modeling. Matching two splines with different numbers of knots is not computationally trivial and has been solved using a sequence of ad-hoc approaches. In Gruen and Akca (2005), the optimal displacement of two cubic spline curves are obtained by minimizing the sum of squared Euclidean distances. The minimization is nonlinear so an iterative updating scheme is used. On the other hand, there is no need for any numerical optimization in curve matching in Fourier descriptors due to the nature of the Hilbert space framework. Instead of using the squared distance of coordinates, others have used the curvature and torsion as features to be minimized to match curves (Corouge et al., 2004; Gueziec et al., 1997; Kishon et al., 1990; Leemans et al., 2006). Instead of applying the cosine series representation to functional time series, we apply to x-, y- and z-coordinates separately (Figure 4.3).

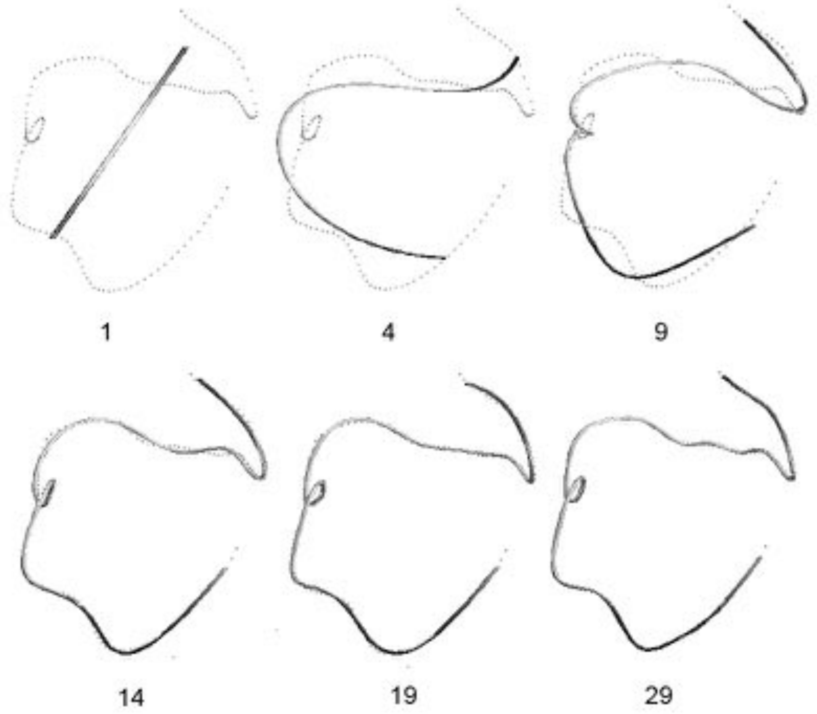

Figure 4.3 Cosine representation of a white matter fiber tract at various degrees. Dots are control points obtained from a streamline based tractography. The degree 1 representation is a straight line that fits all the control points in a least squares fashion. The degree 19 representation is used through the paper. It is unclaer what optimal degree we should use.

4.6.4 Stepwise model selection

One major problem in the cosine series representation is that the expansion has to be truncated at some degree. In Fourier descriptor and spherical harmonic representation literature, the issue of the optimal degree has not been addressed properly and the degree is simply selected based on a prespecified error bound (Bulow, 2004; Gerig et al., 2001; Gu et al., 2004; Shen and Chung, 2006; Shen et al., 2004). This model selection framework for Fourier descriptors was first presented in (Chung et al., 2007, 2008a). Although increasing the degree of the representation increases the goodness-of-fit, it also increases the number of estimated coefficients linearly. It is necessary to stop the series expansion at the degree where the goodness-of-fit and the number of coefficients balance out.

Suppose we have the k-th degree expansion of signal $\zeta_i(t)$:

$$\zeta_i(t) = \sum_{l=0}^{k} \widehat{c_{li}} \psi_l(t),$$

where $\widehat{c_{li}}$ are the least squares estimation. Assuming up to the $(k-1)$-degree representation is reasonably fitting data well, we determine if adding the k-degree term is statistically significant by testing

$$H_0 : c_{ki} = 0 \text{ vs. } H_1 : c_{ki} \neq 0.$$

Let the k-th degree *sum of squared errors* (SSE) be

$$\text{SSE}_k = \sum_{j=1}^{n} \Big[\zeta_i(t_j) - \sum_{l=0}^{k} \widehat{c_{li}} \psi_l(t_j) \Big]^2 .$$

As the degree k increases, SSE decreases until it flattens out. It is reasonable to stop the series expansion when the decrease in SSE is no longer statistically significant. We use the test statistic F given by

$$F = \frac{\text{SSE}_{k-1} - \text{SSE}_k}{\text{SSE}_{k-1}/(n-k-2)} \sim F_{1,n-k-2},$$

which is distributed as the F-distribution with 1 and $n-k-2$ degrees of freedom under H_0. We compute the F statistic at each degree and stop increasing the degree of expansion if the corresponding p-value first becomes bigger than the prespecified significance, which we can put at $\alpha = 0.01$ for instance.The forward model selection framework hierarchically builds the cosine series representation from lower to higher degree.

4.6.5 Distance between signals

It is often necessary to measure distance or similarity be- tween functions and time series. Using the cosine series representation, we can determine the optimal distance between the collections of functional tim series, which avoids brute-force style numerical optimization schemes often used in the functional data analysis field (Gruen and Akca, 2005; Gueziec et al., 1997; Kishon et al., 1990; Leemans et al., 2006; Ramsay and Silverman, 1997). This simplicity makes the cosine series representation more well suited than more often used splines or other signal filtering techniques (Gruen and Akca, 2005).

With the abuse of notations, we will interchangeably use functional signal to be estimated and their estimation with the same notations when the meaning is clear. Let the cosine series representation of two collections of time series $\boldsymbol{\eta}$ and $\boldsymbol{\zeta}$ be

$$\boldsymbol{\eta}(t) = \sum_{l=0}^{k} \boldsymbol{\eta}_l \psi_l(t),$$
$$\boldsymbol{\zeta}(t) = \sum_{l=0}^{k} \boldsymbol{\zeta}_l \psi_l(t)$$

where $\boldsymbol{\eta}_l$ and $\boldsymbol{\zeta}$ are the Fourier coefficient vectors. $\boldsymbol{\eta}$ and $\boldsymbol{\zeta}$ are collection of times series in a vectorial form.

Consider a displacement vector $\mathbf{u} = (u_1, u_2, \cdots, u_p)^\top$ that is required to register $\boldsymbol{\zeta}$ to $\boldsymbol{\eta}$ as close as possible. We will determine an optimal displacement $\mathbf{u}$ such that the distance between the deformed curve $\boldsymbol{\zeta} + \mathbf{u}$ and $\boldsymbol{\eta}$ is minimized with respect to a certain distance measure ρ. The distance ρ between $\boldsymbol{\eta}$ and $\boldsymbol{\zeta}$ is defined as the integral of the sum of squared distance:

$$\rho(\boldsymbol{\zeta}, \boldsymbol{\eta}) = \int_0^1 \|\boldsymbol{\zeta}(t) - \boldsymbol{\eta}(t)\|^2 \, dt.$$

The distance ρ can be simplified as

$$\begin{aligned}\rho(\boldsymbol{\zeta}, \boldsymbol{\eta}) &= \int_0^1 \sum_{j=1}^{p} \left[\sum_{l=0}^{k} (\zeta_{lj} - \eta_{lj}) \psi_l(t) \right]^2 \, dt \\ &= \sum_{j=1}^{p} \sum_{l=0}^{k} (\zeta_{lj} - \eta_{lj})^2.\end{aligned}$$

We have used the orthogonality condition

$$\int_0^1 \psi_l(t) \psi_m(t) \, dt = \delta_{lm}$$

to simplify the expression. It is left as an exercise to show $\rho(\boldsymbol{\zeta}, \boldsymbol{\eta})$ is a proper metric.

Theorem 4.5 *Let $\mathcal{H}_k = \{\sum_{l=0}^{k} c_l \psi_l(t) : c_l \in \mathbb{R}\}$ be the subspace spanned by up to k-th basis. Then we have*

$$\arg \min_{u_1, \cdots, u_p \in \mathcal{H}_k} \rho(\boldsymbol{\zeta} + \mathbf{u}, \boldsymbol{\eta}) = \sum_{l=0}^{k} (\boldsymbol{\eta}_l - \boldsymbol{\zeta}_l) \psi_l(t)$$

Proof. Let $\mathbf{u}^*(t)$ be the optimal displacement, which has a form

$$\mathbf{u}^*(t) = \sum_{l=0}^{k} \mathbf{u}_l \psi_l(t)$$

for some unknown parameter vector $\mathbf{u}_l = (u_{l1}, u_{l2}, \cdots, u_{lp})^\top$. Then

$$\rho(\boldsymbol{\zeta} + \mathbf{u}^*, \boldsymbol{\eta}) = \sum_{j=1}^{p} \sum_{l=0}^{k} (\zeta_{lj} + u_{lj} - \eta_{lj})^2,$$

which is an unconstrained positive definite quadratic program with respect to variables u_{lj}. The global minimum always exists and is obtained when $\rho(\boldsymbol{\zeta} + \mathbf{u}^*, \boldsymbol{\eta}) = 0$. Thus, we have $u_{lj} = \eta_{lj} - \zeta_{lj}$. □

The simplicity of Theorem 4.5 is that function registration is done by simply matching the corresponding Fourier coefficients without any sort of numerical optimization as in spline curve matching. In (4.20), the distance between two collections of time series is given as a function of degree k. Thus, the *multiscale distance* that captures both low and high frequency similarity can be given by

$$\sum_{k=0}^{K} \sum_{j=1}^{p} \sum_{l=0}^{k} (\zeta_{lj} - \eta_{lj})^2,$$

where K is the preselected degree.

White matter fiber tract registration and clustering. The method can be applied to linearly registering the fiber tracts and use it for clustering. The cosine series representation can be used to analyze a collection of fiber bundles consisting of similarly shaped curves. The ability to register one tract to another tract is necessary to establish anatomical correspondence for a subsequent population study. Since curves are represented as combinations of cosine functions, the registration will be formulated as a minimization problem in the subspace $\mathcal{H}_k$ which avoids brute-force style numerical optimization schemes (Gruen and Akca, 2005; Gueziec et al., 1997; Kishon et al., 1990; Leemans et al., 2006; Ramsay and Silverman, 1997). This simplicity makes the cosine series representation more well suited than the usual spline representation of curves in subsequent statistical analysis (Gruen and Akca, 2005).

4.6.6 Inference on a collection of functional signals

Based on the idea of computing distance between functional signals by matching coefficients, we can construct the average functional signals of p functional

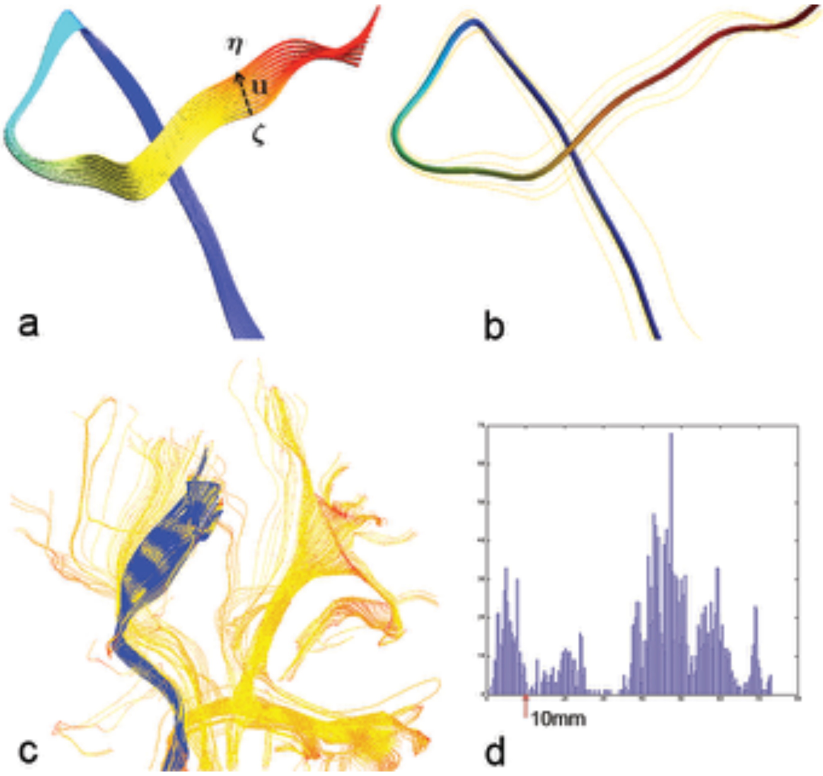

Figure 4.4 (a) the curve $\boldsymbol{\zeta}$ is registered to $\boldsymbol{\eta}$ by the displacement vector field $\mathbf{u}$, which is estimated in the least squares fashion using the cosine series representation. The other intermediate curves are generated by plotting $\boldsymbol{\zeta} + \alpha\mathbf{u}$ with $\alpha \in [0, 1]$ to show how the different amount of displacement deforms the curve $\boldsymbol{\zeta}$. (b) The average of a fiber bundle consisting of 5 tracts obtained by averaging the cosine coefficients. (c) Fiber tract clustering based on the distance between tracts. (d) The histogram of distances from a single tract (one of blue tracts) to all other tracts. By thresholding the histogram at 10mm, we cluster tracts.

signals $\boldsymbol{\zeta}^1, \cdots, \boldsymbol{\zeta}^p$ by finding the optimal function that minimizes the sum of all discrepancies in subspace $\mathcal{H}_k$:

$$\overline{\boldsymbol{\zeta}}(t) = \arg \min_{\zeta_1, \cdots, \zeta_p \in \mathcal{H}_k} \sum_{j=1}^{p} \rho(\boldsymbol{\zeta}^j, \boldsymbol{\zeta}).$$

The algebraic manipulation can show that the optimum signal is obtained by the average of representation:

$$\overline{\boldsymbol{\zeta}}(t) = \frac{1}{p}\sum_{j=1}^{p}\sum_{l=0}^{k}\boldsymbol{\zeta}_l^j\psi_l(t) = \sum_{l=0}^{k}\overline{\boldsymbol{\zeta}}_l\psi_l(t), \tag{4.20}$$

where $\overline{\boldsymbol{\zeta}}_l$ is the average coefficient vector

$$\overline{\boldsymbol{\zeta}}_l = \frac{1}{p}\sum_{j=1}^{p}\boldsymbol{\zeta}_l^j.$$

This simplicity is the consequence of the Fourier series having the best representation in the Hilbert space. Similarly we can define the sample variance of p signals and it will turn out to be the cosine representation with the coefficient vector consisting of the sample variance of p coefficients. The construction of the sample variance of p signals should be fairly straightforward and we will not go into detail.

Given another collection of functional signals $\boldsymbol{\eta}^1, \cdots, \boldsymbol{\eta}^q$, we can perform statistical inference on the equality of functional signals in the two populations. The null hypothesis of interest is

$$H_0 : \overline{\boldsymbol{\zeta}} = \overline{\boldsymbol{\eta}}. \tag{4.21}$$

Here we again abused the notation so we are testing the equality of mean representations of populations. From the very property of the Fourier series in Hilbert space, the uniqueness of the cosine series representation is guaranteed so the two representations are equal if and only if the coefficients vectors match. Therefore, the equivalent hypothesis to (4.21) is given by

$$H_0' : \overline{\boldsymbol{\zeta}}_1 = \overline{\boldsymbol{\eta}}_1, \cdots, \overline{\boldsymbol{\zeta}}_k = \overline{\boldsymbol{\eta}}_k.$$

Obviously this is a multiple comparisons problem. Under the Gaussian assumption in (4.16), testing the equality of the mean coefficient vector can be done using the Hotelling's T-square statistic. For correcting for the multiple comparisons, the Bonferroni correction can be used.

4.6.7 Gibbs phenomenon

Limitation of cosine series representation. A downside of using Fourier descriptors is that they are not local and it is not possible to make a statement about a specific portion of the functional signal. Although the Fourier coefficients are global and mainly used for globally classifying shapes (Shen et al., 2004), it is still possible to obtain local shape information and make a statement about local shape characteristics (Chung et al., 2007).

Splines have also been widely used for modeling functional signals (Clayden et al., 2007; Gruen and Akca, 2005; Kishon et al., 1990). Unfortunately, splines are not easy to model and to manipulate explicitly compared to Fourier descriptors, due to the introduction of internal knots. In Clayden et al. (Clayden et al., 2007; Corouge et al., 2004), the cubic-B spline is used to parameterize the median of a set of functions. Matching two splines with different numbers of knots is not computationally trivial and has been solved using a sequence of ad-hoc approaches. In Gruen et al. (Gruen and Akca, 2005), the optimal displacement of two cubic spline curves is obtained by minimizing the sum of squared Euclidean distances. The minimization is nonlinear so an iterative updating scheme is used. On the other hand, there is no need for any numerical optimization in obtaining the matching in our method due to the very nature of the Hilbert space framework. Instead of using the squared distance of coordinates, others have used the curvature and torsion as features to be minimized to match curves (Gueziec et al., 1997; Kishon et al., 1990; Leemans et al., 2006).

Gibbs phenomenon (ringing artifacts) often arises in Fourier series expansion of discontinuous data. It is named after American physicist Josiah Willard Gibbs. In representing piecewise continuously differentiable data using the Fourier series, the overshoot of the series happens at a jump discontinuity. The overshoot does not decease as the number of terms increases in the series expansion, and it converges to a finite limit called the Gibbs constant. The Gibbs phenomenon was first observed by Henry Willbraham in 1848 (Wilbraham, 1848) but it did not attract any attention at that time. Then a Nobel prize laureate Albert Michelson constructed an harmonic analyzer, one of the first mechanical analogue computers, that was used to plot Fourier series and observed the phenomenon. He thought the phenomenon was caused by mechanical error but Josiah Willard Gibbs correctly explained the phenomenon as mathematical in 1899. Josiah Willard Gibbs rediscovered the phenomenon in 1898 (Gibbs, 1898). Later mathematician Maxime Bocher named it the Gibbs phenomenon and gave a precise mathematical analysis in 1906 (Bocher, 1906). The Gibbs phenomenon associated with spherical harmonics were first observed by Herman Weyl in 1968. The history and the overview of Gibbs phenomenon can be found in the literature (Foster and Richards, 1991; Jerri, 1998).

There are few available techniques for reducing Gibbs phenomenon (Brezinski, 2004) (Gottlieb and Shu, 1997). Most techniques are a variation on some sort of kernel methods. For instance, consider Fejer kernel K_n defined as

$$K_n(u) = \frac{1}{n} \sum_{j=0}^{n-1} D_j(u),$$

where D_j is the Dirichlet kernel

$$D_j = \sum_{k=-j}^{j} e^{iku}.$$

Then it can be shown that

$$K_n(u) = \frac{1}{n}\Big(\frac{\sin\frac{nu}{2}}{\sin\frac{u}{2}}\Big)^2.$$

The kernel is symmetric and positive. Then it can be shown that

$$K_n * f \to f$$

for any, even discontinuous, $f \in \mathcal{L}^2[-\pi, \pi]$ as $n \to \infty$. It has the effect of smoothing the discontinuous signal f and in turn the convolution will not exhibit the ringing artifacts for sufficiently large n. Particularly related to Fourier and spherical harmonic descriptors, we have introduced an exponential weighting scheme (Chung et al., 2007, 2008a). By weighting Fourier coefficients with exponentially decaying weights, the series expansion can converge faster and reduce the Gibbs phenomenon significantly.

Instead of the k-th degree expansion (4.15), we define the weighted Fourier expansion as

$$\sum_{l=0}^{k} e^{-\lambda_l\sigma}\langle f, \psi_l\rangle\psi_l \tag{4.22}$$

for some smoothing parameter σ. Then it can be shown that (4.22) is the finite series expansion of heat kernel smoothing $K_\sigma * f$, where the heat kernel is defined as

$$K_\sigma(t, s) = \sum_{l=0}^{\infty} e^{-\lambda_l\sigma}\psi_l(t)\psi_l(s).$$

The expansion (4.22) can be further shown to be the finite approximation to the solution of heat diffusion

$$\frac{\partial}{\partial\sigma} g = \Delta g,\ g(t, \sigma = 0) = f(t).$$

Since the weighting scheme makes the expansion converges to heat diffusion, the estimation at the jump discontinuity is smoothed out reducing the Gibbs phenomenon.

4.7 Correlating functional signals

We can also use correlations to measure distance between functions. The concept of correlation here can be viewed as the generalization of Pearson correlation that is applied to discrete vector data to functional data. Consider functional signal

$$\zeta_1(t), \cdots, \zeta_p(t) \in L^2[0,1].$$

Definition 4.7 *The mean of functional signal ζ_i over time is given by*

$$\mathbb{E}_t \zeta_i = \int_0^1 \zeta_i(t)\, dt.$$

The cross-covariance *between functional signals ζ_i and ζ_j over interval $[0,1]$ is then given by*

$$\mathbb{V}_t(\zeta_i, \zeta_j) = \mathbb{E}_t\big[(\zeta_i - \mathbb{E}_t\zeta_i)(\zeta_j - \mathbb{E}_t\zeta_j)\big].$$

The variance *of ζ_i is*

$$\mathbb{V}_t\zeta_i = \mathbb{V}_t(\zeta_i, \zeta_i) = \int_0^1 \big[\zeta_i(t) - \mathbb{E}_t\zeta_i(t)\big]^2\, dt.$$

The cross correlation coefficient between functional signals ζ_i and ζ_j over interval $[0,1]$ is

$$\rho_{ij} = \frac{\mathbb{V}_t(\zeta_i, \zeta_j)}{\sqrt{\mathbb{V}_t(\zeta_i)\mathbb{V}_t(\zeta_j)}}.$$

Often we subtract $\mathbb{E}_t\zeta_i$ from functional signal ζ_i, i.e. $\zeta_i - \mathbb{E}_t$ such that ζ_i is centered. This is often done operation to normalize signals since subjects have different baseline average signal. From now on, we will simply assume $\mathbb{E}_t\zeta_i = 0$. If not, simply subtract by its mean. To reduce the confusion, let $\eta_i = \zeta_i - \mathbb{E}_t$ be the centered data of ζ_i. Consider the cosine series representation of centered η_i.

$$\eta_i = \sum_{l=0}^{k} c_{li}\psi_l(t) + e_i(t),\ t \in [0,1].$$

$e_i(t)$ is the residual function of the fit. If we use sufficiently large k, it is expected that e_i is small and ignorable.

The covariance of η_i and η_j is given by

$$\begin{aligned}\mathbb{V}_t(\eta_i,\eta_j) &= \int_0^1 \eta_i(t)\eta_j(t)\ dt\\ &\approx \sum_{l,m=0}^{k} c_{li}c_{mj}\int_0^1 \psi_l(t)\psi_m(t)\ dt\\ &= \sum_{l=0}^{k} c_{li}c_{lj} = \mathbf{c}_i^\top \mathbf{c}_j,\end{aligned}$$

where $\mathbf{c}_i = (c_{0i}, c_{1i}, \cdots, c_{ki})^\top$. The variance of η_i is then given by

$$\mathbb{V}_t\eta_i = \int_0^1 \eta_i^2(t)\ dt \approx \sum_{l=0}^{k} c_{li}^2 = \mathbf{c}_i^\top \mathbf{c}_i.$$

We used the fact that ψ_l are orthonormal, i.e.,

$$\int_0^1 \psi_l(t)\psi_m(t)\ dt = \delta_{lm}.$$

The cross-correlation between functional signals η_i and η_j is then given by

$$\rho_{ij} = \frac{\mathbf{c}_i^\top \mathbf{c}_j}{[\mathbf{c}_i^\top \mathbf{c}_i \mathbf{c}_j^\top \mathbf{c}_j]^{1/2}}.$$

If we let $\mathbf{b}_i = \mathbf{c}_i/\sqrt{\mathbf{c}_i^\top \mathbf{c}_i}$, the correlation is simplified as

$$\rho_{ij} = \mathbf{b}_i^\top \mathbf{b}_j.$$

Let $B_{k\times p} = [\mathbf{b}_1, \mathbf{b}_2, \cdots, \mathbf{b}_p]$ be the matrix of scaled Fourier coefficients. Then the correlation matrix $\rho = (\rho_{ij})$ is given by $\rho = B^\top B$. This will provide a faster but approximate computation of large-scale correlation matrices. Note that (4.23) is exact only if k is larger than the number of sampling points n. As k increases, the accuracy is expected to increase (Figure 4.5).

Example 4.2 *Consider fMRI time series η_1 and η_2 obtained in two voxels 1 and 2 in Figure 5.11 example. The Pearson correlation between η_1 and η_2 is -0.018. The baselines are $\mathbb{E}_t\eta_1 = 1763.0$ and $\mathbb{E}_t\eta_2 = 1724.8$. Now perform degree 59 cosine series expansion on the mean subtracted signals:*

$$\begin{aligned}\eta_1 - 1763.0 &= \sum_{l=0}^{59} c_{l1}\psi_l(t) + e_1(t)\\ \eta_2 - 1724.8 &= \sum_{l=0}^{59} c_{l2}\psi_l(t) + e_2(t),\end{aligned}$$

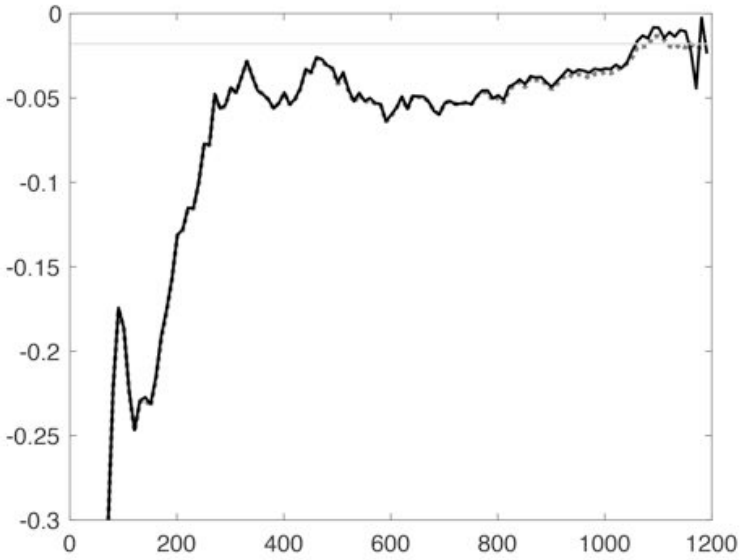

Figure 4.5 The straight line is the correlation between resting-state fMRI in two voxels. The dotted line is the correlation between the cosine series representations. The solid line is using the product of Fourier coefficients $\mathbf{b}_i^\top \mathbf{b}_j$. As the degree of expansion increases, they converge to each others.

where $e_1(t)$ and $e_2(t)$ are residual model fit. In this example,

$$\mathbb{E}_t e_1 = 3.05 \cdot 10^{-7}, \quad \mathbb{E}_t e_2 = 5.72 \cdot 10^{-6}.$$

The Pearson correlation becomes -0.3329.

4.8 Thresholding correlation networks

The majority of brain network analyses have been based on thresholding correlation edge weights in detecting focal regions of correlated voxels (Cao and Worsley, 1999b; Koch et al., 2002). On the other hand, Worsley et al. (2005a) used 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. Unless it is large sample size studies, we expect the *large p small n problem.* 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 Σ is:

$$\Sigma = UWU',$$

where U is an orthonormal matrix and W is a diagonal matrix of component weights.Worsley et al. (2005a) proposed to estimate Σ by setting the smaller weights in W to be zero. This is exactly the principal component analysis or partial least squares (PLS) (McIntosh et al., 1996; McIntosh and Lobaugh, 2004). 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 Σ can be prohibitive in small computers. (Worsley et al., 2005a) proposed to bypass the problem by matrix decompositions. Afterward, the statistical inference is done either using permutation tests (Nichols and Holmes, 2002) or the random field theory (Cao and Worsley, 1999b; Worsley et al., 1998). Since the brain networks are known to be sparse and highly clustered (Achard and Bullmore, 2007; He et al., 2007), 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 (Jolliffe et al., 2003; Zou et al., 2006). LASSO is a widely used variable selection technique that produces sparse models (Tibshirani, 1996). 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 connectivity analyses based on correlation or covariance matrices is that it fails to explicitly factor out the confounding effect of other regions. To remedy this limitation, partial correlation has been naturally introduced in factoring out the dependencies of other regions (He et al., 2007; Marrelec et al., 2006) or eliminating the effect of the experimental design (McIntosh et al., 1996). Since the partial correlation corresponds to the off-diagonal entries of the inverse covariance matrix, sparse PCA can be used to the inverse covariance matrix. A similar frameworks found applications in image classification (Berge et al., 2007), gene expression (Dobra et al., 2004), flow cytometry data (Friedman et al., 2008) and functional brain network model (Huang et al., 2009, 2010).