The MEMGENE package for R: Tutorial

Paul Galpern[1,2] and Pedro Peres-Neto[3]

[1] *Faculty of Environmental Design, University of Calgary*
[2] *Natural Resources Institute, University of Manitoba*
[3] *Departement des sciences biologiques, Universite du Quebec a Montreal*

Contents

1 Introduction

MEMGENE is a tool for spatial pattern detection in genetic distance data. It uses a multivariate regression approach and Moran's Eigenvector Maps (MEM) to identify the spatial component of genetic variation. MEMGENE variables are the output, and can be used in visualizations or in subsequent inference about ecological or movement processes that underly genetic pattern.

Please see the publication associated with the MEMGENE package (Galpern et al., 2014) for more information.

Two tutorials are presented here. The first shows a MEMGENE analysis of a simulated data set produced for the publication associated with the MEMGENE package (Galpern et al., 2014) and a second demonstrates an analysis for field-collected caribou data contained in the same paper.

2 Tutorial 1: Simulated data set

This tutorial demonstrates how to use MEMGENE under typical analysis conditions. It is possible to reproduce these examples directly in R. The tutorial focuses on the radial data set, which is also provided with the package.

2.1 The data set

A full description of how the radial spatial genetic data were simulated is available in the publication associated with this package (Galpern et al., 2014). Briefly, we simulated the moving and mating of 1000 individuals over 300 non-overlapping generations. Movement across the arms of the radial structure (Figure 1) was less likely than within the three regions of the landscape, due to landscape resistance to movement imposed on the simulated individuals. This makes the radial structure into a semi-permeable barrier, reducing dispersal and therefore gene flow. Given a sufficient number of generations for genetic drift under reduced gene flow, we expect a spatial genetic pattern that reflects the landscape resistance pattern in Figure 1.

The data set provided with the package (`radial.csv` installed in the `extdata` folder) represents a spatially stratified sampling of 200 individuals at generation 300 of this simulation. It includes 200 rows, one for each individual, two columns giving coordinates at which the individual was "sampled", and 30 paired columns giving the alleles at 15 codominant loci.

2.2 MEMGENE analysis

Step 1 Produce a genetic distance matrix

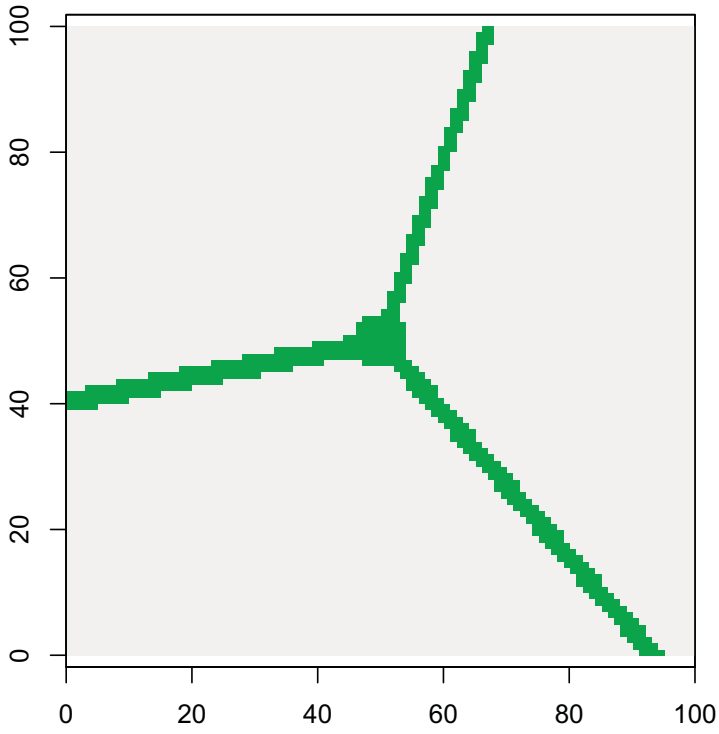

Figure 1: The radial resistance surface used to generate the spatial genetic data set used in this tutorial

MEMGENE requires a genetic distance matrix giving the pairwise genetic distances among individual genotypes. Any genetic distance metric can be used. In principle the method will also work with a population genetic distance matrix (e.g. pairwise Fst).

In this first step we find the genetic distance matrix using the proportion of shared alleles among individuals (Bowcock et al., 1994) as the metric. We use a convenience function included in the package to produce this that wraps functions in the `adegenet` package (Jombart, 2008).

```
> ## Load the radial genetic data
> radialData <- read.csv(system.file("extdata/radial.csv", package="memgene"))

> ## Create objects for positional information and genotypes
> radialXY <- radialData[ ,1:2]
> radialGen <- radialData[, 3:ncol(radialData)]

> ## Produce a proportion of shared alleles genetic distance matrix
> ## using the convenience wrapper function provided with the package
> radialDM <- codomToPropShared(radialGen)
```

Step 2 Extract MEMGENE variables

In this second step we extract the MEMGENE variables, using the typical interface to the MEMGENE package (the `mgQuick` function). The analysis framework is discussed in detail in the publication associated with this package.

The `mgQuick` function does the following: (1) Finds the Moran's eigenvectors given the sampling locations of the individuals (`mgMEM` function); (2) Uses these eigenvectors to identify significant spatial genetic patterns (`mgForward` and `mgRDA` functions); (3) Returns MEMGENE variables that describe these significant patterns on a reduced set of axes (`mgRDA` function). For additional detail on these functions, and for more control over the MEMGENE analysis see the R help files.

```
> ## Run the MEMGENE analysis
> ## May take several minutes
> if (!exists("radialAnalysis")) radialAnalysis <- mgQuick(radialDM, radialXY)
```

Step 3 Visualize MEMGENE variables

The MEMGENE variables represent orthonormal patterns of significant spatial genetic variation, and are ordered in terms of the amount of variation they explain from most to least. Typically, much of the variation is summarized in the first two variables, so it can often be convenient to visualize these two initially.

```
> ## Visualize the first two MEMGENE variables
> ## by providing only the first two columns of the $memgene matrix
> mgMap(radialXY, radialAnalysis$memgene[, 1:2])
```

However, it is often more interesting to visualize the MEMGENE variables superimposed over some map or other. In Figure 2 we superimpose the first MEMGENE variable

(MEMGENE1) over the resistance surface used to create the spatial genetic data. This can be done using the `add.plot=TRUE` parameter.

Although visualization may often be an end in itself, the MEMGENE variables can also be used singly or in combination to test hypotheses about the creation of the spatial genetic neighbourhoods they describe.

3 Tutorial 2: Wildlife data set

This tutorial demonstrates the use of MEMGENE with a data set for boreal woodland caribou, a North American ungulate.

3.1 The data set

A full description of how these spatial genetic data were collected and genotyped can be found in the publication associated with this package (Galpern et al., 2014). Briefly, these are genotypes for 87 caribou sampled on both sides of the Mackenzie River (Northwest Territories, Canada). The Mackenzie is a major North American river that varies betwee 1 and 4.5 km through the study area. Caribou have occasionally been reported crossing the river.

Boreal woodland caribou are a threatened species under Canada's Species at Risk Act. For this reason the caribou data included with the package have obfuscated sampling locations produced by reprojecting them in a way that maintains the Euclidean distance matrix among the points, but is not easily assignable to a precise location on the Earth's surface.

3.2 MEMGENE analysis

Step 1 Produce a genetic distance matrix

```
> ## Load the caribou genetic data
> caribouData <- read.csv(system.file("extdata/caribou.csv", package="memgene"))

> ## Create objects for positional information and genotypes
> caribouXY <- caribouData[ ,1:2]
> caribouGen <- caribouData[, 3:ncol(caribouData)]

> ## Produce a proportion of shared alleles genetic distance matrix
> ## using the convenience wrapper function provided with the package
> caribouDM <- codomToPropShared(caribouGen)
```

Step 2 Extract MEMGENE variables

```
> ## Run the MEMGENE analysis
> ## May take several minutes
> if (!exists("caribouAnalysis")) caribouAnalysis <- mgQuick(caribouDM, caribouXY)
```

```
> library(raster)
> radialRas <- raster(system.file("extdata/radial.asc", package="memgene"))
> plot(radialRas, legend=FALSE)
> mgMap(radialXY, radialAnalysis$memgene[, 1], add.plot=TRUE, legend=TRUE)
```

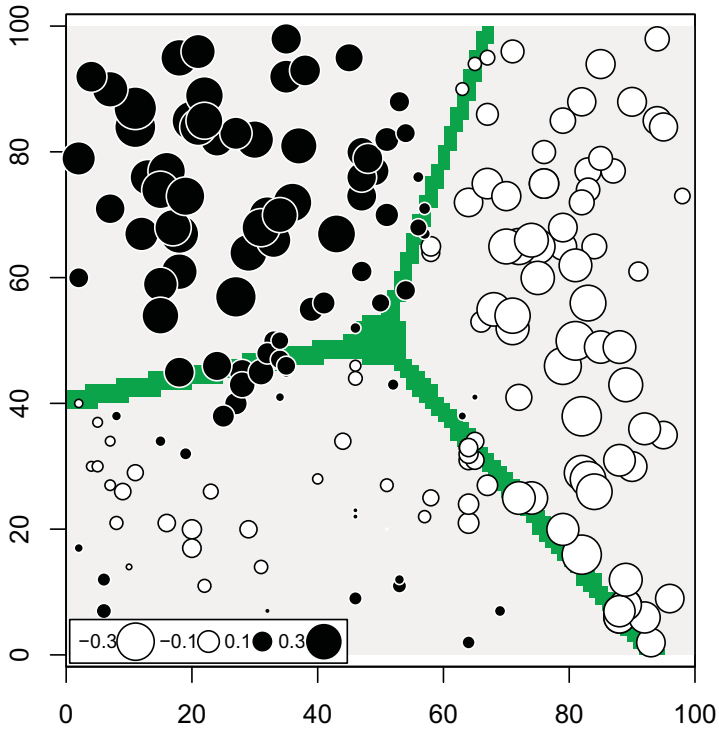

Figure 2: The scores of individuals on the MEMGENE1 axis superimposed on the resistance surface used to create the spatial genetic data. Circles of similar size and colour represent individuals with similar scores on this axis. Note how the pattern of spatial genetic variation in MEMGENE1 (spatial genetic neighbourhoods) reflects the structure of the landscape used to create it.

Step 3 Visualize MEMGENE variables

The results of the visualization of MEMGENE1 is shown in Figure 3. This figure also appears in the publication associated with this package, sumperimposed over a map of the region.

Step 4 Additional interpretation

Finding the adjusted R-squared (i.e. the genetic variation explained by spatial pattern) is just a matter of referencing the list element in the `caribouAnalysis` object as follows:

```
> caribouAnalysis$RsqAdj
```

```
[1] 0.02905906
```

Note that this low value should be interpreted not as an inadequacy of the regression to explain variation, but rather that there is only a small proportion of all genetic variation that can be attributed to spatial patterns; or more specifically, to the `N-1` (where `N` is the number of sampling locations) MEM spatial eigenfunctions that were extracted. It is important to note, however, that adjustments to how the MEM eigenfunctions are extracted have the potential to subtly change which spatial patterns are captured, as well as increase R squared. Further work is required to explore the effects of these modelling decisions.

Then determining the proportion of the this variation that is explained by each of the MEMGENE variables is also straightforward:

```
> ## Find the proportional variation explained by each MEMGENE variable
> caribouMEMGENEProp <- caribouAnalysis$sdev/sum(caribouAnalysis$sdev)
```

```
> ## Neatly print proportions for the first three MEMGENE variables
> format(signif(caribouMEMGENEProp, 3)[1:3], scientific=FALSE)
```

```
      MEMGENE1       MEMGENE2       MEMGENE3
"0.7220000000" "0.2780000000" "0.0000000153"
```

It is clear that there are only two distinctive patterns in these data, and the dominant pattern is that created by the Mackenzie River (i.e. MEMGENE1)

References

Bowcock AM, Ruizlinares A, Tomfohrde J, et al. (1994) High resolution of human evolutionary trees with polymorphic microsatellites. *Nature*, 368, 455-457.

Galpern, P., Peres-Neto, P., Polfus, J., and Manseau, M. (2014) MEMGENE: Spatial pattern detection in genetic distance data. *Submitted*.

Jombart T. (2008) adegenet: a R package for the multivariate analysis of genetic markers *Bioinformatics* 24: 1403-1405.

```
> plot(caribouXY, type="n", xlab="", ylab="", axes=FALSE)
> mgMap(caribouXY, caribouAnalysis$memgene[, 1], add.plot=TRUE, legend=TRUE)
> box()
```

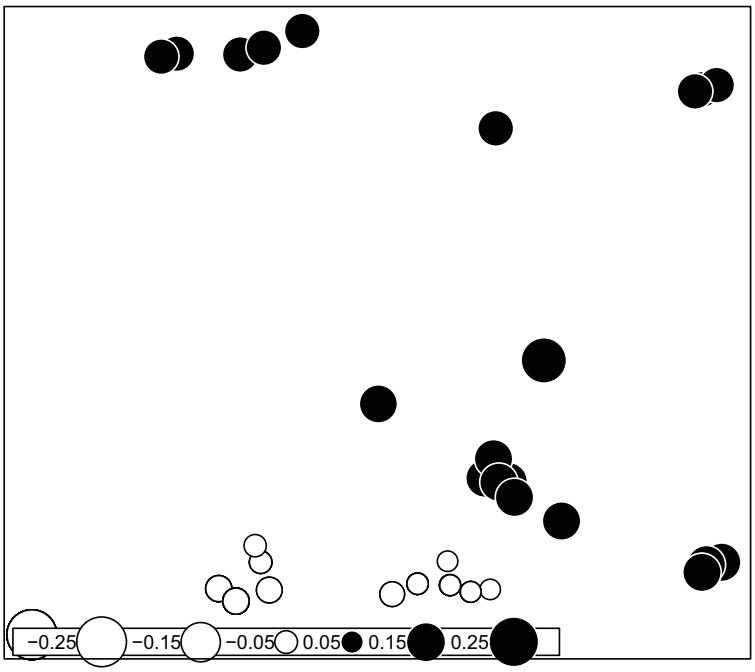

Figure 3: The scores of individual caribou on the MEMGENE1 axis. The Mackenzie River separates the white and black circles diagonally through the lower half of the map (not shown). For the full presentation of these results see the publication associated with this package.