# LocalClusteringTheory

2023-07-17

## Introduction to the Methodology

This is a brief summary of what we plan to do and how we plan to do it.

The goal is to develop pathway analysis methods that are more robust to pathway definitions that differ across databases.

---

**General set-up:**

1. We have a list of genes that we care about (e.g., genes over-expressed in tumors vs controls).
2. Those genes are components of genetic pathways (graphs).
3. The goal is to better interpret the results by finding genetic pathways that are over-represented in this list.
4. Analysis: Basic prob. (combinatorics). Calculate prob. of observing $g_p$ genes in pathway $\pi$ for each pathway. Depends on length of list and number of genes in pathway $\pi$.
5. Rather than relying on specific pathway definitions, which differ greatly across databases, we begin by forming a superset $\mathcal{S}$ that is the union of all pathways.
6. This superset ends up having one giant, connected, component, and a small number of genes that fall outside that component. For now, we reduce $\mathcal{S}$ to include just the connected component.
7. We aim to find local regions in $\mathcal{S}$ that are over-represented in our initial gene list.

---

;

## On methods for detecting local clustering...

First, let's orient ourselves by reminding ourselves of Moran's I, a global test for clustering...

;

**Global tests - Moran's $I$**

Suppose we have a network, $\mathcal{S}$, in which the value of node $i$ is denoted by $y_i$.

Moran's $I$ is defined as

$$I = \frac{N}{W} \frac{\sum_i \sum_j w_{ij}(y_i - \bar{y})(y_j - \bar{y}))}{\sum_i (y_i - \bar{y})^2}$$

where $N$ is the number of nodes (which are indexed by i(j)),

$(w_{ij})$ is a matrix of spatial weights with zeroes on the diagonal (e.g., distances between points).

$W$ is the sum of all $w_{ij}$.

Common choices for the weights are:

- weight=1 if two 'nodes' are neighbors, and 0 otherwise,
- weight=1 to $k$ nearest neighbors, some $k$, or 0 otherwise.
- use a distance decay function for assigning weights (e.g., 1/distance).

So it's a generalized correlation coefft.

---

**Theory for Moran's $I$**

Values of $I$ usually range from -1 to +1.

The expected value of Moran's $I$ under the null hypothesis of no spatial autocorrelation is

$$E(I) = \frac{-1}{N-1}$$

which approaches 0 as $N \to \infty$.

If we assume normality of node labels we have:

$$Var(I) = \frac{NS_4 - S_3 S_5}{(N-1)(N-2)(N-3)W^2} - (E(I))^2$$

where

$$S_1 = \frac{1}{2} \sum_i \sum_j (w_{ij} + w_{ji})^2$$

$$S_2 = \sum_i \left( \sum_j w_{ij} + \sum_j w_{ji} \right)^2$$

$$S_3 = \frac{N^{-1} \sum_i (y_i - \bar{y})^4}{(N^{-1} \sum_i (y_i - \bar{y})^2)^2}$$

$$S_4 = (N^2 - 3N + 3)S_1 - NS_2 + 3W^2$$

$$S_5 = (N^2 - N)S_1 - 2NS_2 + 6W^2$$

So testing is straightforward in that context, because this gives us a null distribution. Of course, in reality, we probably don't live in a world in which node labels are normally distributed.

---

**Notes on other global tests**

It's worth noting that there are other measures of global correlation, e.g., Geary's C (Geary 1954). It is claimed that Geary's C, despite being a global test, is more sensitive to local clustering. We will look into that.

Moran's $I$ is inversely related to Geary's C, but it is not identical. (wikipedia)

---

;

**Local Indicators of Spatial Association: LISA.**

Ref: Luc Anselin. Geogr Anal, 1995 vol. 27 (2) pp. 93-115.

http://doi.wiley.com/10.1111/j.1538-4632.1995.tb00338.x

Anselin introduces a new general class of **local indicators of spatial association (LISA)** and show how they allow for the decomposition of global indicators, such as Moran's $I$, into the contribution of each observation.

The LISA statistics serve two purposes:

- They may be interpreted as indicators of local pockets of nonstationarity, or hot spots, similar to the $G_i$ and $G$: statistics of Getis and Ord (1992). *This is the bit that interests us.*
- They may be used to assess the influence of individual locations on the magnitude of the global statistic and to identifiy outliers, as in Anselin's Moran scatterplot (1993a).

---

;

**LISA - Definition**

A Local Moran's $I$ can be defined as:

$$I_i = z_i \sum_j w_{ij} z_j.$$

Here, the $z_i$s are the $y_i$s standardized in terms of deviations from the mean. (i.e., $z_i = y_i - \bar{y}$)

Moments of $I_i$ can be calculated under some assumptions, that won't be applicable in our context. In any case, significance is typically assessed via randomization tests. The details of these tests matter and will require thought (more later).

---

;

**LISA - Features:**

1. The LISA for each observation gives an indication of the extent of significant spatial clustering of similar values around that observation;
2. It can be shown that the average of the $I_i$ equals the global $I$, up to a factor of proportionality.

---

;

**Significance testing**

General results on the distribution of a generic LISA are hard to obtain.

Instead, we propose, as is common, to use a conditional randomization or permutation approach.

The randomization is conditional in the sense that the value $y_i$ at a location $i$ is held fixed (that is, not used in the permutation) and the remaining values are randomly permuted over the locations in the data set.

A complicating factor in the assessment of significance of LISAs is that the statistics for individual locations will tend to be correlated.

Another complicating factor is the question of how to define regions of local clustering based upon the values of $I_i$ obtained at each node (or the measure of marginal significance of $I_i$ across all $i$.

There will be much to think about here!

---

;

**Local Geary**

FYI, a local version of Geary's global statistic can also be defined:

$$c_i = \sum_j w_{ij}(z_i - z_j)^2.$$

See Anselin's paper for more details. So, eventually, we will try that as well.

---

;

**Notes**

What you get out of this, once we've worked out how to do the randomization test, is a p-value for each node.

Small p-values indicate that the node is "at the center of" a local cluster.

Randomization/Permutation tests need to be done carefully. There are papers that do this and they point out that in their context they need to generate spatially correlated data to assess the null, because global spatial correlation often induces local clusters of over(under)representation even if there is no specific local clustering process going on. It's not clear that this will apply to us. This will all require more thought.

---

;

**Notes on moment results**

The moments under the null hypothesis are derived assuming that each value is equally likely at any location, which is inappropriate in the presence of global spatial association.

This is appropriate when the objective is to detect local spatial clusters in the absence of global spatial association.

But it is not correct when global spatial association is present

The conditional randomization (permutation) strategy does not suffer from these problems.

**Key Question:** Is global correlation present in the gene super-network?

---

;

**What should the null hypothesis be?**

The null hypothesis will assume that there is no local clusters.

But should it allow for, or exclude, global correlation?

Anselin looks at producing null data with varying degrees of autocorrelation using a spatial autoregression model. (So regressing each point on its neighbors in some way.) We have implemented something like this.

Anselin notes that *"while the mean and standard deviation are rough in accordance with those for a standard normal distribution, the kurtosis and to a lesser extent skewness are not."* (LISA has fatter tails because of the global autocorrelation.)

*"The implications of these results for inference in practice are that even when no global spatial autocorrelation is present, the significance levels indicated by a normal approximation will result in an over-rejection of the null hypothesis for a given Type I error."*

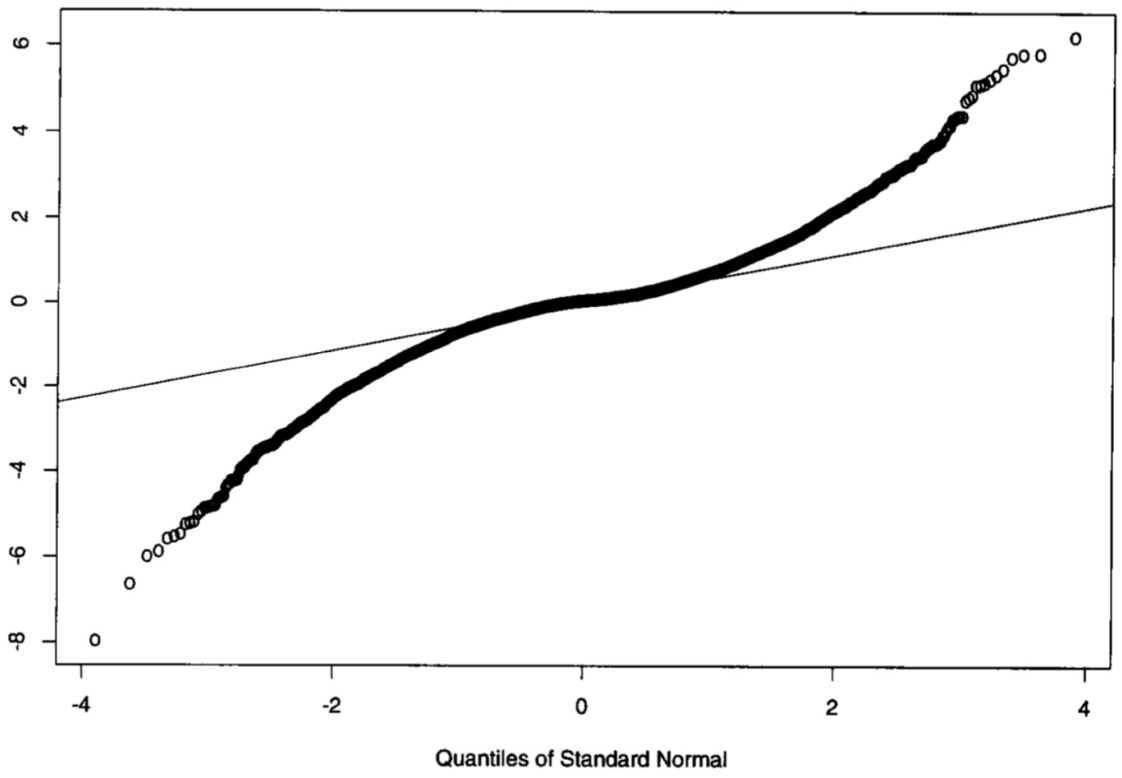


FIG. 6. Quantiles of $z$-values of Local Moran against the Normal Distribution ($n = 42$; 10,000 Replications)

*"The presence of global spatial autocorrelation has a strong influence on the moments of the distribution of the local Moran."*

*"Both mean and standard deviation increase with spatial autocorrelation, but the most significant effect seems to be on the skewness of the distribution."*

*"With increasing global spatial autocorrelation, both the spread and the number of outliers increases."*

*"[Thus,] in the presence of a high degree of spatial autocorrelation, several extreme values of the $I_i$ statistic are to be expected as a normal result of the heterogeneity induced by a spatial autoregressive process."*

;

**Conclusions**

What we get from this method is a permutation test based p-value for each node.

Small p-values indicates that the node is part of a local cluster of extreme values.

The normal approximation is not valid *"at least for the small sample sizes employed here"*. (Ours will be bigger.)

"*Furthermore, the uncritical use of the null distribution in the presence of global spatial autocorrelation will give incorrect significance levels.*"

---

;

**Questions for us...**

- Are these kinds of methods sensible to apply here? (many other methods exist)
- What is the right way to do the randomization test to assess the distribution of $I_i$ under the null?
- What kind of correlation do we see in the super-network in general?
- How do we use the node-specific p-values to produce the "enriched" network component(s)?
- Are there methods that directly label clusters rather than nodes?
- Once we have the enriched components (or lists of nodes), how do we map them back to the original pathways for interpretation? (And is that the right thing to do?)

---