Moran's Autocorrelation Coefficient in Comparative Methods

Emmanuel Paradis

July 15, 2014

This document clarifies the use of Moran's autocorrelation coefficient to quantify whether the distribution of a trait among a set of species is affected or not by their phylogenetic relationships.

1 Theoretical Background

Moran's autocorrelation coefficient (often denoted as I) is an extension of Pearson product-moment correlation coefficient to a univariate series [2, 5]. Recall that Pearson's correlation (denoted as ρ) between two variables x and y both of length n is:

$$\rho = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\left[\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2\right]^{1/2}},$$

where $\bar{x}$ and $\bar{y}$ are the sample means of both variables. ρ measures whether, on average, x_i and y_i are associated. For a single variable, say x, I will measure whether x_i and x_j, with $i \neq j$, are associated. Note that with ρ, x_i and x_j are *not* associated since the pairs (x_i, y_i) are assumed to be independent of each other.

In the study of spatial patterns and processes, we may logically expect that close observations are more likely to be similar than those far apart. It is usual to associate a *weight* to each pair (x_i, x_j) which quantifies this [3]. In its simplest form, these weights will take values 1 for close neighbours, and 0 otherwise. We also set $w_{ii} = 0$. These weights are sometimes referred to as a *neighbouring function*.

I's formula is:

$$I = \frac{n}{S_0} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \tag{1}$$

where w_{ij} is the weight between observation i and j, and S_0 is the sum of all w_{ij}'s:

$$S_0 = \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}.$$

Quite not so intuitively, the expected value of I under the null hypothesis of no autocorrelation is not equal to zero but given by $I_0 = -1/(n-1)$. The expected variance of I_0 is also known, and so we can make a test of the null hypothesis. If the observed value of I (denoted $\hat{I}$) is significantly greater than I_0, then values of x are positively autocorrelated, whereas if $\hat{I} < I_0$, this will indicate negative autocorrelation. This allows us to design one- or two-tailed tests in the standard way.

Gittleman & Kot [4] proposed to use Moran's I to test for "phylogenetic effects". They considered two ways to calculate the weights w:

- With phylogenetic distances among species, e.g., $w_{ij} = 1/d_{ij}$, where d_{ij} are distances measured on a tree.
- With taxonomic levels where $w_{ij} = 1$ if species i and j belong to the same group, 0 otherwise.

Note that in the first situation, there are quite a lot of possibilities to set the weights. For instance, Gittleman & Kot also proposed:

$$\begin{array}{ll} w_{ij} = 1/d_{ij}^{\alpha} & \text{if } d_{ij} \leq c \\ w_{ij} = 0 & \text{if } d_{ij} > c, \end{array}$$

where c is a cut-off phylogenetic distance above which the species are considered to have evolved completely independently, and α is a coefficient (see [4] for details). By analogy to the use of a spatial correlogram where coefficients are calculated assuming different sizes of the "neighbourhood" and then plotted to visualize the spatial extent of autocorrelation, they proposed to calculate I at different taxonomic levels.

2 Implementation in ape

From version 1.2-6, `ape` has functions `Moran.I` and `correlogram.formula` implementing the approach developed by Gittleman & Kot. There was an error in the help pages of `?Moran.I` (corrected in ver. 2.1) where the weights were referred to as "distance weights". This has been wrongly interpreted in my book [6, pp.˜139–142]. The analyses below aim to correct this.

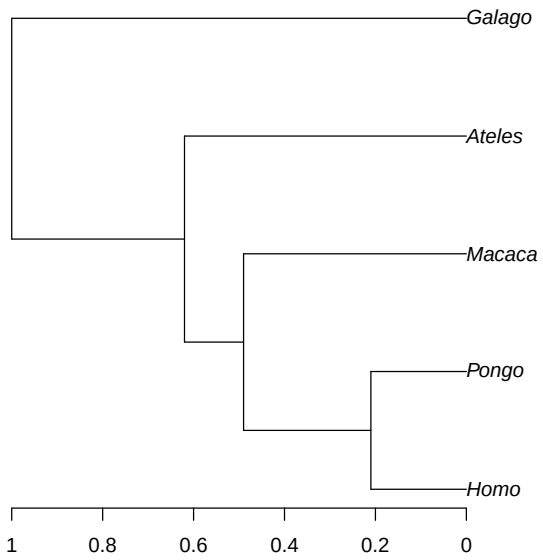
2.1 Phylogenetic Distances

The data, taken from [1], are the log-transformed body mass and longevity of five species of primates:

```
> body <- c(4.09434, 3.61092, 2.37024, 2.02815, -1.46968)
> longevity <- c(4.74493, 3.3322, 3.3673, 2.89037, 2.30259)
> names(body) <- names(longevity) <- c("Homo", "Pongo", "Macaca", "Ateles", "Galago")
```

The tree has branch lengths scaled so that the root age is one. We read the tree with ape, and plot it:

```
> library(ape)
> trnwk <- "((((Homo:0.21,Pongo:0.21):0.28,Macaca:0.49):0.13,Ateles:0.62)"
> trnwk[2] <- ":0.38,Galago:1.00);"
> tr <- read.tree(text = trnwk)
> plot(tr)
> axisPhylo()
```



We choose the weights as $w_{ij} = 1/d_{ij}$, where the d's is the distances measured on the tree:

```
> w <- 1/cophenetic(tr)
> w
```

```
           Homo     Pongo    Macaca    Ateles Galago
Homo         Inf 2.3809524 1.0204082 0.8064516    0.5
Pongo  2.3809524       Inf 1.0204082 0.8064516    0.5
Macaca 1.0204082 1.0204082       Inf 0.8064516    0.5
Ateles 0.8064516 0.8064516 0.8064516       Inf    0.5
Galago 0.5000000 0.5000000 0.5000000 0.5000000    Inf
```

Of course, we must set the diagonal to zero:

```
> diag(w) <- 0
```

We can now perform the analysis with Moran's I:

```
> Moran.I(body, w)

$observed
[1] -0.07312179

$expected
[1] -0.25

$sd
[1] 0.08910814

$p.value
[1] 0.04714628
```

Not surprisingly, the results are opposite to those in [6] since, there, the distances (given by `cophenetic(tr)`) were used as weights. (Note that the argument `dist` has been since renamed `weight`.[1]) We can now conclude for a slighly significant positive phylogenetic correlation among body mass values for these five species.

The new version of `Moran.I` gains the option `alternative` which specifies the alternative hypothesis (`"two-sided"` by default, i.e., H_1: $I \neq I_0$). As expected from the above result, we divide the P-value be two if we define H_1 as $I > I_0$:

```
> Moran.I(body, w, alt = "greater")

$observed
[1] -0.07312179

$expected
[1] -0.25

$sd
[1] 0.08910814

$p.value
[1] 0.02357314
```

The same analysis with `longevity` gives:

```
> Moran.I(longevity, w)

$observed
[1] -0.1837739

$expected
[1] -0.25
```

[1] The older code was actually correct; nevertheless, it has been rewritten, and is now much faster. The documentation has been clarified. The function `correlogram.phylo`, which computed Moran's I for a tree given as argument using the distances among taxa, has been removed.

```
$sd
[1] 0.09114549

$p.value
[1] 0.4674727
```

As for `body`, the results are nearly mirrored compared to [6] where a non-significant negative phylogenetic correlation was found: it is now positive but still largely not significant.

2.2 Taxonomic Levels

The function `correlogram.formula` provides an interface to calculate Moran's I for one or several variables giving a series of taxonomic levels. An example of its use was provided in [6, pp.˜141–142]. The code of this function has been simplified, and the graphical presentation of the results have been improved.

`correlogram.formula`'s main argument is a formula which is "sliced", and `Moran.I` is called for each of these elements. Two things have been changed for the end-user at this level:

1. In the old version, the rhs of the formula was given in the order of the taxonomic hierarchy: e.g., `Order/SuperFamily/Family/Genus`. Not respecting this order resulted in an error. In the new version, any order is accepted, but the order given it is then respected when plotted the correlogram.

2. Variable transformations (e.g., log) were allowed on the lhs of the formula. Because of the simplification of the code, this is no more possible. So it is the responsibility of the user to apply any tranformation before the analysis.

Following Gittleman & Kot [4], the autocorrelation at a higher level (e.g., family) is calculated among species belonging to the same category and to different categories at the level below (genus). To formalize this, let us write the different levels as $X^1/X^2/X^3/\ldots/X^n$ with X^n being the lowest one (`Genus` in the above formula):

$$\left.\begin{array}{ll} w_{ij} = 1 & \text{if } X_i^k = X_j^k \text{ and } X_i^{k+1} \neq X_j^{k+1} \\ w_{ij} = 0 & \text{otherwise} \end{array}\right\} k < n$$

$$\left.\begin{array}{ll} w_{ij} = 1 & \text{if } X_i^k = X_j^k \\ w_{ij} = 0 & \text{otherwise} \end{array}\right\} k = n$$

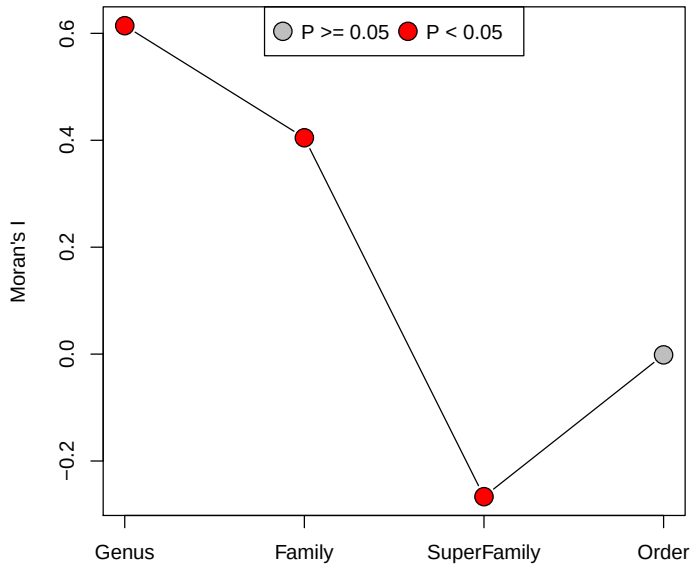
This is thus different from the idea of a "neighbourhood" of different sizes, but rather similar to the idea of partial correlation where the influence of the lowest level is removed when considering the highest ones [4].

To repeat the analyses on the `carnivora` data set, we first $\log_{10}$-transform the variables mean body mass (`SW`) and the mean female body mass (`FW`):

```
> data(carnivora)
> carnivora$log10SW <- log10(carnivora$SW)
> carnivora$log10FW <- log10(carnivora$FW)
```

We first consider a single variable analysis (as in [6]):

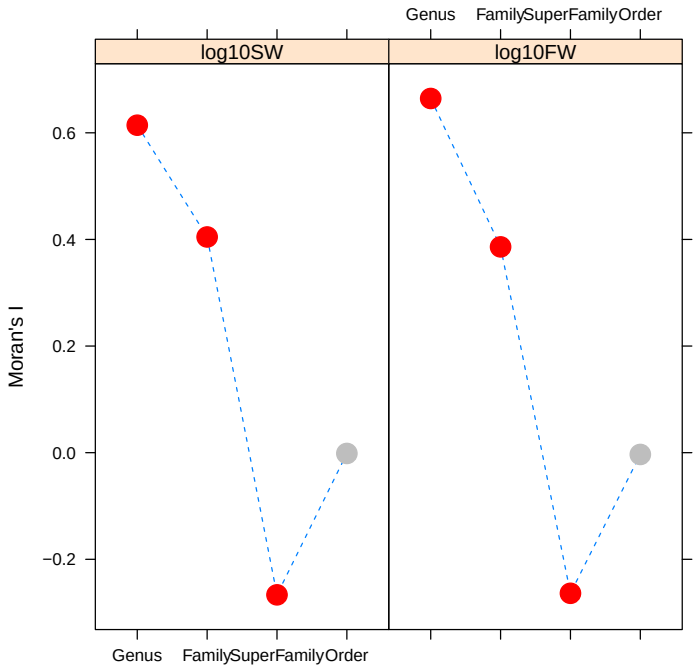
```
> fm1.carn <- log10SW ~ Order/SuperFamily/Family/Genus
> co1 <- correlogram.formula(fm1.carn, data = carnivora)
> plot(co1)
```



A legend now appears by default, but can be removed with `legend = FALSE`. Most of the appearance of the graph can be customized via the option of the plot method (see `?plot.correlogram` for details). This is the same analysis than the one displayed on Fig.˜6.3 of [6].

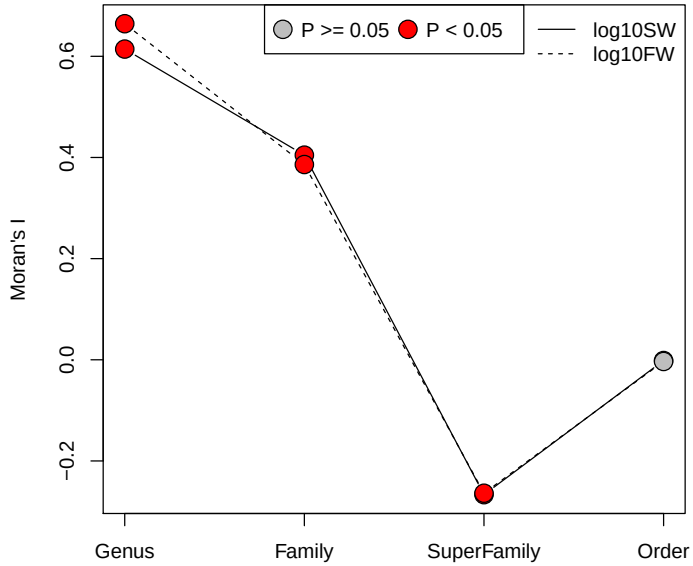
When a single variable is given in the lhs in `correlogram.formula`, an object of class `"correlogram"` is returned as above. If several variables are analysed simultaneously, the object returned is of class `"correlogramList"`, and the correlograms can be plotted together with the appropriate plot method:

```
> fm2.carn <- log10SW + log10FW ~ Order/SuperFamily/Family/Genus
> co2 <- correlogram.formula(fm2.carn, data = carnivora)
> print(plot(co2))
```



By default, lattice is used to plot the correlograms on separate panels; using `lattice = FALSE` (actually the second argument, see `?plot.correlogramList`) makes a standard graph superimposing the different correlograms:

```
> plot(co2, FALSE)
```



The options are roughly the same than above, but do not have always the same effect since lattice and base graphics do not have the same graphical parameters. For instance, `legend = FALSE` has no effect if `lattice = TRUE`.

3 Implementation in ade4

The analysis done with ade4 in [6] suffers from the same error than the one done with `Moran.I` since it was also done with a distance matrix. So I correct this below:

```
> library(ade4)
> gearymoran(w, data.frame(body, longevity))
```

```
class: krandtest
Monte-Carlo tests
Call: as.krandtest(sim = matrix(res$result, ncol = nvar, byr = TRUE),
    obs = res$obs, alter = alter, names = test.names)

Test number:   2
Permutation number:   999
Alternative hypothesis: greater

       Test         Obs   Std.Obs Pvalue
1      body -0.06256789 2.1523342  0.001
2 longevity -0.22990437 0.3461414  0.414

other elements: NULL
```

The results are wholly consistent with those from ape, but the estimated coefficients are substantially different. This is because the computational methods are not the same in both packages. In ade4, the weight matrix is first transformed as a relative frequency matrix with $\tilde{w}_{ij} = w_{ij}/S_0$. The weights are further transformed with:

$$p_{ij} = \tilde{w}_{ij} - \sum_{i=1}^{n} \tilde{w}_{ij} \sum_{j=1}^{n} \tilde{w}_{ij},$$

with p_{ij} being the elements of the matrix denoted as P. Moran's I is finally computed with $x^{\mathrm{T}}Px$. In ape, the weights are first row-normalized:

$$w_{ij} \Big/ \sum_{i=1}^{n} w_{ij},$$

then eq.˜1 is applied.

Another difference between both packages, though less important, is that in ade4 the weight matrix is forced to be symmetric with $(W + W^{\mathrm{T}})/2$. In ape, this matrix is assumed to be symmetric, which is likely to be the case like in the examples above.

4 Other Implementations

Package sp has several functions, including `moran.test`, that are more specifically targeted to the analysis of spatial data. Package spatial has the function `correlogram` that computes and plots spatial correlograms.

Acknowledgements

I am thankful to Thibaut Jombart for clarifications on Moran's I.

References

[1] J.˜M. Cheverud, M.˜M. Dow, and W.˜Leutenegger. The quantitative assessment of phylogenetic constraints in comparative analyses: sexual dimorphism in body weight among primates. *Evolution*, 39:1335–1351, 1985.

[2] A.˜D. Cliff and J.˜K. Ord. *Spatial Autocorrelation*. Pion, London, 1973.

[3] A.˜D. Cliff and J.˜K. Ord. Spatial and temporal analysis: autocorrelation in space and time. In E.˜N. Wrigley and R.˜J. Bennett, editors, *Quantitative Geography: A British View*, pages 104–110. Routledge & Kegan Paul, London, 1981.

[4] J.˜L. Gittleman and M.˜Kot. Adaptation: statistics and a null model for estimating phylogenetic effects. *Systematic Zoology*, 39:227–241, 1990.

[5] P.˜A.˜P. Moran. Notes on continuous stochastic phenomena. *Biometrika*, 37:17–23, 1950.

[6] E.˜Paradis. *Analysis of Phylogenetics and Evolution with R*. Springer, New York, 2006.