Hypothesis Testing For Multilayer Network Data

Jun Li

Dept of Mathematics and Statistics, Boston University

Joint work with Eric Kolaczyk

BOSTON UNIVERSITY

Outline

- Background and Motivation
- Geometric structure of multilayer networks
- Frechet mean and its general central limit theorem
- Central limit theorem for multilayer network
- Simulation Study
- Potential directions

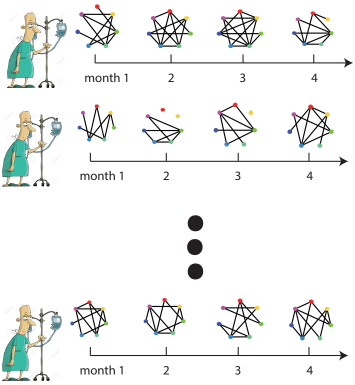
Background and Motivation

There is a trend to analyze large collections of networks. In recent work by our group[1], a formal notion of a space of unilayer network Graph Laplacians has been introduced and a central limit theorem has been developed based on it.

In many natural and engineered systems, collections of multiple networks best describe them, and multilayer network representations arise naturally.

[1] Ginestet, C. E., Balanchandran, P., Rosenberg, S., & Kolaczyk, E. D. (2014). Hypothesis Testing For Network Data in Functional Neuroimaging. arXiv preprint

Background and Motivation: A Motivating Example



n patients; d Regions of interest (ROI)

- Assume patients received treatment between the 2nd and the 3rd month: if the treatment has an effect.
- Assume we measured two samples from different populations: if there exists a difference between the two populations.

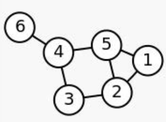
Outline

- Geometric structure of multilayer networks
- Frechet mean and its general central limit theorem
- Central limit theorem for multilayer network
- Simulation Study
- Potential directions

Geometric Structure of Multilayer Networks

Laplacian matrix and Supra-Laplacian

The Laplacian matrix of a network G is defined by $L = D - A$

Labeled graph	Degree matrix	Adjacency matrix	Laplacian matrix
	$\begin{pmatrix} 2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$	$\begin{pmatrix} 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$	$\begin{pmatrix} 2 & -1 & 0 & 0 & -1 & 0 \\ -1 & 3 & -1 & 0 & -1 & 0 \\ 0 & -1 & 2 & -1 & 0 & 0 \\ 0 & 0 & -1 & 3 & -1 & -1 \\ -1 & -1 & 0 & -1 & 3 & 0 \\ 0 & 0 & 0 & -1 & 0 & 1 \end{pmatrix}$

For a multilayer network $\mathcal{M}$, we can list all its nodes and treat it as a network $G_{\mathcal{M}}$. Then the Supra-Laplacian for $\mathcal{M}$ is defined the same as the Laplacian matrix for $G_{\mathcal{M}}$.

Geometric Structure of Multilayer Networks (cont.)

Geometric structure of unilayer networks' Graph Laplacians[2]:

Theorem 1 Let $\mathcal{L}_d$ be $d \times d$ matrices L satisfying:

- (1) Symmetry, $L^{'} = L$
- (2) The entries in each row sum to 0
- (3) The off-diagonal entries are non-positive, $e_{ij} \leq 0$
- (4) $Rank(L) = d - 1$

Then the matrix L should also satisfy:

- (5) Positive semi-definiteness, $L \geq 0$

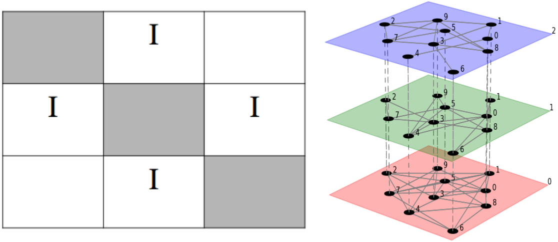
The matrices with these properties form a submanifold of $\mathbb{R}^{d^2}$ of dimension $\frac{d(d-1)}{2}$ with corners. In addition, $\mathcal{L}_d$ is a convex subset in $\mathbb{R}^{d^2}$.

[2] Ginestet, C. E., Balanchandran, P., Rosenberg, S., & Kolaczyk, E. D. (2014). Hypothesis Testing For Network Data in Functional Neuroimaging. arXiv preprint

Geometric Structure of Multilayer Networks (cont.)

Two classes of multilayer network Supra-Laplacian

Class 1:

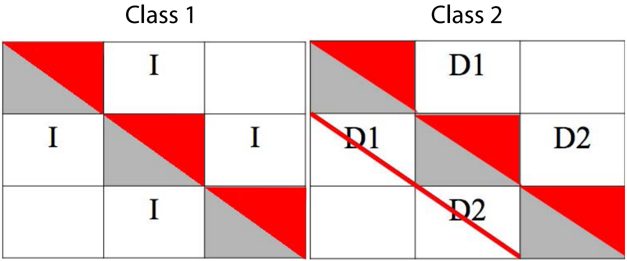


Class 2: Extend Class 1 Case by letting inter-layer links be any positive weights.

Geometric Structure of Multilayer Networks (cont.)

Geometric structure of Two classes of multilayer network Supra-Laplacian

Theorem 2 Class 1 Supra-Laplacians form a submanifold of $\mathbb{R}^{(nd)^2}$ of dimension $\frac{nd(d-1)}{2}$.



Class 2 Supra-Laplacians form a submanifold of $\mathbb{R}^{(nd)^2}$ of dimension $\frac{nd(d-1)}{2} + (n-1)d$.

Both of the submanifolds are convex subsets in $\mathbb{R}^{(nd)^2}$.

Outline

- Frechet mean and its general central limit theorem
- Central limit theorem for multilayer network
- Simulation Study
- Potential directions

Definition of Frechet mean

On a metric space (S, ρ) there is a notion of the mean μ of a distribution Q, as the minimizer of the expected squared distance from a point,

$$\mu = \underset{p}{\operatorname{argmin}} \int \rho^2(p, q) Q(dq)$$

assuming the integral is finite (for some p) and the minimizer is unique, in which case one says that the *Frechet mean of Q exists*.

Frechet Mean and its General Central Limit Theorem (cont.)

Theorem 3 On the metric space S, under some regularity conditions[3], we have the general CLT for Frechet mean:

$$n^{1/2}[J(\mu_n) - J(\mu)] \to N(0, \Lambda^{-1} C \Lambda^{-1}), as\ n \to \infty$$

Notations:

- h $x \to h(x; q) := \rho^2(J^{-1}(x), q)$
- μ_n the Frechet sample mean of the empirical distribution
- J a homeomorphism from a measurable subset of S to an open subset of $\mathbb{R}^s$
- C the covariance matrix of $\{D_r h(J(\mu); Y_1), r = 1, ..., s\}$
- Λ $[ED_{r,r'} h(J(\mu); Y_1)]_{r,r'=1,\ldots,s}$
- Y_i's i.i.d. S-valued random variables

[3] Bhattacharya, R., & Lin, L. (2013). An omnibus CLT for Fréchet means and nonparametric inference on non-Euclidean spaces. arXiv preprint

BOSTON UNIVERSITY

Outline

- Central limit theorem for multilayer network
- Simulation Study
- Potential directions

Central Limit Theorem for Multilayer Network

Theorem 4 Let $\mathcal{M}_1$, ..., $\mathcal{M}_n$ denote n multilayer networks and let L_1, ..., L_n be the corresponding Supra-Laplacians. $\hat{L}_n$ is their empirical mean. The L_i's are assumed to be independent and identically distributed according to a distribution Q.

If the expectation, $\Lambda := \mathbb{E}[L]$, does not lie on the boundary of $\mathcal{L}_d$, and $\mathbb{P}[U] > 0$, where U is an open subset of $\mathcal{L}_d$ with $\Lambda \in U$, and under the condition that each element of $L_i, i = 1, ...n$ has finite variance; we obtain the following convergence in distribution,

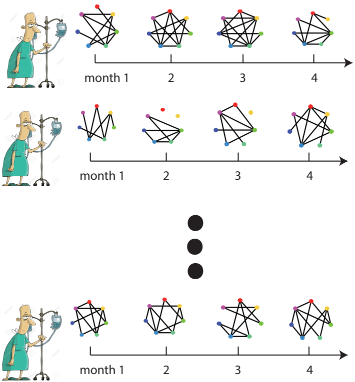
$$n^{1/2}(J(\hat{L}_n) - J(\Lambda)) \to N(0, \Sigma)$$

where $\Sigma := Cov[J(L)]$ and $J(\cdot)$ denotes the supra-half-vectorization of its matrix argument, that is, J aligns the upper diagonal of a symmetric matrix to vectorize it.

Outline

- Simulation Study
- Potential directions

Simulation Study: Motivating Example Review



n patients; d Regions of interest (ROI)

- Assume patients received treatment between the 2nd and the 3rd treatment: if the treatment has an effect.
- Assume we measured two samples from different populations: if there exists a difference between the two populations.

Simulation Study: Methods

1. Hypothesis Testing procedure based on central limit theorem only

- One-sample case: Each multilayer network (patient) could be spreaded into a $\frac{d(d-1)}{2} * T$ dimensional vector through the supra-half-vectorization J. Suppose the vectors are Y_1, ..., Y_n. Let $a = (1, 1, ..., 1, -1, ..., -1, -1)$ and $X_i = a \cdot Y_i$. H_0: X's distribution has 0 mean $\leftrightarrow$ H_1: X's distribution doesn't have 0 mean.

- Two-sample case: In this case, we have two sample of such vectors to represent patients: Y_{11}, ..., Y_{1n} and Y_{21}, ..., Y_{2n}. H_0:The two population have the same mean $\leftrightarrow$ H_1: The two population have different means.

  We can construct $T = (J(\hat{Y}_1) - J(\hat{Y}_2))^T \hat{\Sigma}^{-1} (J(\hat{Y}_1) - J(\hat{Y}_2))$ which has an asymptotic χ^2_m distribution under the null hypothesis, where $m = \binom{d}{2} * T$[4]

[4]Ginestet, C. E., Balanchandran, P., Rosenberg, S., & Kolaczyk, E. D. (2014). Hypothesis Testing For Network Data in Functional Neuroimaging. arXiv preprint

2. Hypothesis Testing procedure based on the bootstrap

In many situations, the bootstrap can be used to perform hypothesis tests that are more reliable in finite samples than tests based on asymptotic theory. If the bootstrap is to work well, the original test statistic must be asymptotically pivotal.[5]

In the one dimensional case, a two-sided test based on normal approximation has level error of order n^{-1}, which is reduced to order n^{-2} by using bootstrap with asymptotic pivotal statistic.[6] Fot the multi dimensional case, the level of error should be related to n and d. As shown in our simulation, the bootstrap leads to a higher power.

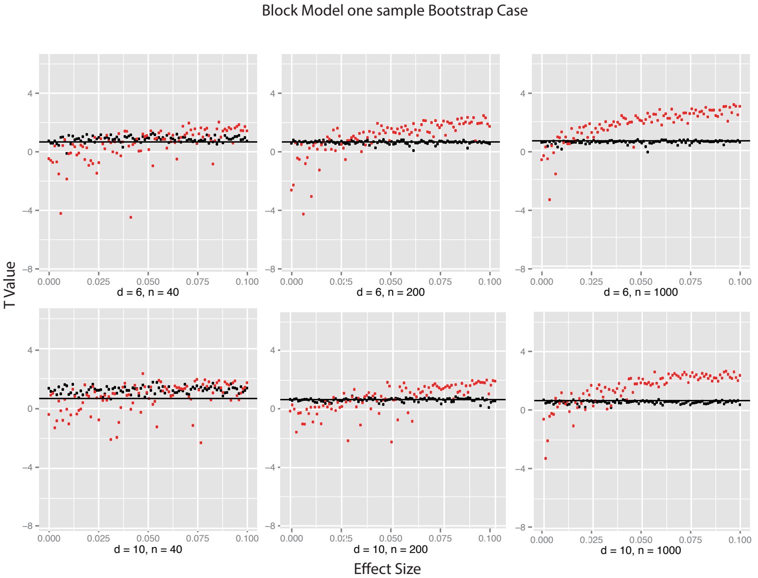
[5] Davidson, R., & MacKinnon, J. G. (1996). The power of bootstrap tests. Queens Institute for Economic Research Discussion Paper

[6] Hall, P. (2013). The bootstrap and Edgeworth expansion. Springer Science & Business Media.

2. Hypothesis Testing procedure based on bootstrap

- One-sample case: Test statistic is defined as $T_n = \| \sqrt{n}\frac{\bar{X}_n}{\hat{\sigma}} \|$. Bootstrap is from empirical distribution G_n based on $\{X_i - \bar{X}_n, i = 1, ..., n\}$. Here X's are scalars.
- Two-sample case: G_n is the same as that in the one sample case. The test statistic is defined as $T_n = \| \sqrt{n}\hat{\Sigma}^{-\frac{1}{2}}(\bar{X}_n^{(1)} - \bar{X}_n^{(2)}) \|$, where $\bar{X}_n^{(1)}$ and $\bar{X}_n^{(2)}$ indicate two sample mean. Here X's are vectors, $\hat{\Sigma}$ is the pooled sample covariance matrix.

Simulation Study: Results



Potential Directions

- Estimate rate of convergence of our central limit theorem for multilayer network as function of d and n.
- Power analysis: to know how many patients we need to measure, if we want to ensure power at a certain level. Needs the above convergence rate in the multivariate case.
- More applications based on this CLT on multilayer network, e.g. the regression on network.