Distance Correlation as a Measure for Dependence of Distance Matrices with Complex Spatial Patterns: An Alternative to Mantel Test

C. Deniz Yenigün [*] Maria Rizzo [†]

Abstract

The Mantel test is routinely used in many areas of ecology to test the independence of the elements in two distance matrices. Despite its popularity, the Mantel test has some well known disadvantages, and the fact that the test is only sensitive to linear or monotonic dependencies is often overlooked by the practitioners. In this paper we focus on resemblance matrices with nonlinear structures and point out that partial distance correlation is a powerful alternative to the Mantel statistic for measuring association of such matrices.

Key Words: Mantel test, partial Mantel test, distance correlation, partial distance correlation

1. Introduction

The Mantel test is a statistical test for the association between two matrices whose rows and columns correspond to the same objects. The test is an example of the well-known permutation tests, and the matrices of interest are typically resemblance matrices containing dissimilarities or distances computed from multivariate data tables. Routinely used in many areas such as ecology, genetics and biology, the Mantel test is a widely accepted tool for testing linear or monotonic independence between the matrices. The test was introduced in an epidemiological study by Mantel (1967) for identifying the dependence between a matrix of spatial distances and a matrix of temporal distances. Smouse, Long and Sokal (1986) proposed the partial Mantel test, which extends the idea to assessing the dependence between two matrices while controlling the effect of a third matrix.

Despite its tremendous popularity, the Mantel test is not free of criticism. The main criticism is the lack of a clear specification of the null and alternative hypotheses, resulting in practitioners overlooking the fact that the test is only sensitive to linear or monotonic dependencies in the similarity matrices. As will be described below, in case of nonlinear spatial patterns, the test may be misleading. For a detailed discussion on this drawback and for some attempts for solution, see, for example, Goslee and Urban (2007), Guillot and Rousset (2013) and Legendre, Fortin and Borcard (2015). Another criticism on the Mantel test states that it should be used only in situations where it is appropriate to express the hypothesis in terms of the distance matrices. As noted by Dutilleul et al. (2000), even if the covariance of columns in the raw multivariate data table is zero, the Mantel statistic computed from the distance matrices may be non-zero. The third notable issue with the Mantel test is on the construction of the permutation test. Typically objects in one of the resemblance matrices are permuted in order to calculate the p-value of the Mantel test. However, Legendre (2000) points out that different permutation procedures must be preferred under different conditions, which introduces another complexity from the practitioner's point of view.

In this paper we focus on resemblance matrices with complex nonlinear structures and point out that distance correlation is a powerful tool for testing independence of such matrices. Proposed by Székely, Rizzo and Bakirov, N.K., (2007), distance correlation characterises independence of random vectors in arbitrary dimensions. In Section 2 we give an overview of simple and partial Mantel test, distance correlation, and partial distance correlation. We focus on a data set with a nonlinear spatial component in Section 3, and point out that distance correlation is superior to Mantel statistic in terms of identifying these associations. Section 4 concludes.

[*] Istanbul Bilgi University, Department of Industrial Engineering, 34060, Istanbul, Turkey

[†] Bowling Green State University, Department of Mathematics and Statistics, 43403, Bowling Green, Ohio, USA

2. Background

2.1 Simple and Partial Mantel Test

Consider two multivariate data sets contained in an $n \times p$ matrix X and an $n \times q$ matrix Y. Here, for example, X may be p environmental characteristics for n locations, and Y may be q species compositions for these locations such that the i-th row of Y describes the same location as the i-th row of X. Now consider pairwise dissimilarity (or distance) matrices D_X and D_Y computed from X and Y using an appropriate distance metric such as *Jaccard*, *Bray-Curtis*, or *Euclidean* distance. Clearly, D_X and D_Y are $n \times n$ symmetric matrices with zero diagonals, whose rows and columns correspond to same set of objects. If X and Y contain the data sets in the above example, then D_X displays the dissimilarities between the environmental characteristics of n locations, and D_Y displays the dissimilarities between the species composition of the same locations. Then the *Mantel statistic*, denoted by r_M, is the Pearson correlation between the upper (or lower) triangular portions of D_X and D_Y.

The null hypothesis of the Mantel test states that distances among D_X are linearly independent of the distances among the same objects in D_Y. In the context of the above example, this hypothesis states that the environmental dissimilarities are linearly independent of the species composition dissimilarities. Since same individual observations are used repeatedly in generating the distance matrices, the matrix entries are correlated with each other and therefore the usual large sample results for Pearson correlation cannot be used for obtaining the null distribution of the Mantel statistic. Instead, the significance of the test statistic is assessed by a permutation test. The usual approach in the Mantel test is to randomly permute the rows and columns of one of the matrices, say D_X, and compute the Mantel statistic between D_Y and the permuted matrix. Note that the permutation of rows and columns must agree and the resulting matrix should also be symmetric. This procedure is repeated a large number of times to obtain the null permutation distribution of the Mantel statistic. Then the p-value of the Mantel test is the proportion of the Mantel statistics in the null permutation distribution which are larger than the original Mantel statistic. This procedure is known as the *Simple Mantel Test*.

The simple Mantel test only considers the relationship between two dissimilarity matrices. Another possibility is to study the relationship between two matrices, but taking into account the effect of a third one. For instance, in the above example the researcher may want to find out if the species composition dissimilarities are indeed related with the environmental dissimilarities, or if the observed relationship appears only because both variables are spatially structured by intrinsic effects. In other words, the researcher may want to measure the dependence between two matrices, after the effect of a third matrix containing the geographical distances have been removed. Proposed by Smouse et al. (1986), the *Partial Mantel Test* addresses this issue using the partial correlation.

Consider three distance matrices D_X, D_Y, and D_Z whose rows and columns correspond to the same set of objects. The partial Mantel statistic $r_M(D_X, D_Y; D_Z)$, measuring the relationship between the matrices D_X and D_Y while controlling for the effect of D_Z, is computed the same way as the first order partial correlation coefficient as follows.

$$r_M(D_X, D_Y; D_Z) = \frac{r_M(D_X, D_Y) - r_M(D_X, D_Z) r_M(D_Y, D_Z)}{\sqrt{1 - r_M(D_X, D_Y)^2}\sqrt{1 - r_M(D_Y, D_Y)^2}}, \tag{1}$$

where $r_M(A, B)$ is the simple Mantel test statistic between matrices A and B. Similar to the simple Mantel test, partial Mantel test is also a permutation test where typically the objects in one of the original dissimilarity matrices are permuted and others are left unpermuted. However, some alternative permutation methods have been proposed and shown to have higher power under certain conditions. Note that all these methods produce the same value for the Mantel statistic, but may produce different p-values for the test. See, for example, Legendre (2000) for a detailed discussion on the effect of permutation methods in partial Mantel test.

Constructing a Mantel test controlling for more than one matrices is straightforward. Note that the first order partial correlation in (1) is a special case of the broader idea of partial correlation. In general, partial correlation between data vectors A and B controlling for the effect of vectors $C_1, ..., C_n$ requires regressing A and B separately on $C_1, ..., C_n$. The Pearson correlation between the residuals of these two regressions is defined to be the partial correlation, and when $n = 1$ this amounts to (1). Therefore, when constructing partial Mantel test controlling for more than one matrices, one may employ the general definition of partial correlation on the upper (or lower) triangular portions of the matrices. Again, significance of the test is obtained by a permutation test where typically one of the original matrices is permuted.

Another useful extension of the Mantel statistic is the Mantel correlogram, a graphic displaying autocorrelations within subsets of the distance matrix of interest corresponding to *similar* items in terms of another distance matrix. To illustrate, in our example above it may be of interest to study the relationship between species composition distances (D_X) and geographical distances (D_Z) across space. In order to do this, the matrix D_Z can be divided into several sub-matrices, each one describing pairs within an interval of geographical distances. Let us define these sub-matrices by Z_k, matrices with binary entries returning a value of 1 if pairs are within a geographic distance referred to as distance class k, and returning a value of 0 otherwise. In order to analyse correlations across space one may create multiple non-overlapping and contiguous distance classes, and compute the Mantel statistic between D_X and Z_1, Z_2,...,Z_k. The Mantel correlogram is constructed by plotting Mantel statistics against the mid-point of the distance classes k. This tool is especially useful for displaying nonlinear relationships between D_X and D_Z. The definition of the distance classes in terms of number and boundaries is somewhat arbitrary and depends on the spatial distribution of data.

2.2 Distance Correlation and Partial Distance Correlation

Distance correlation is a relatively new and powerful dependence measure introduced by Székely, Rizzo and Bakirov, N.K., (2007). For all distributions with finite first moments, distance correlation generalizes the idea of correlation in two fundamental ways. Firstly, distance correlation is defined for variables in arbitrary dimensions, it is not limited to the bivariate case. Secondly, distance correlation vanishes if and only if the variables are independent.

Consider random vectors X in $\mathbf{R}^p$ and Y in $\mathbf{R}^q$. The characteristic functions of X and Y are denoted by f_X and f_Y, respectively, and the joint characteristic function of X and Y is $f_{X,Y}$. The *distance covariance* between X and Y is

$$\begin{aligned} V^2(X,Y) &= \|f_{X,Y}(t,s) - f_X(t)f_Y(s)\|^2 \\ &= \frac{1}{c_p c_q}\int_{\mathbf{R}^{p+q}} \frac{|f_{X,Y}(t,s) - f_X(t)f_Y(s)|^2}{|t|_p^{1+p}|s|_q^{1+q}} dtds, \end{aligned}$$

where $c_d = \pi^{(1+d)/2}/\Gamma\{(1+d)/2\}$, Γ is the complete gamma function, and $|a|_d$ is the Euclidean norm of a in $\mathbf{R}^d$. Similarly, the *distance variance* of X is

$$V^2(X) = \|f_{X,X}(t,s) - f_X(t)f_X(s)\|^2, \tag{2}$$

and the *distance correlation* between X and Y is the positive square root of

$$R^2(X,Y) = \begin{cases} \frac{V^2(X,Y)}{\sqrt{V^2(X)V^2(Y)}}, & V^2(X)V^2(Y) > 0 \\ 0, & V^2(X)V^2(Y) = 0 \end{cases}. \tag{3}$$

Székely and Rizzo (2014) extend the idea of distance correlation to define *partial distance correlation*. The partial distance correlation between X and Y, given Z is

$$R^*(X,Y;Z) = \begin{cases} \frac{R^2(X,Y) - R^2(X,Z)R^2(Y,Z)}{\sqrt{1-R^4(X,Z)}\sqrt{1-R^4(Y,Z)}}, & \\ & R(X,Y) \neq 1 \text{ and } R(X,Y) \neq 1; \\ & \\ 0, & R(X,Y) = 1 \text{ or } R(X,Y) = 1, \end{cases} \tag{4}$$

where $R(X,Y)$ denotes the distance correlation. The empirical distance correlation and empirical partial distance correlation are defined in the referred papers, and independence tests have been proposed. The R package *energy* by Rizzo and Székely (2011) can be used for implementation of these methods.

	Forest Cover (Y)	Location (Z)
Composition (X)	0.34	0.23
Forest Cover (Y)	-	0.05

Table 1: Pairwise Mantel statistics for plant community composition data.

3. Understanding Nonlinear Spatial Structures

While studying spatial autocorrelation, ecologists often measure the significance of correlation between compositional or environmental dissimilarity and geographic distance. In such cases partial Mantel statistic is expected to correct any spatial autocorrelation as it considers the partial correlation between composition and environmental dissimilarities, given geographical distance. When the underlying spatial structure is nonlinear, the Mantel statistic may be misleading as it implicitly assumes linearity. In this section we will first reveal this drawback of Mantel statistic using a real data set. Then we will illustrate that distance correlation may capture and correct for such nonlinear spatial structures, providing a strong alternative to Mantel statistic.

In order to illustrate how Mantel test may fail to respond to complex spatial patterns, we follow the steps of an illustration given in Goslee and Urban (2007), which analyses the *Plant Community Composition Data* compiled by Tracy and Sanderson (2000). The data set contains 50 vegetation samples taken on 12 farms in northern New York. The observed variables are *Composition* (X), mean values of canopy cover estimates from 10 quadrats located within a pasture; *Forest Cover* (Y), percentage forest cover in the area; and *Location* (Z), a function of longitude and latitude since sites fall along southwest to northeast line. The main question here is whether or not there is an association between plant community composition and the percent forest cover within a circle of one kilometer surrounding the farm. There is a change in species identity from north to south, so the effect of location must be accounted for when considering any relationship between community composition and other factors.

3.1 Analysis Using Simple and Partial Mantel Statistic

The pairwise Mantel statistics between the variables are given in Table 1. All pairwise Mantel statistics indicate statistically significant association with p-values less than 10^{-8}. Since the question of interest is "Is there a relationship between community composition (X) and surrounding forest cover (Y), once spatial effects (Z) have been accounted for?" the usual approach would be to use partial Mantel test on composition and forest cover, given the effects of location.

The partial Mantel statistic between community composition and surrounding forest cover, given spatial effects turns out to be 0.3417. In other words, once the effect of location is removed, there is a significant relationship between community composition and forest cover. We know that Mantel test only removes the linear component of the variation, not more complex spatial patterns. In this example, we see that the simple Mantel statistic and the partial mantel statistic are very close, thus, the inclusion of the partial had no effect. In fact, the percentage reduction in the Mantel statistic after the removal of the location effect is only 0.43%.

3.2 Identifying the Nonlinear Spatial Patterns

As observed above, the conventional use of partial Mantel statistic is unable to remove the potential spatial component of the relationship between plant community composition and forest cover. This may be due to the fact that the spatial component is nonlinear. In this section we further analyse the data in order to understand the nature of the spatial effect.

In spatial statistics a *correlogram* displays the value of Mantel statistic of observations with a specified distance (lag) apart from each other. In Figure 1 we display the spatial correlograms in order to identify potential nonlinear relationships. The correlogram on the left panel indicates that the plant community composition shows a roughly linear relationship at closer distances, with space being unimportant at farther distances. However, when we focus on forest cover displayed on the right panel, we see a strong nonlinear pattern. Recall that Mantel test was insignificant because there is little or no linear component to the relationship. Using the Mantel statistic alone would have given a false impression on the structure of the data.

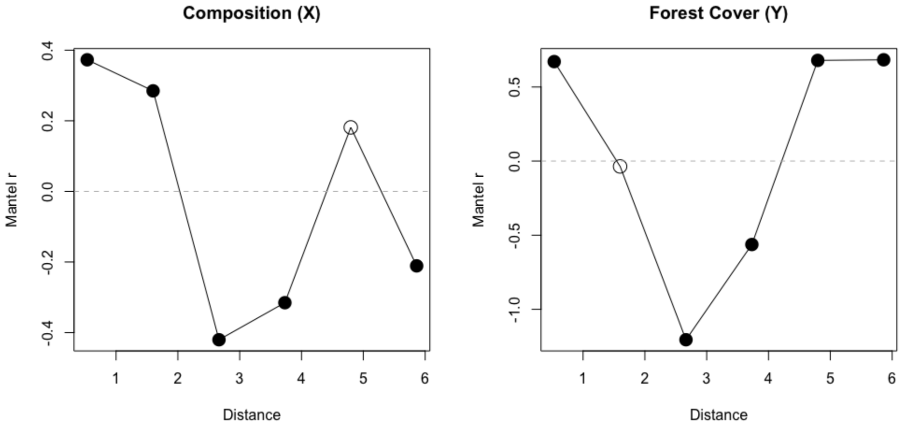

Figure 1: Correlograms for plant community composition data.

Next, we vectorize all observed distances and create a matrix plot to see the nonlinear spatial association from a different perspective. The plots in Figure 2 verify the linear spatial characteristic of community composition and the nonlinear spatial characteristic of forest percentage indicated by the correlogram analysis. Both figures provide some evidence that the nonlinear spatial characteristic of forest percentage may be the reason that the partial Mantel statistic cannot account for the spatial effect while quantifying the association between community composition and forest cover.

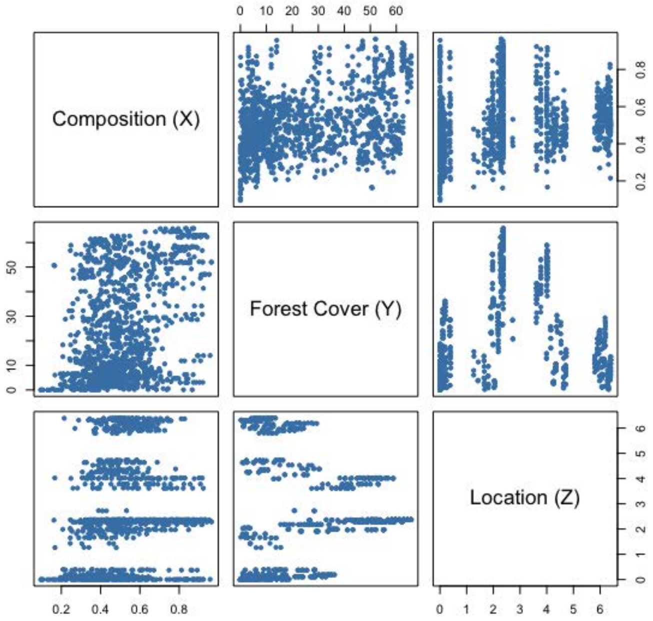

Figure 2: Scatter Plots.

	Forest Cover (Y)	Location (Z)
Composition (X)	0.18	0.45
Forest Cover (Y)	-	0.27

Table 2: Pairwise Distance Correlations for plant community composition data.

3.3 Analysis Using Partial Distance Correlation

Having observed the nonlinear spatial characteristic of data, partial distance correlation is considered to be a strong alternative to the commonly used Mantel statistic. In this section we repeat every step of the Mantel analysis described above using bias-corrected distance correlation and partial distance correlation. The distance correlation matrix is given in Table 2. The partial distance correlation between composition and forest percentage, given the location is observed to be 0.068. We see that, removing the effect of location greatly reduces the distance correlation between composition and forest percentage, namely, from 0.18 to 0.068. This is what we wanted to achieve in this example as we observed that a strong nonlinear relationship exists between forest cover and location. The percentage reduction in distance correlation after the removal of the location effect is observed to be 62.3%.

4. Conclusion

In this paper we focus on resemblance matrices with nonlinear structures and point out that distance correlation is a powerful tool for measuring association of such matrices. The illustrative example we focus on reveals that in case of nonlinear spatial associations, partial distance correlation is superior to the commonly used partial Mantel statistic, in terms of measuring the association between two distance matrices while accounting for the effect of geographical distances.

REFERENCES

Dutilleul, P., Stockwell, J.D., Frigon, D., Legendre, P., (2000), “The Mantel test versus Pearson’s correlation analysis: assessment of the differences for biological and environmental studies,” *Journal of Agricultural, Biological, and Environmental Statistics*, 5, 131-150.

Goslee, C.G., Urban, D.L., (2007), “The ecodist Package for Dissimilarity-based Analysis of Ecological Data,” *Journal of Statistical Software*, Volume 22, Issue 4, 1-19.

Guillot, G., Rousset , F., (2013), “Dismantling the Mantel tests,” *Methods in Ecology and Evolution*, 4, 336344.

Legendre, P., (2000), “Comparison of permutation methods for the partial correlation and partial mantel tests,” *Journal of Statistical Computation and Simulation*, 67, 37-73.

Legendre, P., Fortin, M.J., Borcard, D., (2015), “Should the Mantel test be used in spatial analysis?,” *Methods in Ecology and Evolution*, 6, 12391247.

Mantel, N., (1967), “The detection of disease clustering and a generalised regression approach,” *Cancer Research*, 27, 209-220.

Rizzo, M. L. and Székely, G. J. (2018). energy: E-statistics (energy statistics). R package version 1.7-5.

Smouse, P.E., Long, J.C., Sokal, R.R., (1986), “Multiple regression and correlation extensions of the Mantel test of matrix correspondence,” *Systematic Zoology*, 35, 627-632.

Székely, G.J., Rizzo, M.L.,(2014), “Partial distance correlation with methods for dissimilarities,” *The Annals of Statistics*, 42, 2382-2412 .

Székely, G.J., Rizzo, M.L., Bakirov, N.K., (2007), “Measuring and testing dependence by correlation of distances,” *The Annals of Statistics*, 35, 2769-2794.

Tracy, B.F., Sanderson, M.A., (2000), “Patterns of plant species richness in pasture lands of the Northeast United States,” *Plant Ecology*, 149, 200-212.