Squared Euclidean distance as genetic dissimilarity

At a single SNP k, the squared Euclidean distance between individuals i and j is

$$d_{ij} = \left(z_{ik} - z_{jk}\right)^2$$

Across a (genome-wide) set of p markers, the squared Euclidean distance between individuals i and j is

$$D_{ij} = \sum_{k=1}^{p} \left(z_{ik} - z_{jk}\right)^2$$

[`bed2diffs` computes the average Euclidean distance, D_{ij}/p. Scaling by a positive constant does not affect the properties I discuss next.]

The EEMS program checks that the input pairwise `diffs` matrix, $D = \left(D_{ij}\right)$, is a valid Euclidean distance matrix.

Such a matrix has some properties that are obvious from the definition:

- There are only 0s on the main diagonal.
- The off-diagonal entries are nonnegative, and if no individual is an exact copy of another, the off-diagonal entries are strictly positive.
- The matrix is symmetric.

However, it turns out that a nonnegative symmetric matrix with 0s on the diagonal is not guaranteed to be a distance matrix. [That is, the three conditions above are not sufficient.]

Sufficient conditions for a Euclidean distance matrix

The matrix is nonnegative, symmetric, with 0s on the main diagonal and

- There is one positive eigenvalue.
- The other $n - 1$ eigenvalues are negative.

If the distance matrix is not full-rank, then some eigenvalues that should be negative are actually equal to 0.

To summarize, the EEMS program check that

- D is symmetric.
- The diagonal entries are 0.
- The off-diagonal entries are nonnegative.
- There is exactly one positive eigenvalue.
- There are no eigenvalues equal to 0.

Is the matrix of F_{ST}s a valid distance matrix?

I used GENEPOP to compute pairwise F_{ST}s for the POPRES dataset, and furthermore, only between 15 Western European populations.

References:

- POPRES (Population Reference Sample) project, `https://www.ebi.ac.uk/ega/studies/phs000145.v2.p2`
- Rousset. GENEPOP'007: a complete re-implementation of the GENEPOP software for Windows and Linux. *Mol Ecol Resour*, 8:103–106, 2008

Is the matrix of F_{ST}s a valid distance matrix?

However, it turns out that the F_{ST} matrix is not a distance matrix.

This can happen if there do not exist "features" $X = (x_{ik})$ such that

$$D_{ij} = \sum_k \left(x_{ik} - x_{jk}\right)^2$$

[How many features? For n items we need at most $n - 1$ features.]

Use multidimensional scaling to produce a distance matrix

Let F be the matrix of pairwise F_{ST} values.

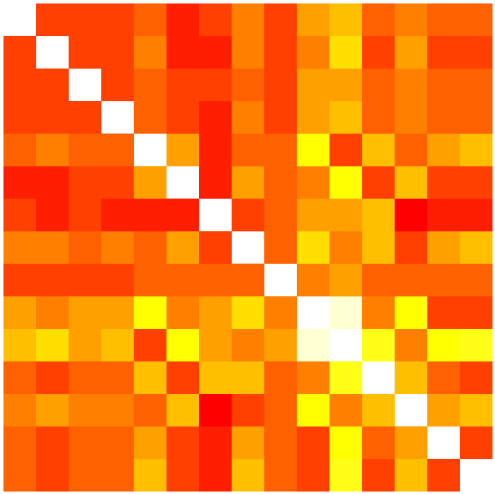
If the observed F is not a valid Euclidean distance matrix, we can approximate it with its “closest” Euclidean distance matrix, D, using multidimensional scaling.

Here is how to do it in R.

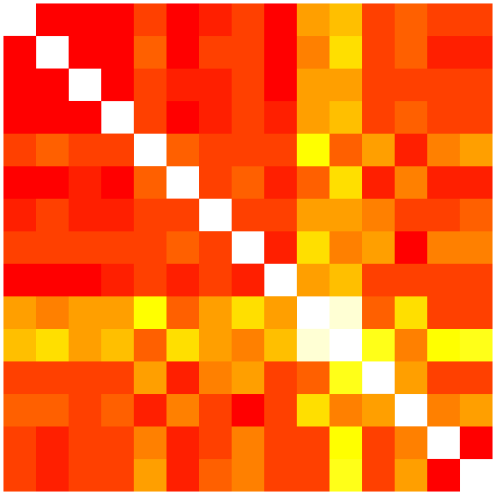
```r
## Perform metric multidimensional scaling
## Let F be the matrix of pairwise FSTs
n = nrow(F)

## The X are the features that generate the D that is "closest" to F
X = cmdscale(F,k = n-1)

## You might get a warning:
## In cmdscale(F, k = n - 1) : only k of the first n-1 eigenvalues are > 0

## k is the maximum dimension of the space which the data are to be
## represented in; must be in {1, 2, ..., n-1}

## Compute the Euclidean distance matrix
D = as.matrix(dist(X,method = "euclidean"))
```

Use multidimensional scaling to produce a distance matrix

The observed F matrix

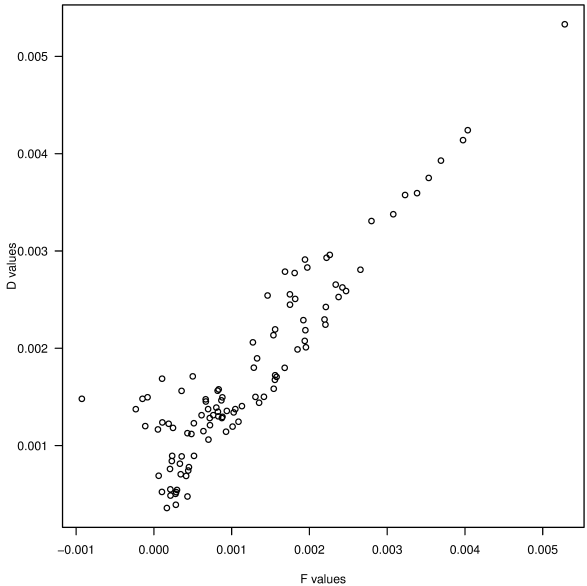


The approximation D



I have plotted the F_{ST} matrix F and its approximation D as heat maps. But it might be better to make a scatter plot of F vs D values.

Scatter plot for F vs D values



Now we have a difference matrix D. The question whether it is a good idea to use it with EEMS is still open.

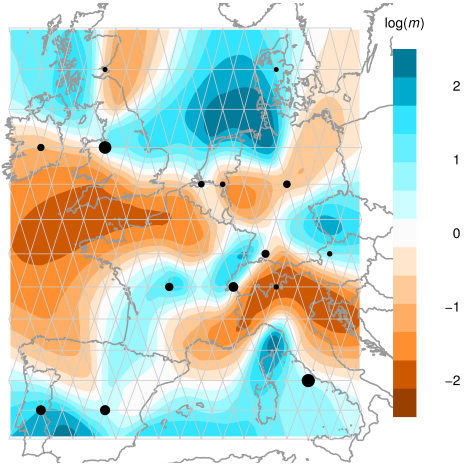
One obvious difference is that the matrix of F_{ST}s is a matrix of dissimilarities between populations, not individuals.

On the other hand, EEMS is designed to model the dissimilarities between individuals, without the need to group the individuals into groups.

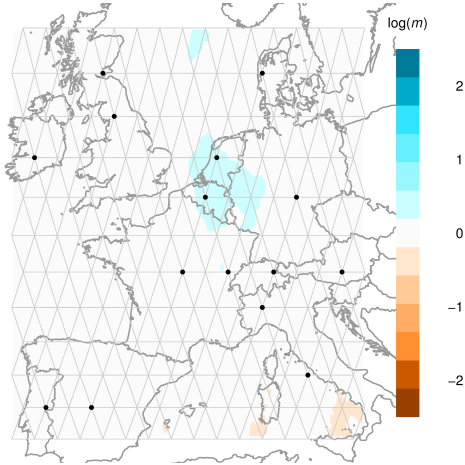
Anyway, I proceed by treating each population as an individual. [That is, as if there is a single observation from 15 distinct locations.]

EEMS results with two different dissimilarity matrices

`diffs` (from `bed2diffs`)



D (approximation to F_{ST}s)



It seems that it is not a good idea to use the approximation D with EEMS.

Population structure in terms of Wright's F-statistics

At this point it is useful to know about the 3 F-statistics: F_{ST}, F_{IS}, F_{IT}.

- Pairwise F_{ST}s contain information about genetic dissimilarities between different locations in the habitat.
- But no information about genetic dissimilarities between different individuals from the same location.

Wright's F-statistics decompose population structure into different components: genetic variation between individuals within subpopulations (F_{IS}) and genetic variation between subpopulations (F_{ST}).

EEMS also decomposes the genetic dissimilarity into two components, between demes (B) and within demes (W):

$$\begin{aligned}\Delta_{\alpha\beta} &= \underbrace{\Delta_{\alpha\beta} - \big(\Delta_{\alpha\alpha} + \Delta_{\beta\beta}\big)/2}_{\text{between demes}} + \underbrace{\big(\Delta_{\alpha\alpha} + \Delta_{\beta\beta}\big)/2}_{\text{within demes}} \\ &= \quad B_{\alpha\beta} \quad + \big(W_\alpha + W_\beta\big)/2\end{aligned}$$

Spatial population parameters in EEMS

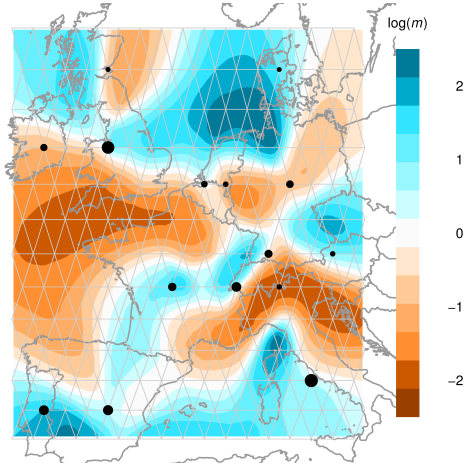
- The between-demes component is modeled by the effective migration rates m: $B = f(m)$.
- The within-demes component is modeled by the effective diversity rates q: $W = g(q)$.

The problem with using F_{ST}s in EEMS as it is might be that F_{ST}s contain information about the between-demes component (and thus about m) but not about the within-demes component (and thus not about q).

I have modified the EEMS program, so that there is only a between-demes component of genetic differentiation: $\mathrm{E}\{D\} = \Delta = B$.

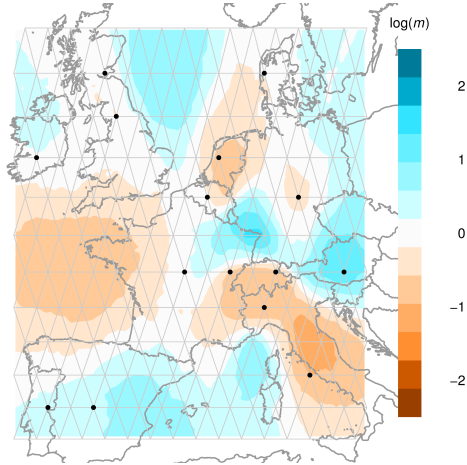
`diffs` (from `bed2diffs`)

with `runeems_snps`



D (approximation to F_{ST}s)

with `runeems_Fsts_v0`



We have reaped some success!

However, the colors are “lighter”, most likely due to loss of information in the observed data and/or the multidimensional scaling approximation.

In particular, a smaller matrix contains less information about spatial population structure:

- There is more uncertainty about spatial patterns of genetic variation in the observed data.
- The likelihood would be less peaked and hence the prior (that migration is uniform) contributes more to the posterior.
- The posterior would be “closer” to the prior, which corresponds to exact isolation by distance.

You can run the modified version `runeems_Fsts_v0` in exactly the same way as the original version `runeems_snps`.

`runeems_Fsts_v0` assumes that each deme is assigned at most one sample. If the grid is too coarse and two samples (= two subpopulations in the case of F_{ST}s) are assigned to the same deme, you will get the error:

```
Use this version only if there is at most one sample in every deme
```

If this happens, you can try increasing nDemes but this is not a general solution because of the computational cost involved.

Next I describe a general version of EEMS for pairwise F_{ST} matrices.

I assume that the matrix of pairwise F_{ST}s is processed as described previously, i.e., it is a valid Euclidean distance matrix.

The added wrinkle is that the population sampling is dense and it happens that two populations are assigned to the same deme in the EEMS graph.

I use a small example to illustrate the problem.

Suppose that there are five individuals labeled from i_1 to i_5. Let $\alpha_{\delta(i)}$ denote the deme (vertex) that individual i is assigned to.

Suppose further that individuals i_1 and i_2 are assigned to the same deme.

i	i_1	i_2	i_3	i_4	i_5
$\alpha_{\delta(i)}$	α_1	α_1	α_3	α_4	α_5

In `runeems_Fsts_v0` we model only the between-demes component of genetic dissimilarity: $\Delta_{\alpha\beta} = B_{\alpha\beta}$ rather than $\Delta_{\alpha\beta} = B_{\alpha\beta} + (W_\alpha + W_\beta)/2$.

What is the expected dissimilarity matrix for the example with five individuals assigned to four distinct demes?

$$\Delta = \begin{matrix} 0 & 0 & B_{\alpha_1\alpha_3} & B_{\alpha_1\alpha_4} & B_{\alpha_1\alpha_5} \\ 0 & 0 & B_{\alpha_1\alpha_3} & B_{\alpha_1\alpha_4} & B_{\alpha_1\alpha_5} \\ B_{\alpha_1\alpha_3} & B_{\alpha_2\alpha_3} & 0 & B_{\alpha_3\alpha_4} & B_{\alpha_3\alpha_5} \\ B_{\alpha_1\alpha_4} & B_{\alpha_2\alpha_4} & B_{\alpha_3\alpha_4} & 0 & B_{\alpha_4\alpha_5} \\ B_{\alpha_1\alpha_5} & B_{\alpha_2\alpha_5} & B_{\alpha_3\alpha_5} & B_{\alpha_4\alpha_5} & 0 \end{matrix}$$

The matrix is symmetric, nonnegative, with 0s on the main diagonal. However, it is not full-rank because the first and the second rows are exactly the same.

EEMS when D is a matrix of F_{ST}s and Δ is rank-deficient

Recall that the "full" EEMS likelihood is Wishart:

$$-LDL' \mid k, m, q, \sigma^2 \sim \mathrm{W}\left(k, -\frac{\sigma^2}{k} L\Delta(m,q)L'\right)$$

If only the between-demes component is modeled:

$$-LDL' \mid k, m \sim \mathrm{W}\left(k, -\frac{1}{k} L\Delta(m)L'\right)$$

[A mathematical detail: Note that $\sigma^2 = 1$. Effectively the scale σ^2 is incorporated into the expected dissimilarity matrix, so that $\Delta = \sigma^2 B$.]

It turns out that if Δ is not full-rank, then $-L\Delta L'$ is not full-rank. [In other words, $-L\Delta L'$ is singular.]

Consequently, $-L\Delta L'$ is not strictly positive definite (all eigenvalues are positive) but positive semi-definite (some eigenvalues are 0, the rest are positive).

In the example with five individuals assigned to four distinct demes, $-L\Delta L'$ is a 4×4 matrix, which has three positive eigenvalues and the last eigenvalue is 0.

Fortunately, the Wishart distribution can be generalized to work with a singular expected matrix.

The generalized Wishart likelihood is implemented in `runeems_Fsts` (let's hope – correctly). The standard Wishart likelihood is implemented in `runeems_Fsts_v0` but it requires at most one population in every deme.

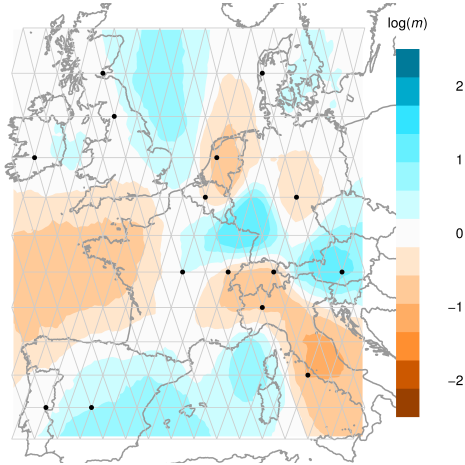
The second version is more expensive computationally. So it will run slower. Start with a smaller number of demes, say `nDemes = 200`.

Reference:

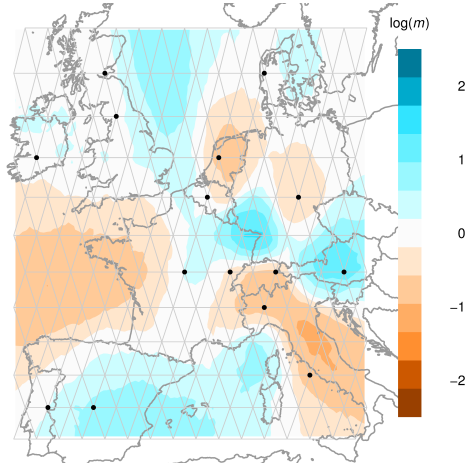
- Díaz-García, Jáimez and KV Mardia. Wishart and pseudo-Wishart distributions and some applications to shape theory. *J Multivar Anal*, 63:73–87, 1997

I run the two versions of EEMS for pairwise F_{ST} matrices with the same graph.

D with `runeems_Fsts_v0`



D with `runeems_Fsts`



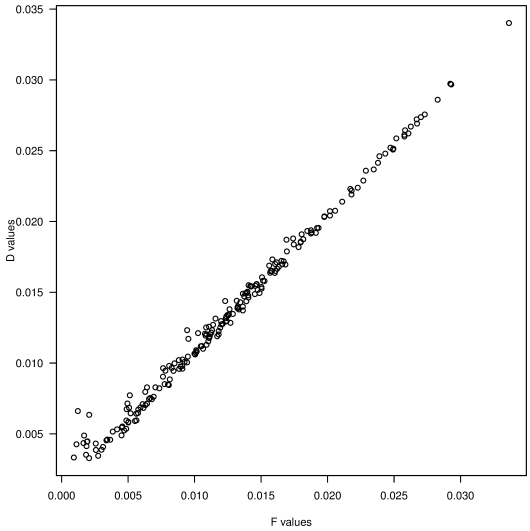
The estimated contour surfaces are very similar – this is the extent of my testing, so we should be cautious.

I have also tried the modified EEMS for pairwise F_{ST}s on a second dataset that comprises 21 ethnic groups from Sub-Saharan Africa.

Reference:

- Wang, Zöllner, and N. Rosenberg. A quantitative comparison of the similarity between genes and geography in worldwide human populations. *PLoS Genet*, 8:e1002886, 2012

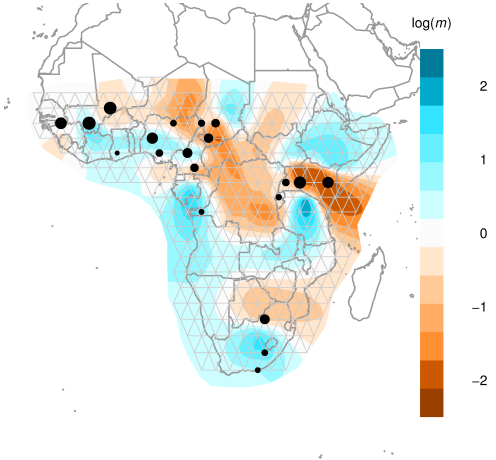
Scatter plot for F vs D values, after multidimensional scaling



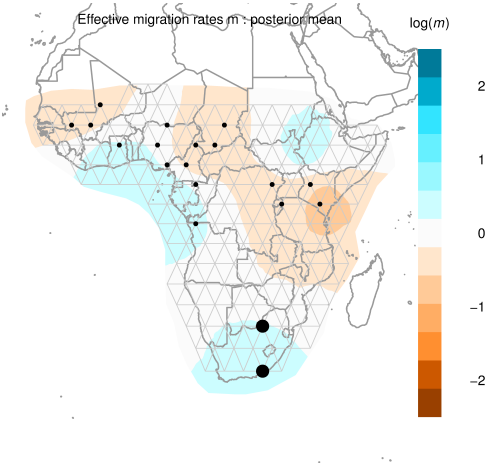
The valid distance matrix D is a very good approximation to the matrix F of pairwise F_{ST}s.

EEMS results with two different dissimilarity matrices

`diffs` with `runeems_snps`



D with `runeems_Fsts`



Note that the population grids are different – the grid is coarser on the right.

The plotting command using the rEEMSplots package

```
eems.plots(mcmcpath,
           plotpath,
           longlat = TRUE,
           add.grid = TRUE,
           add.demes = TRUE,
           add.title = FALSE,
           m.colscale = c(-2.5,+2.5),
           remove.singletons = FALSE)
```