Eigenvector selection guidelines in R

David Bauman

6 septembre 2017

Objective of this document

This document aims at giving general guidelines for selecting an optimal subset of spatial eigenvectors (EVs) depending on the purpose of the researcher and on the uni- or multivariate nature of the response.

Useful packages

```
library(vegan)
library(adespatial)
library(spdep)
```

Data input

The oribatid mite dataset will be used to illustrate the eigenvector selection procedures (see Borcard et al. 1992, 1994 for details on the data).

```
data(mite)
data(mite.xy)

Y <- mite
C <- mite.xy
```

If Y is multivariate, we transform the data with the Hellinger transformation (more details in Legendre and Gallagher 2001):

```
Y <- decostand(Y, method = "hellinger")
```

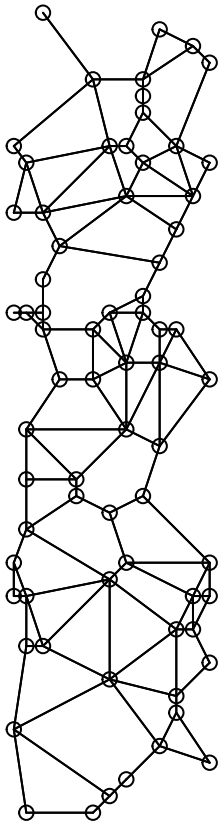
I. Moran's eigenvector maps (MEM): constructing the spatial variables

The MEM variables (or spatial eigenvectors, EVs) can be built from a huge variety of spatial weighting matrices (W). The W matrix is constructed by the Hadamard product of a connectivity matrix (B) defining which sites are connected and which are not, and a weighting matrix (A) either binary (no weighting) or continuous (more realistic W), often making connectivity decrease with the distance. The decrease can be linear, or follow a concave-down or -up curve (see Dray et al. 2006 for details). This R code does not provide the way to select an optimal W matrix, as this procedure still needs to be thoroughly adressed through an unbiased procedure (Bauman et al., unpublished).

The function createlistw of the adespatial package allows visualising different connectivity criteria. This can help deciding which connectivity scheme is more realistic, considering the response under study, the field reality (resistance to long distance movements etc).

Here, we use a Gabriel graph with no weighting to build the spatial EVs, as an example to illustrate the unbiased eigenvector selection procedures (which are the focus of this document). However, any W matrix can be used for the continuation of the procedure.

```
C <- as.matrix(C)
createlistw()
```



Gabriel graph of the mite dataset:
We begin by defining some parameters.

Define whether we want the spatial EVs modelling positively autocorrelated patterns ("positive"), negatively autocorrelated patterns ("negative"), or all n-1 EVs ("all"):

```
MEM_model <- "positive"
```

Standardisation scheme of the listw object (see help of nb2listw):

```
style <- "B"

nb <- graph2nb(gabrielneigh(as.matrix(C), nnmult = 5), sym = TRUE)
listw <- nb2listw(nb, style = style)
MEM <- scores.listw(listw, MEM.autocor = MEM_model)
```

II. Eigenvector selection

II.1. Spatial filters - Controlling the spatial autocorrelation of an OLS or GLM model residuals

If the response is univariate, and if the purpose of the spatial EVs is to control the spatial autocorrelation of an OLS or GLM model, then an suited EV selection is that of Griffith and Peres-Neto (2006). This procedure selects the smallest EV subset best minimising the Moran's I of the model residuals.

As this selection procedure is restricted to univariate data, we only consider the second species of the community dataframe.

```
Y <- mite
Y <- Y[, 2]
```

II.1.1. Selection of the EVs based on the residuals of an explanatory model of Y on a set of explanatory (often environmental) variables (X):

```
data(mite.env)
X <- mite.env[, 1:2]

select <- ME(Y ~., data = as.data.frame(X), listw = listw, family = gaussian, nsim = 99,
             alpha = 0.05)
sort <- sort(as.numeric(select$selection[, 1][2:length(which(select$selection[, 3] <= 0.05))]))
MEM.select <- MEM[, sort]
```

MEM.select can be used to control the spatial autocorrelation of our model by adding the EVs to the explanatory variables.

II.1.2. Selection of the EVs based on Y only (MIR approach in Bauman et al. 2017, Fig. 1, step 3.3.).

This adapted version of the function ME focuses on controling the spatial autocorrelation of Y, instead of the spatial autocorrelation of the residuals of the model of Y on X.

Call the MEM.moransel.R (supplementary material of Bauman et al. 2017):

```
source("MEM.moransel.R")
moransel <- MEM.moransel(Y, C, listw, MEM.autocor = MEM_model, nperm = 999, alpha = 0.05)

if (class(moransel) == "list") {
  MEM.select <- moransel$MEM.select
} else print("No significant spatial autocorrelation was detected in the response")
```

II.2. Selecting EVs while describing space as accurately as possible

If the response is multivariate, and/or if the purpose of the spatial EVs is to capture as much spatial structure as possible in Y, whether it is related to a set of explanatory (often environmental) variables (X) or not, then the forward selection (FS) of Blanchet et al. (2008) is to be prefered.

A first mandatory step before performing the FS is to check the significance of the global model, that is, the model of Y as a function of all the predictors. The FS can only be performed if this global test is significant at a predefined alpha threshold (here, 0.05). This step was shown to control the Type I error rate that otherwise can be highly inflated (Blanchet et al. 2008).

If MEM_model = "all" (we are interested in both positively and negatively autocorrelated patterns), then two separate global tests are performed, on the EVs displaying positive and negative eigenvalues, respectively. A p-value correction for multiple testing is then applied (Sidak correction) and the FS is performed only if at least one of the two tests is significant.

The FS with two stopping criteria consists in 1) searching the EV that best explains Y (highest R^2 adjusted by the Ezekiel correction, 1929), then 2) to search for the next EV best explaining the residuals of Y on the first selected EV, etc. At each selection step, two stopping criteria are used to accept the next best EV or stop the procedure: a) the p-value of the added EV (as in the classical forward selection), and b) the adjusted R^2 of the global model (including all predictors). This second criterion was shown by Blanchet et al. to avoid model overfitting (one of the main issues of the classical forward selection).

Here, we consider the whole community dataframe (multivariate response):

```
Y <- mite
Y <- decostand(Y, "hellinger")
```

Eigenvector selection using the forward selection with double stopping criterion:

```
if (MEM_model != "all") {

  if (anova.cca(rda(Y, MEM), permutations = 9999)$Pr[1] <= 0.05) {
    R2adj <- RsquareAdj(rda(Y, MEM))$adj.r.squared
    fsel <- forward.sel(Y, MEM, adjR2thresh = R2adj, nperm = 999)
    sorted_sel <- sort(fsel$order)
    MEM.select <- as.data.frame(MEM)[, c(sorted_sel)]
  } else print("No significant spatial autocorrelation was detected in the response")

} else {
  mem.sign <- vector("list", 2)  # List to save the positive and negative selected EVs
  signif <- c("FALSE", "FALSE")
  for (i in 1:2) {
    if (i == 1) {
      mem <- MEM[, which(attributes(MEM)$values > 0)]
    } else {
      mem <- MEM[, which(attributes(MEM)$values < 0)]
    }
    if (anova.cca(rda(Y, mem), permutations = 9999)$Pr[1] <= (1-(1-0.05)^0.5)) {   # Sidak corr.
      R2adj <- RsquareAdj(rda(Y, mem))$adj.r.squared
      fsel <- forward.sel(Y, mem, adjR2thresh = R2adj, nperm = 999)
      sorted_sel <- sort(fsel$order)
      mem.sign[[i]] <- as.data.frame(mem)[, c(sorted_sel)]
      signif[i] <- "TRUE"
    }
  }
  if (length(which(signif == "FALSE")) != 2) {
    if (length(which(signif == "TRUE")) == 2) {
      MEM.select <- cbind(mem.sign[[1]], mem.sign[[2]])
    } else if (signif[1] == "TRUE") {
      MEM.select <- mem.sign[[1]]
    } else MEM.select <- mem.sign[[2]]
  } else print("No significant spatial autocorrelation was detected in the response")
}
```

The EVs of MEM.select can be used as (co)variables in an OLS or GLM model (univariate reponse), or in an RDA or CCA (multivariate response).

References

Blanchet, F. G. et al. 2008. Forward Selection of Explanatory Variables. - Ecology 89: 2623–2632.

Borcard, D., P. Legendre and P. Drapeau. 1992. Partialling out the spatial component of ecological variation. Ecology 73: 1045-1055.

Borcard, D. and P. Legendre. 1994. Environmental control and spatial structure in ecological communities: an example using Oribatid mites (Acari, Oribatei). Environmental and Ecological Statistics 1: 37-61.

Dray, S. et al. 2006. Spatial modelling: a comprehensive framework for principal coordinate analysis of

neighbour matrices (PCNM). - Ecol. Modell. 196: 483–493.

Ezekiel, M. 1929. The application of the theory of error to multiple and curvilinear correlation. - J. Am. Stat. Assoc. 24: 99–104.

Griffith, D. A. and Peres-Neto, P. R. 2006. Spatial modeling in Ecology: the flexibility of eigenfunction spatial analyses. - Ecology 87: 2603–2613.

Legendre, P., and Gallagher, E. D. 2001. Ecologically meaningful transformations for ordination of species data. - Oecologia 129(2): 271-280.