Statistical methods in genetic relatedness and pedigree analysis

Oslo, January 2020

Magnus Dehli Vigeland and Thore Egeland

Exercise set V. Relatedness inference and pedigree reconstruction

Exercise V-1 (Reconstructing a pedigree from pairwise estimates)

The goal of this exercise is to reconstruct the pedigree relating 4 individuals, using genotype data from 1000 SNPs. The genotypes are contained in the ped file `reconstruction1.ped` which you can download from the course home page.

For this exercise you need to load the `forrel` package:

```
library(forrel)
```

a) Familiarise yourself with the file, by opening it in a good text editor (I use `Notepad++`) or a spreadsheet program like Excel. The file consist of four very wide lines, and has no column names.

   i. Can you identify the pedigree columns? (You can assume that they follow the conventional ped format.)
   ii. What are the genders?
   iii. How are the genotypes given?

b) Load the pedfile into R (change to the correct path on your computer):

```
x = readPed("reconstruction1.ped", famid = 1, id = 2, fid = 3, mid = 4, sex = 5,
            sep = "/", locusAttributes = list(alleles = c("A", "B")))
```

Make sure you understand the meaning of the arguments in the above call. Note in particular the `locusAttributes` argument, which tells `readPed()` that all the markers are SNPs with alleles A and B. Without this information the function would misinterpret loci showing only a single allele.

c) Use the following code to estimate the IBD coefficients between each pair of individuals and plot the results in the IBDtriangle.

```
ibd = IBDestimate(x)
ibd
showInTriangle(ibd, labels = TRUE)
```

Note that if the labels are hard to discriminate, you can select which rows of `ibd` to plot, and also play with the optional arguments of `showInTriangle()`. For example, the `pos` argument indicates the position of labels relative to the points (1 = below, 2 = left, 3 = above, 4 = right), while `cex_labels` may be used to increase the font size. Try this:

```
showInTriangle(ibd[1:3, ], labels = TRUE, pos = 3, cex_labels = 2)
```

d) Reconstruct the pedigree!

Exercise V-2 (Pedigree reconstruction with `pedbuildr`)

In this exercise you will need the `pedbuildr` package, which you can download from github if you haven't done it already:

```
devtools::install_github("magnusdv/pedbuildr")
```

If the above command fails, try this:

```
install.packages("http://familias.name/norbisRelatedness/pedbuildr_0.1.0.tar.gz", repos = NULL)
```

To get started, load the package:

```
library(pedbuildr)
```

We will apply the `reconstruct()` function of `pedbuildr` to the data set we analysed in Exercise 1. What `reconstruct()` does is the following:

1. Generate a list of all possible pedigrees connecting the given number of individuals
2. Compute the likelihood of each pedigree, given the genotype data
3. Sort the output so that the most likely pedigree comes first.

As this is very much a brute force approach, it is usually a good idea to put restrictions on the pedigree space. For example, by putting `pairwise = TRUE`, the program will run a preliminary pairwise analysis (much like we did in Exercise 1) to identify all certain parent-child relationships, and only allow pedigrees compatible with these. Another sensible restriction is `maxLinearInbreeding = 0`. This disallows *linear incest*, i.e. procreation between parent-child, grandparent-grandchild, a.s.o., which is usually a safe option in human applications.

The essential input objects to `reconstruct()` are:

- `alleleMatrix`: A matrix with alleles (two columns per marker) and one row per individual. The easiest way to get this is to run `getAlleles(x)`.
- `loci`: A list of marker attributes, i.e. alleles, frequencies a.s.o.
- `sex`: A vector indicating the sex of each individual. We can code this manually, or simply use `getSex()`.

a) Use the following code to create the objects needed by `reconstruct()`. Here `x` should be unchanged from the previous exercise.

```
als = getAlleles(x)
loci = list(afreq = c(A = 0.5, B = 0.5))
sex = getSex(x)
```

b) Run `reconstruct()`:

```
res = reconstruct(alleleMatrix = als,
                  loci = loci,
                  sex = sex,
                  pairwise = TRUE,
                  maxLinearInbreeding = 0,
                  genderSym = TRUE)
```

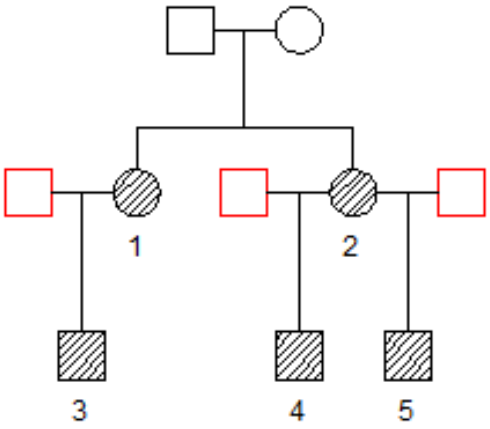
c) When the compuation finishes, use `plotBestPeds()` to inspect the most likely pedigrees:

```
plotBestPeds(res)
```

Did you get the same result as in Exercise 1?

Exercise V-3 (A question about fathers)

The following is based on a true case from Australia. Genotypes are available from two sisters and their children: The first sister has one child, the other has two children. The question we must answer is the following: *Do any of the children have the same father?*



At the course home page you will find the data needed for this exercise: The file `AustraliaData.ped` contains the genotypes for 24 forensic markers for the 5 individuals. In addition you need the R file `AustraliaFreqs.rds`, which contains a database with Australian allele frequencies for these markers.

a) Load the genotype data and set the correct frequencies

```
x = readPed("AustraliaData.ped", sep = "/")

# Attach the frequency database
db = readRDS("AustraliaFreqs.rds")
x = setLocusAttributes(x, locusAttributes = db)

# Inspect the result
x
```

b) Estimate the pairwise relationships, as in Exercise 1c), and plot the results in the IBD triangle. Comment the results. What are you most certain about? Why do you think the results this time are more difficult to interpret than in the previous exercise?

c) Use `pedbuildr` to estimate the most likely pedigree:

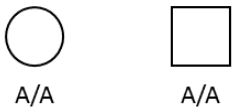
```
r = reconstruct(
  alleleMatrix = getAlleles(x),
  loci = getLocusAttributes(x),
  sex = getSex(x),
  knownPO = list(c(1,3), c(2,4), c(2,5)), # Undisputed PO relationships
  allKnown = TRUE,                        # No other PO's among these 5
  connected = TRUE,                       # Only consider connected pedigrees
  maxLinearInbreeding = 0,                # Don't allow parent-offspring incest
  genderSym = TRUE                        # Remove gender symmetries in added parents
)
```

d) Plot the top pedigrees. What is your conclusion?

```
plotBestPeds(r)
```

Exercise V-4 (ML-estimation by hand - for the mathematically inclined)

This exercise walks you through the computations of a maximum likelihood estimation of the relationship between two non-inbred individuals. To enable hand calculation, the data is the simplest possible: Genotypes from a single marker, for which both individuals are homozygous A/A. Denote by p the population frequency of the A allele, and assume $0 < p < 1$.



a) What do you think is the most likely relationship given this data?
Recall that maximum likelihood estimation of pairwise relatedness works by finding the value of $k = (k_0, k_1, k_2)$ that maximises the likelihood function

$$\begin{aligned} L(k) &= P(data \mid k) && (1)\\ &= P(data \mid UN) \cdot k_0 + P(data \mid PO) \cdot k_1 + P(data \mid MZ) \cdot k_2, && (2) \end{aligned}$$

where the *data* are the observed genotypes, and UN, PO and MZ denote the relationships of unrelated, parent-child and MZ twins, repsectively.

b) Show that the likelihood function in this case becomes

$$L(k) = P(AA, AA \mid k) = p^4 k_0 + p^3 k_1 + p^2 k_2. \quad (3)$$

c) Explain that for optimisation purposes you can get rid of a factor p^2. Furthermore, use the relation $k_0 + k_1 + k_2 = 1$ to eliminate k_1, giving the simpler function

$$L_1(k_0, k_2) = (p^2 - p)k_0 + p + (1 - p)k_2. \quad (4)$$

d) Remove another factor, $(1 - p)$, and conclude that all we have to do is to maximize the function

$$L_2(k_0, k_2) = k_2 - pk_0 + C, \quad (5)$$

where C is a constant.

e) Which point (k_0, k_2) in the IBD triangle gives the highest value of the function L_2 obtained in the previous step? What is the estimated relationship?