Guidelines of Spatial Eigenvector Selection in R

Disentangling good from bad practices in the selection of spatial or phylogenetic eigenvectors. - Ecography

Bauman et al. 2018

Objective of this document

This document provides general guidelines for selecting an optimal subset of spatial eigenvectors (MEM variables) depending on the purpose of the study and on the univariate 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
```

We transform the species 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 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 often causing connectivity to decrease with distance. The decrease can be linear, or follow a concave-down or concave-up curve (see Dray et al. 2006 for details), or any other function defined by the user. 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.

The function `createlistw` of the `adespatial` package offers an interactive way to create and visualize different connectivity criteria. This can help deciding which connectivity scheme is more adapted for a given study case.

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

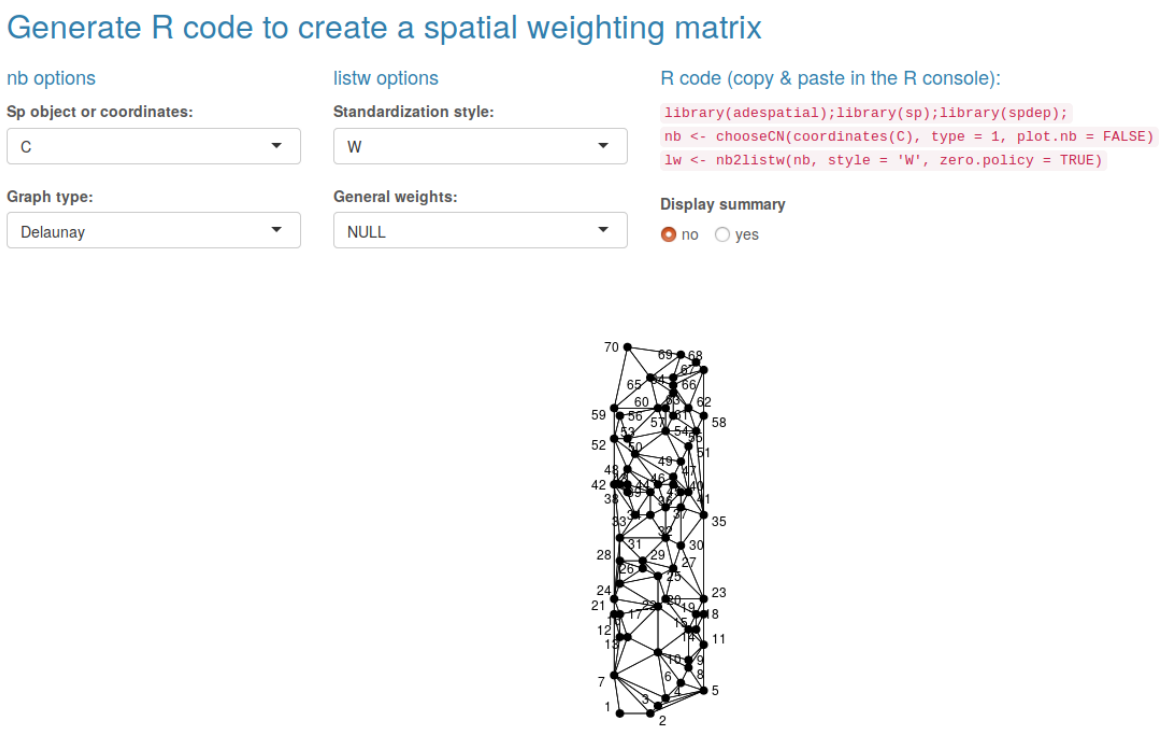
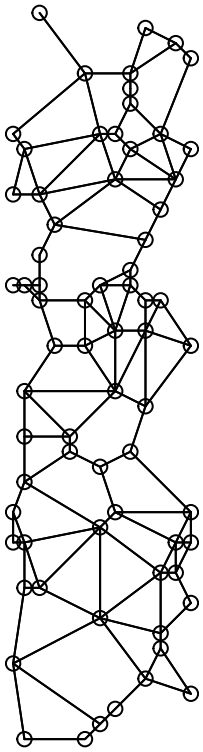


Figure 1: Screenshot of the interactive tool provided by the function `createlistw`.

Here, we use a Gabriel graph with no weighting to build the MEM variables, as an example to illustrate the unbiased eigenvector selection procedures. However, any W matrix can be used for the continuation of the procedure.

Gabriel graph of the mite dataset:

```
nb <- chooseCN(C, type = 2, plot.nb = FALSE)
lw <- nb2listw(nb, style = 'B', zero.policy = TRUE)
par(mar = c(0, 0, 0, 0))
plot(nb, C)
```



We start by defining some parameters. First, we define whether we want the MEM variables modelling positively autocorrelated patterns (`"positive"`), negatively autocorrelated patterns (`"negative"`), or all n-1 MEM variables (`"all"`):

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

Then, we define the standardisation scheme of the `listw` object (see help of `nb2listw` function), and then build the set of spatial eigenvectors (stored in object `MEM`). Depending on the value of `MEM_model`, the object `MEM` contains all n-1 eigenvectors, or only the eigenvectors associated with positive or negative eigenvalues (corresponding to positively and negatively autocorrelated patterns, respectively).

```
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 OLS or GLM model residuals

When the response data is univariate, if the purpose of the spatial predictors is to control the spatial autocorrelation of an OLS or GLM model, then the most suited MEM variable selection is that of Griffith and Peres-Neto (2006). This procedure selects the smallest MEM subset minimising spatial autocorrelation (Moran's I) in the residuals.

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

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

II.1.1. Selection of the spatial predictors based on the residuals of a model relating Y to a set of explanatory variables (X):

The Moran eigenvector filtering function ME allows removing spatial autocorrelation from the residuals of generalised linear models (see help of ME for details).

```
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)
MEM.select <- select$vectors
```

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

II.1.2. Selection of the spatial predictors based on Y only (MIR approach in Bauman et al. 2018, Fig. 1, Step 3.3.).

The `MEM.moransel` function focuses on controling the spatial autocorrelation of Y, instead of the spatial autocorrelation of the residuals of the model relating Y to X. The function does the same as ME, but with a model that contains only an intercept term.

Call the MEM.moransel.R (Appendix A3 in Bauman et al. 2018), construct the spatial predictors and select a subset of them following the MIR procedure.

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

`MEM.moransel` returns two dataframes containing all the spatial predictors (MEM.all) and the subset of predictors selected by the procedure (MEM.select). If no significant spatial autocorrelation could be detected in Y, then an informing message is printed.

II.2. Selecting MEM variables to describe space as accurately as possible

If the response data is multivariate, and/or if the purpose of the spatial predictors is to capture as much spatial structure as possible in Y (i.e., maximise the spatial fit), whether it is related to a set of explanatory variables (X) or not, then the forward selection (FWD) of Blanchet et al. (2008) should be preferred.

A first mandatory step before performing the FWD is to check the significance of the global model, that is, the model of Y as a function of all the spatial predictors. The FWD can only be performed if this global test is significant at a predefined threshold of null hypothesis rejection (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 MEM displaying positive and negative eigenvalues, respectively. A p-value correction for multiple testing is then applied (Sidak correction) and the FWD is performed only if at least one of the two tests is significant (see Blanchet et al. 2008).

The FWD with two stopping criteria consists in 1) searching the MEM variable that best explains Y (highest adjusted R^2 adjusted by the Ezekiel correction, 1929), then 2) to search for the next MEM best explaining the

residuals of Y on the first selected MEM, etc. At each selection step, two stopping criteria are used to accept the next best MEM or stop the procedure: a) the p-value of the added MEM (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. (2008) to avoid model overfitting (one of the main issues of the classical forward selection).

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

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

Eigenvector selection using the forward selection with double stopping criterion:

```
if (MEM_model != "all") {   # We consider only positively or negatively autocorrelated MEM

  if (anova.cca(rda(Y, MEM), permutations = 9999)$Pr[1] <= 0.05) {
    # Global adjusted R-squared of the model
    R2adj <- RsquareAdj(rda(Y, MEM))$adj.r.squared
    # FWD with two stopping criteria
    fsel <- forward.sel(Y, MEM, adjR2thresh = R2adj, nperm = 999)
    # We order the selected MEM by decreasing eigenvalue
    sorted_sel <- sort(fsel$order)
    # Object containing the selected MEM
    MEM.select <- as.data.frame(MEM)[, c(sorted_sel)]
  } else print("No significant spatial autocorrelation was detected in the response")

} else {   # We consider both positively and negatively autocorrelated predictors
  # List to save the positively and negatively autocorrelated MEM separately
  mem.sign <- vector("list", 2)
  signif <- c("FALSE", "FALSE")
  # We select the positive and negative MEM separately after testing the global
  # significance of both models at a corrected threshold value of null hypothesis
  # rejection (Sidak correction)
  for (i in 1:2) {
    if (i == 1) {   # Positive MEM
      mem <- MEM[, which(attributes(MEM)$values > 0)]
    } else {        # Negative MEM
      mem <- MEM[, which(attributes(MEM)$values < 0)]
    }
    # Global test of significance with the Sidak correction for multiple tests
    if (anova.cca(rda(Y, mem), permutations = 9999)$Pr[1] <= (1-(1-0.05)^0.5)) {
      # Global adjusted R-squared of the model
      R2adj <- RsquareAdj(rda(Y, mem))$adj.r.squared
      # FWD with two stopping criteria
      fsel <- forward.sel(Y, mem, adjR2thresh = R2adj, nperm = 999)
      # We order the selected MEM by decreasing eigenvalue
      sorted_sel <- sort(fsel$order)
      # We save the selection of MEM
      mem.sign[[i]] <- as.data.frame(mem)[, c(sorted_sel)]
      signif[i] <- "TRUE"
    }
  }
  # MEM.select will contain both positive and negative MEM, only positive or only
  # negative MEM, depending on the significance of the global tests.
  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")
}
```

```
## Testing variable 1
## Testing variable 2
## Testing variable 3
## Testing variable 4
## Testing variable 5
## Testing variable 6
## Testing variable 7
## Testing variable 8
## Testing variable 9
## Testing variable 10
## Testing variable 11
## Testing variable 12
## Testing variable 13
## Testing variable 14
## Procedure stopped (alpha criteria): pvalue for variable 14 is 0.084000 (> 0.050000)
```

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

References

Bauman, D. et al. 2018. Disentangling good from bad practices in the selection of spatial or phylogenetic eigenvectors – Ecography

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.