Final report

Maoyu Zhang

Institute of Statistics and Big Data
Renmin University of China

05.08 2021

Overview

- Introduction
- Data information
- Model
- Result
- Discussion

Introduction

Two-sample tests for multivariate data and non-Euclidean data are widely used in many fields. Parametric tests are mostly restrained to certain types of data that meets the assump- tions of the parametric models. In this paper, we study a nonparametric testing procedure that uses graphs representing the similarity among observations. It can be applied to any data types as long as an informative similarity measure on the sample space can be defined. We summary the three test based on minimal spanning trees, and apply them to our two network datasets:

1. Two group brain networks of patients with and without Parkinson's disease;
2.Flight network data at U.S. airports in 2015.

Data information

- Brain network data sets with 81 Parkinson patients and 36 controls. There are 264 vertices representing 264 brain acupoints
- We have flight data for all airports in the United States in 2015. We select the 50 busiest airports as the vertices of the network data, and the number of flights between them is used as the weight of the edges of the network data, there are a total of 334 days of data, which means that we have 334 networks, of which there are 239 days on weekdays and 95 days on weekends. These data are reported by the U.S. Department of Transportation's (DOT) Bureau of Transportation Statistics, and can be download in https://www.kaggle.com/usdot/flight-delays

Data information

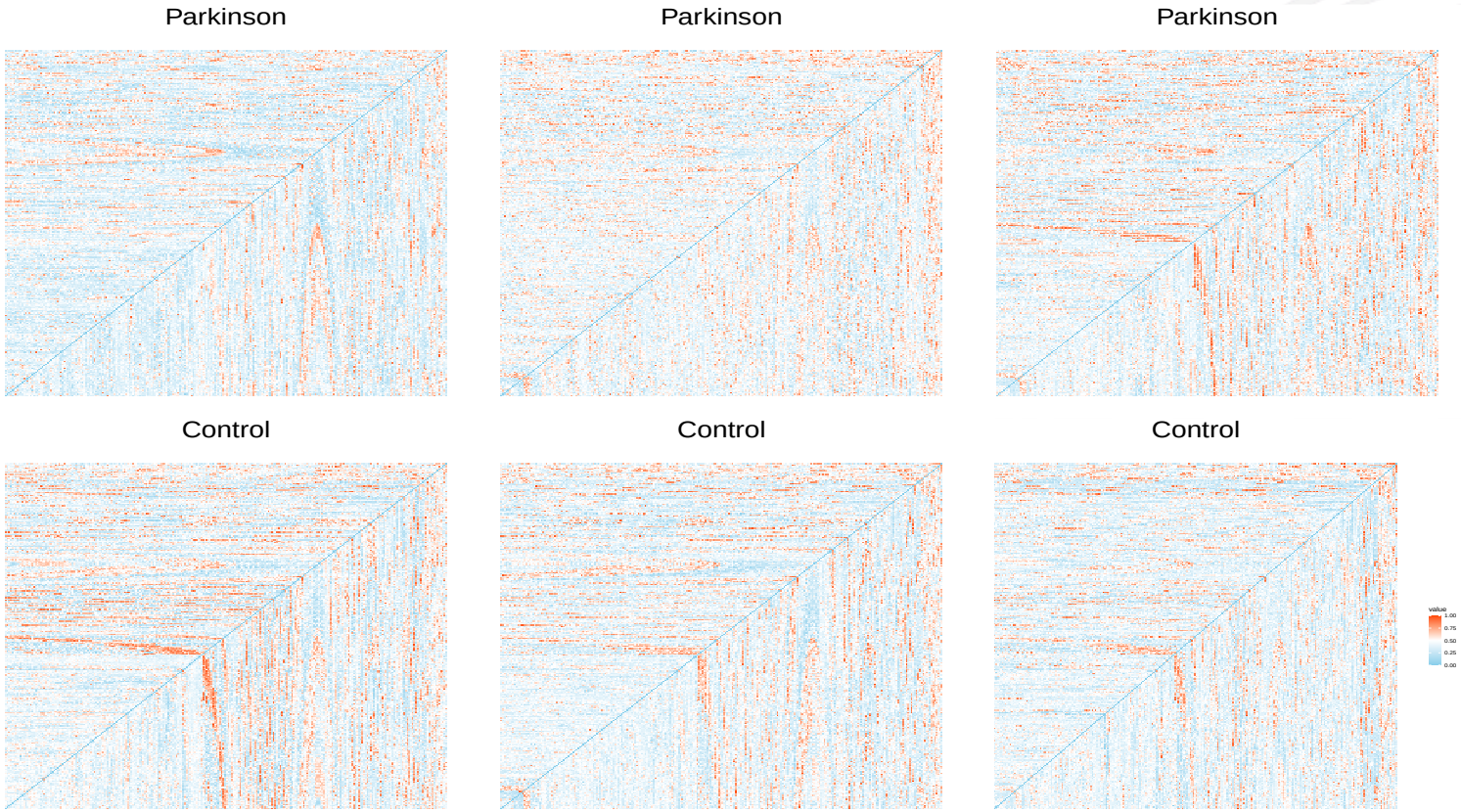


Figure: The heatmap of networks for Parkinson's patients and controls.

Data information

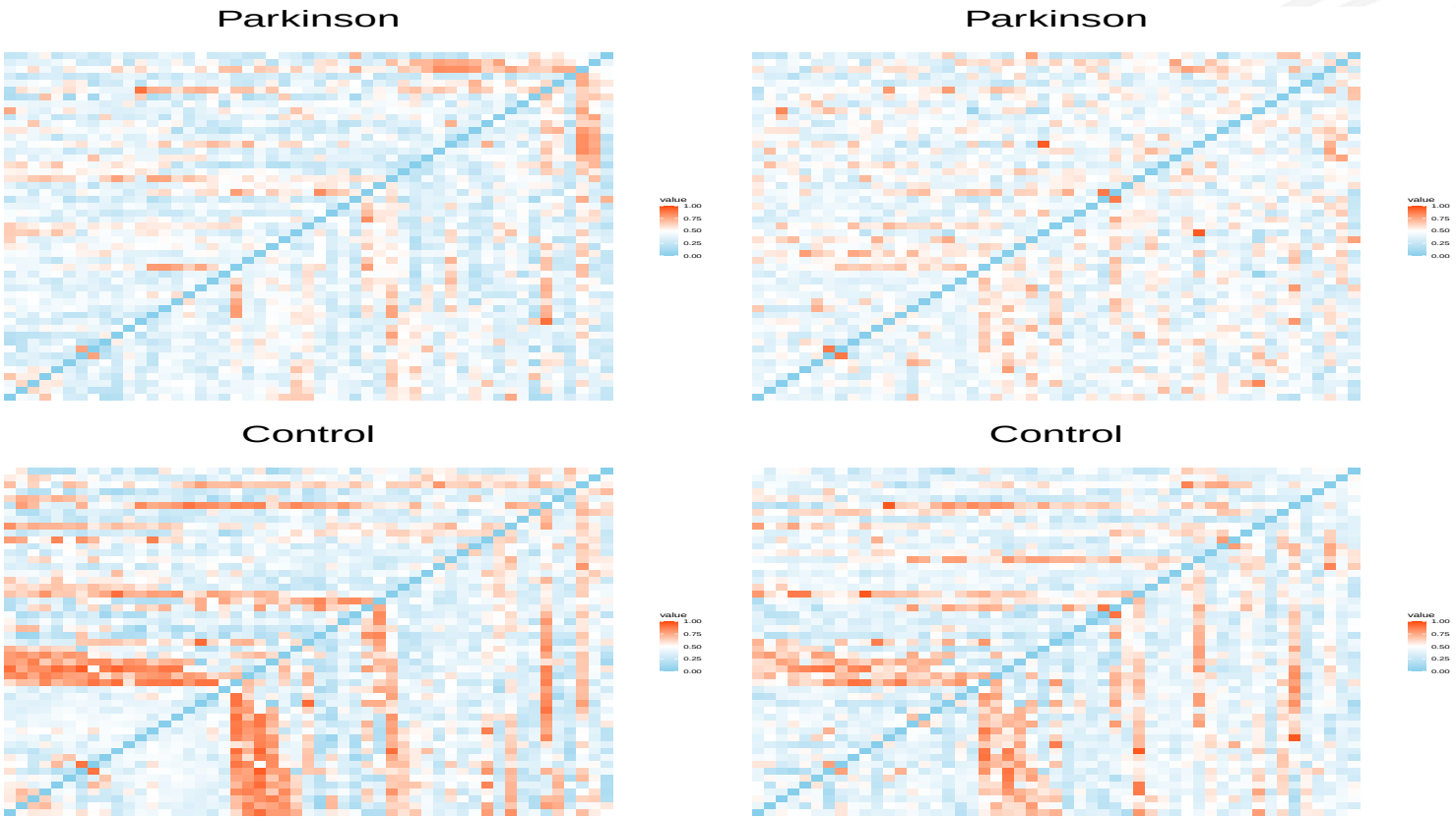
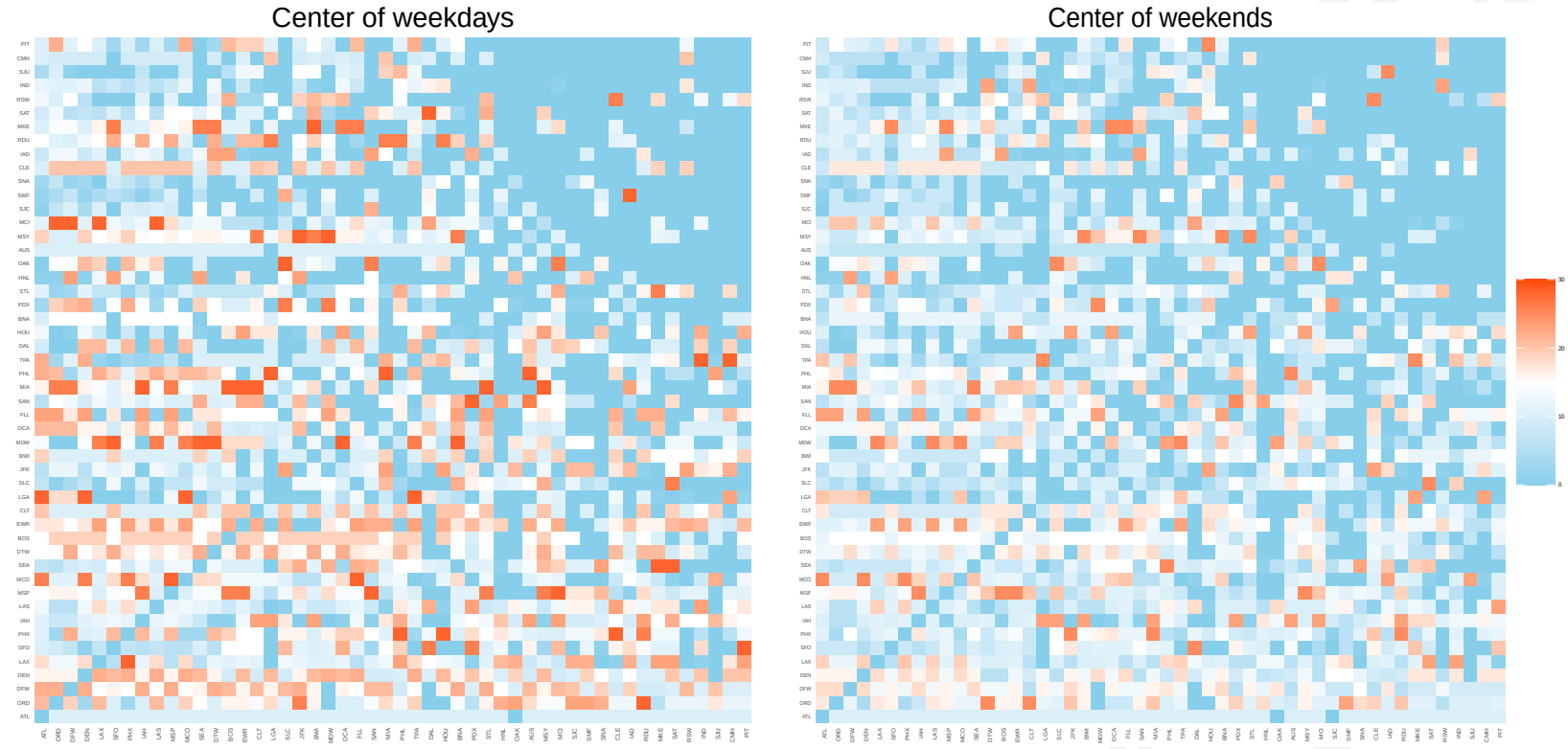


Figure: The heatmap of networks for Parkinson's patients and controls.

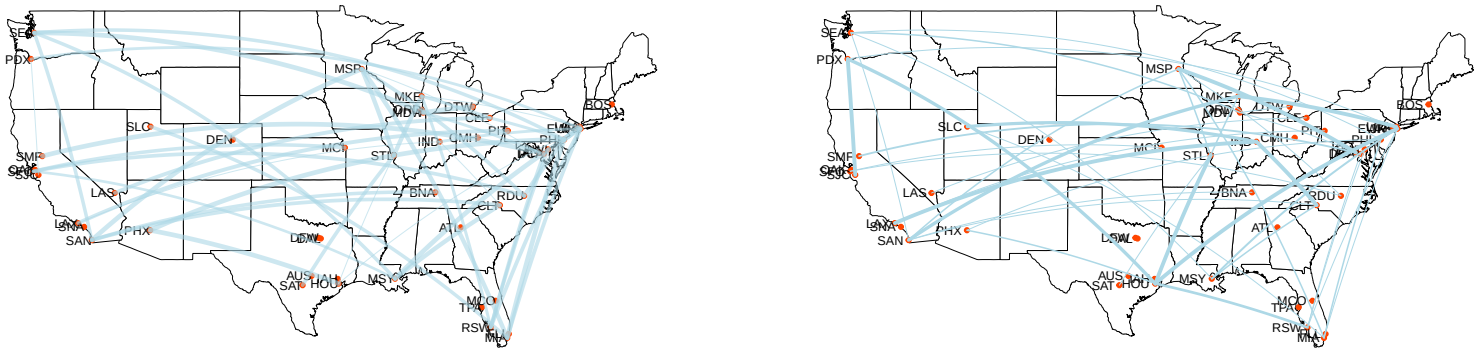
Data information



(a) center of weekdays (b) center of weekends

Figure: The heatmap of two networks from two groups.

Data information



(a) center of weekdays

(b) center of weekends

Figure: The flight map of two networks from two group, the routes shown are displayed by selecting some routes whose absolute value of the difference between two networks is greater than 15.

Model

Minimal spanning tree (MST)

- Edge-Count statistic (runs test)(1979)
- General statistic (2017)
- Weighted edge-Count statistic(2018)
- Cross-rank test
- Inter-rank test

Minimal spanning trees.

A spanning subgraph of a given graph is a subgraph with node set identical to the node set of the given sraph.

A spanning tree of a graph is a spanning subgraph that is a tree. Note that here is a (unique) path between every two nodes in a tree, and thus a spanning tree of a (connected) graph provides a path between every two nodes of the graph.

An edge weighted graph is a graph with a real number assigned to each edge. A minimal spanning tree (MST) of an edge weighted graph is a spanning tree for which the sum of edge weights is a minimum.

Multivariate number of runs test

Number the $N-1$ edges of the MST arbitrarily and define $Z_i, 1 \leqslant i \leqslant N-1$, as follows:

$$Z_i = \begin{cases} 1, \text{if the } i \text{ th edge links nodes from different samples.} \\ 0, \textit{otherwise}. \end{cases}$$

Then

$$R = \sum_{i=1}^{N-1} Z_i + 1 \quad \text{and} \quad E[R] = \sum_{i=1}^{N-1} E\left[Z_i\right] + 1$$

$$W = \frac{R - E[R]}{(\mathrm{Var}[R])^{\frac{1}{2}}}$$

The general statistic

Let $g_i = 0$ if the observation is from sample otherwise. For an edge $e = (i, j)$, we define:

$$J_e = \begin{cases} 0 & \text{if} \quad g_i \neq g_j \\ 1 & \text{if} \quad g_i = g_j = 0 \\ 2 & \text{if} \quad g_i = g_j = 1 \end{cases}$$

$$R_k = \sum_{e \in G} I_{J_e = k}, k = 0, 1, 2$$

Then R_0 is the number of between-sample edges (which is the test statistic for the edge-count test), R_1 is the number of edges connecting observations both from sample X, and R_2 is the number of edges connecting observations both from sample Y.

The general statistic

The new test statistic is defined as follows:

$$S = (R_1 - \mu_1, R_2 - \mu_2)\, \Sigma^{-1} \begin{pmatrix} R_1 - \mu_1 \\ R_2 - \mu_2 \end{pmatrix}$$

where $\mu_1 = \mathrm{E}(R_1), \mu_2 = \mathrm{E}(R_2)$, and Σ is the covariance matrix of the vector $(R_1, R_2)'$ under the permutation null distribution. Under the location-alternative, or the scale-alternative for lowdimensional data, we would expect both R_1 and R_2 to be larger than their null expectations, then S would be large. Under the scale-alternative for moderate/high-dimensional data, the number of within-sample edges for the sample with a smaller variance is expected to be larger than its null expectation, and the number of within-sample edges for the sample with a larger variance is expected to be smaller than its null expectation, then S would also be large. Therefore, the test defined in this way is sensitive to both location and scale alternatives.

Weighted edge-count test

we consider the following test statistic:

$$R_w = qR_1 + pR_2, \quad p = \frac{m}{N}, \quad q = 1 - p$$

where m is the sample size of $\mathcal{F}_1$, n is the sample size of $\mathcal{F}_2$, $m + n = N$. When the test statistic is defined in this way, its variance is well controlled no matter how different m and n are.

Cross-rank test based on MST

From the above three tests, we know that they are all based on the number of edges of the minimum spanning tree to construct statistics, and do not use the information of the distance to the minimum spanning tree. Based on the distance of minimal spanning tree, we propose a new test, Sort the distances of the $n-1$ edges of the minimum spanning tree, and sum the ranks of the between-sample edges to obtain a new statistic. Let d_{ij} is the distance of between $i-$th nodes and $j-$th node of the minimal spanning tree, $Q_{ij} = \text{ranks}(d_{ij})$, which is the rank for the edge connecting $i-$th nodes and $j-$th node, from samllest to highest, giving them a rank from 1 to n, then the new statistic is defined as:

$$W_0 = \sum_{i=1}^{n}\sum_{j=1}^{n} Q_{ij} I_{\{\text{i,j from the different sample}\}}$$

Inter-rank test based on MST

Based on the cross-rank test, we propose another new Inter-rank test, first, we define:

$$W_1 = \sum_{i=1}^{n}\sum_{j=1}^{n} Q_{ij} I_{\{\text{i,j from the sample X}\}}$$

$$W_2 = \sum_{i=1}^{n}\sum_{j=1}^{n} Q_{ij} I_{\{\text{i,j from the sample Y}\}}$$

the new statistic is defined as:

$$T = \left(W_1 - \mu_1^{'}, W_2 - \mu_2^{'}\right) \Sigma_1^{-1} \begin{pmatrix} W_1 - \mu_1^{'} \\ W_2 - \mu_2^{i} \end{pmatrix},$$

where $\mu_1^{'} = \mathrm{E}(W_1)$, $\mu_2^{'} = \mathrm{E}(W_2)$, Σ_1 is the covariance matrix of the vector $(W_1, W_2)'$ under the permutation null distribution.

Result

Table: Size, power and p-value of different methods

methods	t-test	1979	2017	2018	cross-rank test	inter-rank
Size	0.052	0.068	0.05	0.072	0.03	0.06
Power	0.98	0.95	0.83	0.94	0.82	0.8
p-value-park	0.2493	0.05	0.022	0.0030	0.0017	0.05
p-value-flight	$5.7 * 10^{-13}$	$3.4 * 10^{-14}$	$2.9 * 10^{-14}$	$1.6 * 10^{-15}$	$1.5 * 10^{-31}$	$3.7 * 10^{-10}$

Result

https://maoyuzhang.shinyapps.io/shiny/

Data information

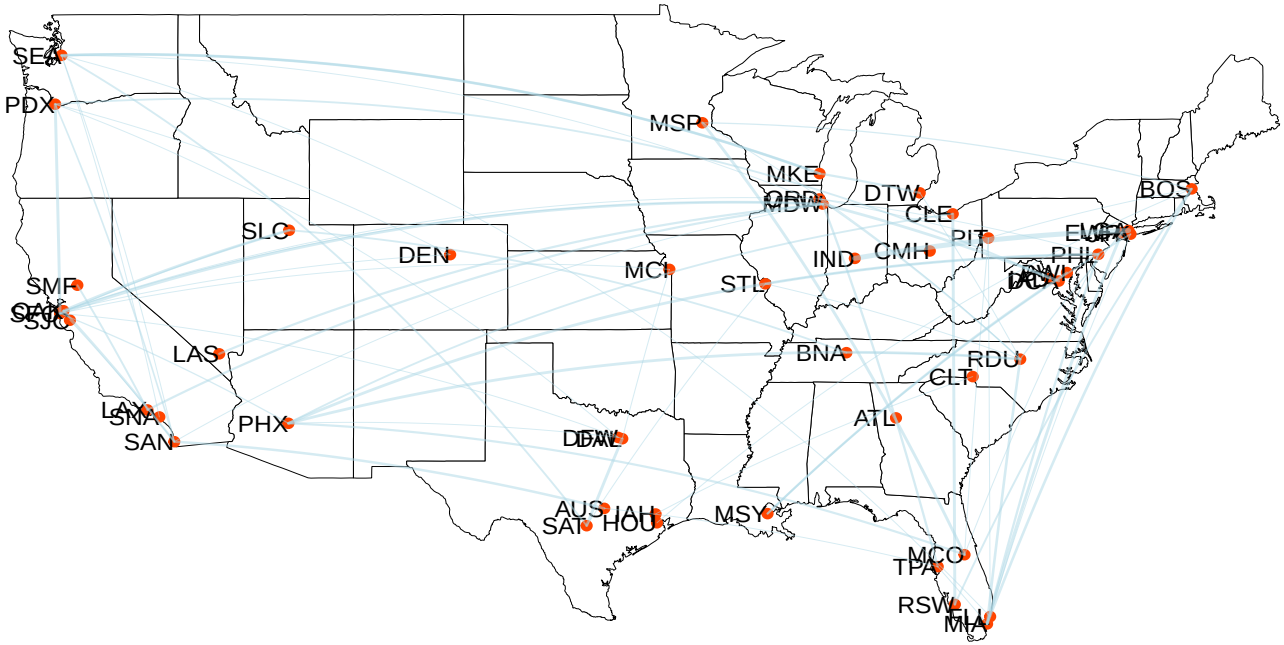


Figure: 11.26(outlier of weekdays).